Work Flow for the Weighting of the German Microcensus Data Using Blaise Bascula

Kirsten Iversen (Federal Statistical Office, Germany)

1. Introduction

The German Microcensus originally focussed on a specified reference week in spring. In January 2005 it started as a continuous survey in order to obtain estimates of quarterly and annual means. The data are collected locally, i.e. the Statistical Offices of the 16 federal Länder in Germany do the field work. Participation in the survey is compulsory.

For every quarter the data are weighted in two steps: Firstly, there is compensation that adjusts the known non-responses. In the second step, the sample distribution of some auxiliary variables (for example age, sex and nationality) is adjusted to the known population totals. The intention of this bounded second-step weighting is to reduce the bias of estimates due to unrecognised non-responses. The variance for variables which are highly correlated to the auxiliary variables is also minimised. Implementation of this two-phase weighting in Blaise Bascula is presented in Chapter 3.1.

Once a year the data for the four quarters are combined for deep regionalised weighting. The weighting program used for this is the subject of section 3.2.

The conclusion is formed by a brief overview of the size of the files and the running times of the programs (Chapter 4). To start with, the content of the processed data will now be briefly explained.

2. The Data

The data of the German Microcensus are collected continually, spread equally over the whole year. The questionnaire used in both the CAPI survey and the paper-and-pencil version comprises roughly 190 questions on the topics of personal data, working life, education, income and housing. In addition, further variables are generated from the information collected which are required as characteristics for the weighting.

An abridged dataset is extracted for the weighting. It covers only 29 variables which are needed directly for compensation and weighting, such as period under report, Federal Land, administrative district, household characteristics, nationality, sex, age and individual characteristics on employment. These datasets, comprising 60 variable labels, form the input file for the weighting in the shape of an ASCII file “MZHR1.asc”.

The “MZHR1” file contains in each case the observations of one quarter. All four quarters in a year are separately compensated and weighted, and are only placed together in the deep regionalised annual weighting. The annual sample covers roughly 1% of the population, that is, depending on the Federal Land, between 670 and 15,000 individuals per quarter (¼ %). This is a proportionally-stratified cluster sample.

Apart from the datasets of the quarter, the bounded weighting also includes benchmark values relating to the population structure. These originate from the

current population adjustment of the population statistics which are based on the 1987 census (West) and on a 1990 register excerpt (East). Figures for the foreign population are added from the Central Foreigners' Register. In addition to the total population size of the Federal Land in question in the respective quarter, figures are also shown separately for the individual administrative districts on sex and nationality. The combinations of these characteristics which arise are shown in a table which is contained in a .pop file, and are incorporated into the weighting as benchmarks. For the deep regionalised annual weighting, figures are additionally included at the level of the regional sub-groups which also come from the current population adjustment.

3. Running of the program

The total weighting consists of five programs, as well as of four batch processes. These are the batch processes of the compensation and actual weighting; they are run in Bascula. The programs, by contrast, run under Blaise and prepare the data for the batch processes: They separate the respondents by individual characteristics, re-sort the datasets as required and then add the weighting factors back to the datasets throughout the whole file.

First of all, the operation of the quarterly weighting is explained (programs 1 to 3, as well as batches 1 to 3), followed by a brief description of the conversion of the annual weighting. A dataflow plan of the whole program run can be found in the Annex (A1).

3.1. Steps in the quarterly weighting

The extraction of the abridged datasets, which now only contain the compensation and weighting-relevant variables, takes place prior to weighting. The "MZHR1.asc" file, which contains the abridged datasets of a quarter, is the input material on which the entire quarterly weighting is built.

Program 1:
Program 1 is a preliminary program which creates further characteristics from the variables of MZHR1, specifically for compensation and weighting, such as nationality (German/EU25/non-EU), a breakdown into age classes, household sizes, questioning at main or secondary residence. Furthermore, three specialised files are created from the input file: One covers all respondents[8], another file contains only respondents from community households (CHH), while the third file holds only the household reference persons of the private households (PHH). The datasets are weighted to cover additional fields for the compensation factors and for the subsequently-calculated weighting factor. For the household reference person in a private household, and for the respondent in a community household unit, the value one is entered in the datasets to provide an inclusion weight for compensation.

After the sub-division, an examination takes place of whether the file of the community households contains at least ten datasets. If this is not the case, no further benchmark values are calculated for the community households, and the Batch 2 process (compensation for non-respondent community household units) is skipped.

[8] "Respondents" always refers to the persons who actually answered, in other words all persons of a sample not including the non-response cases, the share of which is as a rule under 3%.

Program 1 also calculates the reference values to compensate for non-respondent households. To this end, for the individual characteristic combinations of the compensation characteristics, frequency counts are carried out in the input material (in other words with respondent and non-respondent persons). The reference values all have at least the value zero, and are output in table form in a .pop file, the structure of which is specified in a .blg file. Separate benchmark value files are created in each case for private households and community household units. All regional adjustment layer numbers in a Federal Land which are not occupied, as well as characteristics represented neither with respondents nor with non-respondents, are directly assigned the value zero.

After these preparatory steps, the compensation factors for non-respondent households can now be calculated (non-response weights), via which the net sample is weighted to the gross sample. The compensation for private households and community household units is carried out in separate batch processes.

Batch 1:
The Batch 1 process calculates compensation factors for the household reference persons of private households with which a shortfall of other private households is to be compensated for and the resultant distortion of the results reduced. The administrative district, as well as the regional adaptation layer and sub-group, the rotation quarter, the label “new buildings layer”, the household size, the nationality (German/not German), main or secondary residence, the age group (under 60/60 years and older) and sex, are used as compensation characteristics. The combinations of the administrative district and the regional adjustment layer with the further variables are used as a compensation model, the reference values are available in the .pop file created in program 1.

Linear weighting with constant is used as a weighting procedure. The inclusion weight of the value of one entered in program 1 is used as input weight. There must be at least ten observations in each combination of the compensation model. If this is not the case, the model is reduced to a higher-order combination of the characteristics (the strongest differentiating factor of a combination is removed). The lower threshold of the correction weights and of the final weights is 0.001, and 3 is stated as the upper limit. The maximum number of iterations is 10 (more iterations are not possible in batch processes). The permissible inconsistency of the benchmark values should not exceed 1, while the accepted relative deviation of the weighted values from the benchmark values is limited to 0.001.

An example syntax for the batch processes can be found in the Annex (A2).

Batch 2:
The Batch 2 process calculates the factors to compensate for non-respondent households in community households. It is only implemented if the file for community households contains at least ten datasets. Where there are fewer observations, the compensation of non-respondent households in community households is not carried out. The compensation model is limited to the administrative district. Reference values are available in the shape of the .pop file created in program 1.

The post-stratification method with constant is used for weighting, the input weight already having been set in program 1. The minimum cell occupation is ten datasets in this case too. If fewer observations are available in an administrative district, the model is reduced to the constant.

Each batch process outputs a protocol file which also documents, in addition to the (perhaps abridged) model used the adjustments in the individual cells, as well as descriptive statistics on the compensation factors.

Once the batch processes are completed, for the household reference persons in private households, and the persons in community households, in each case a file is now available in which the datasets are weighted by the respective compensation factor. These files are input to program 2.

Program 2:
The second program now allots to all respondent persons the respective compensation factors. Here, in addition to the household reference person in private households, all further household members are also allotted the factor of the reference person calculated in Batch 1.

The number of respondent persons (n) is then calculated. The number of the population total (N) is input as the target value. The compensation factor weighted with the quotients from these two figures (N/n) is used as inclusion weight in the weighting.

Like Program 1, Program 2 also outputs a file with all respondents. The datasets are now supplemented by the compensation factors and the inclusion weights for the weighting. In a copy of this file, the datasets are sorted by household number since these are then used as a cluster code. This file is input directly into the weighting in the Batch 3 process.

Batch 3:
This process calculates the standard weighting factors in the quarter. Administrative district, regional adaptation layer, sex, nationality (German/not German and German/Turkish/EU25/not EU25), an age classification and a label for employees of the Federal Armed Forces, police or at the Federal Border Guard serve as weighting characteristics. The household number is used as a cluster code so that all persons in a private household receive the same weighting factor. The weighting model consists of various combinations of the characteristics. Figures from the current population adjustment are included as reference values which are already available prior to the program launch in table form contained in a .pop file.

The weighting is carried out as in Batch 1 by the linearconsistent procedure with constant. The inclusion weight was calculated in Program 2. The minimum cell occupation is set at 10 (with the exception of one Federal Land), while with fewer observations, the model reduces to a higher-order combination of the characteristics. Once more, the lower threshold for the correction weight is set at 0.001, and the upper limit is 5. Once again, a maximum of 10 iterations should be carried out. The permissible inconsistency of the reference values is 50, while the relative deviation of the weighted values from the reference values should not exceed 0.001.

This batch process also outputs a protocol file which documents the (possibly abridged) model used, the adjustments to the individual combinations of characteristics, as well as descriptive statistics on the correction and final weight.

Once the Batch 3 process has been completed, a file is available containing all respondent persons which is sorted by household numbers and the datasets of which were weighted to include the weighting factor. This file is input into Program 3.

Program 3:
Program 3 calculates the standard weighting factors "year" as ¼ of the standard weighting factors "quarter" which were calculated in Batch 3. Then, both standard weighting factors are applied to the respective datasets in the unsorted file of all respondents (which was not used in Batch 3). Additionally, the compensation factors are attached to the datasets as 7-digit string variables and the standard weighting factors as 4-digit string variables, in each case with no comma.

The data are then output once as an ASCII file and once as a Blaise file. The Blaise file is incorporated into the continued statistical quarter processing. The ASCII file with the name MZHR6-Q*x* is retained for each quarter *x*; all four quarter files are then input together to the deep regionalised annual weighting. The quarterly weighting is completed at this juncture.

3.2. Weighting for deep regionalised annual results

In schematic terms, the annual weighting directly follows the quarterly weighting: The four quarter files MZHR6-Q1 to -Q4 are used as input material. Furthermore, population sizes per regional sub-group by sex and nationality are input from the current population adjustment as benchmark values; these are available in table form as a .pop file.

Since no compensation is required in the annual weighting, this part of the program run only consists of one batch process for weighting and two contiguous programs.

Program 4:
Program 4 first of all combines all four quarter files MZHR6-Q1 to -Q4, and then creates a file of all respondents of a year. Furthermore, a file is created for the Batch 4 process which only contains persons in private households and community households who were asked at their main residence. These are sorted by their household number, which serves as a cluster code in the weighting.

The program (as in Program 1 for the compensation) also calculates further reference values. For this, the standard weighting factors of the individual combinations of the characteristics are added up and output as a table in a .pop file.

Batch 4:
The Batch 4 process calculates the weighting factors for deep regionalised annual results. Administrative district, regional adjustment layer and sub-group, as well as sex, nationality (German/not German, as well as German/Turkish/EU25/not EU25), an age classification, a label of employees of the Federal Armed Forces, police or the Federal Border Guard and the employment status, serve as characteristics. The household number is used as a cluster code once again. The weighting model emerges from the combinations of the characteristics. In addition to the benchmark values calculated in Program 4, the population sizes per regional sub-group from the current population adjustment are also input into the weighting. They are also available in table form as a .pop file.

For weighting, the linearconsistent method with constant is used. The inclusion weight is ¼ of that in Batch 3 (quarterly weighting), i.e. ¼*(*N*/*n*)*compensation factor. Each cell of the model is to hold at least 10 observations. If this is not the case, the model is reduced to a higher-order combination of the characteristics. The interval for the correction weights ranges from 0.001 to 5, while the maximum number of iterations is once again 10. A maximum value of 50 is permitted for the inconsistency of the benchmark values, and the relative deviation of the weighted values from the benchmark values should not exceed 0.001.

To document the procedure, the batch process outputs a protocol file with the (possibly abridged) model used, the adjustments to the individual combinations of characteristics, as well as descriptive statistics on the correction and final weight.

Once Batch 4 has run, a file is created containing all persons who were asked at their main residence. The datasets now also contain the weighting factor for deep regionalised evaluations.

Program 5:
In Program 5, the factors of the deep regionalised annual weighting are applied to all respondents at their main and secondary residence. The weighting factors are once more added to the datasets as 4-digit string variables. Finally, an ASCII file "MZHR7.asc" including all respondents (i.e. private households and community household units, as well as main residence and secondary residence) is output, the datasets of which contain all compensation and weighting factors, both for quarterly and for annual results. The annual weighting is thus completed.

4. File sizes and running times

The size of the MZHR1.asc input files varies greatly from one Federal Land to another since the sample size, and hence also the number of datasets, fluctuates with the number of inhabitants. In Federal Länder with a high population size and roughly 15,000 datasets per quarter, the input file quickly reaches a size of 2,000 KB. With smaller Federal Länder comprising 670 datasets, by contrast, the file size of the input material is only 80 KB.

The benchmark values from the current population adjustment are relatively low, at 1 KB for the total population size, and 5 KB for the .pop file with the tables of the individual combinations of characteristics at the level of the regional adjustment layers. Depending on the size of the Federal Land, and hence on the number of regional adjustment layers, the tables contain 50 to 82 individual values in the .pop file.

The running time of the programs depends on the size of the input files: The quarterly weighting of a small Federal Land only takes one minute, whilst with a Federal Land that has many inhabitants, the running time is roughly eight minutes. The most time is taken in this case by the benchmark value and auxiliary variable calculation.

The input files for the deep regionalised annual weighting MZHR6-Q*x*.asc have a size of roughly 400 KB each with small Federal Länder, and 10,000 KB with high-population Länder. Depending on the number of regional sub-groups in a Federal Land, up to 54 figures are input as benchmark values. The program needs roughly twice as long to run as the quarterly weighting: Almost two minutes for smaller Länder and roughly 15 minutes for a large Federal Land.

5. Abbreviations

CHH	community household
N	number of the population total
n	number of respondent persons
PHH	private households
x	quarter number

6. References

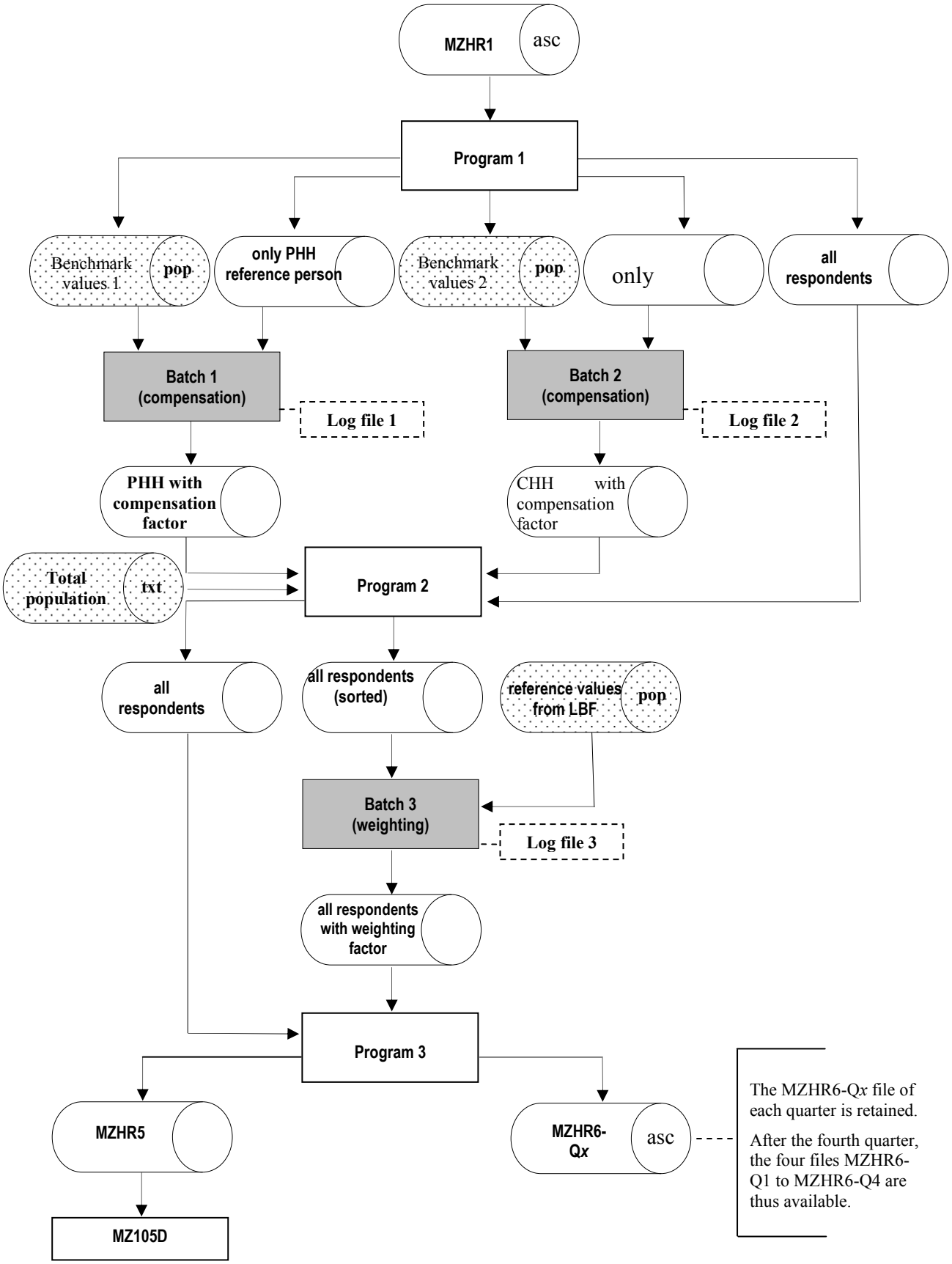
Afentakis, A. (2005): Hochrechnung der Mikrozensusergebnisse ab 2005 – Fachkonzept für die DV-technische Umsetzung. Statistisches Bundesamt Deutschland, Wiesbaden.

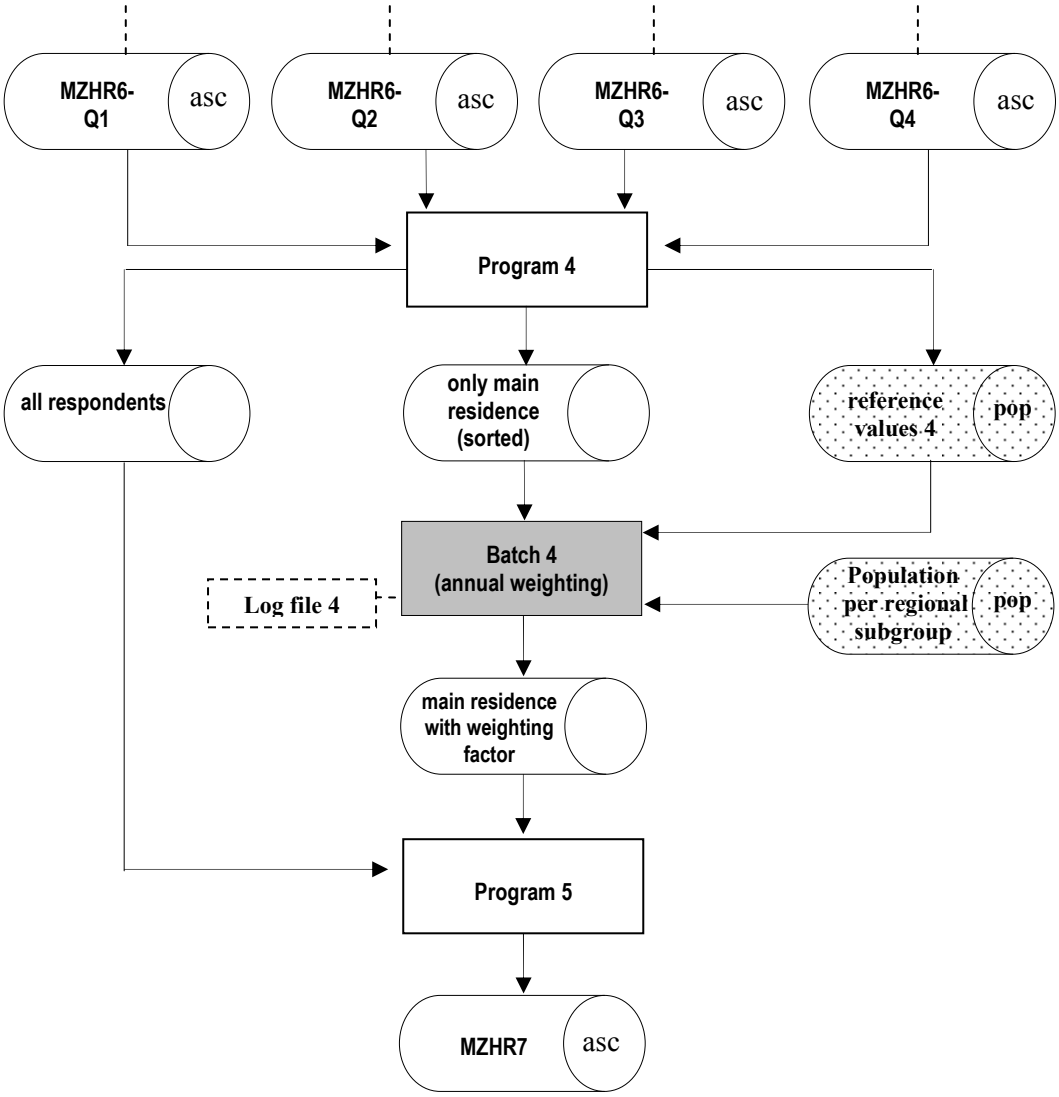
Afentakis, A. & Bihler, W. (2005): Das Hochrechnungsverfahren beim unterjährigen Mikrozensus ab 2005, Wirtschaft und Statistik 10/2005, Statistisches Bundesamt Deutschland, Wiesbaden.

Nieuwenbroek, N. & Boonstra, H.J. (2001): Bascula 4.0 Reference Manual. Statistics Netherlands, Heerlen.

Annex

A1: Data flow plan




MZHR6-Q1
asc
MZHR6-Q2
asc
MZHR6-Q3
asc
MZHR6-Q4
asc
Program 4
all respondents
only main residence (sorted)
reference values 4
pop
Batch 4 (annual weighting)
Log file 4
Population per regional subgroup
pop
main residence with weighting factor
Program 5
MZHR7
asc

A2: Example syntax for the batch processes

```
PROCESS Weighting 'ProcessName' [9]

USES
  MetaID 'MetaFileName' [9]

INPUTFILE SampleFileID: MetaID('MZHR4C',FileType) [9]

AUXFIELDS
  Res: INTEGER

WEIGHT WeightID (SampleFileID)
  VARIABLEFUNCTIONS
    STANDARD = EF5, EF32, EF23, EF100, EF101, EF102, EF103, EF104,
               EF106
    CLUSTERCODE = EF105        {does not apply to compensation (Batches 1+2)}
    INCLUSION = EF122
    CORRECTION = EF119
    OUTPUT = EF120
  ENDVARIABLEFUNCTIONS
  POPULATIONTABLES             {vary by model}
    POPULATIONTABLE
      TABLE = EF106
      POPULATIONFILE = 'MZHR3B.pop' [9]
    ENDPOPULATIONTABLE
    POPULATIONTABLE
      TABLE = EF103 * EF104
      POPULATIONFILE = 'MZHR3B.pop' [9]
    ENDPOPULATIONTABLE
    POPULATIONTABLE
      TABLE = EF23 * EF100 * EF101
      POPULATIONFILE = 'MZHR3B.pop' [9]
    ENDPOPULATIONTABLE
    POPULATIONTABLE
      TABLE = EF23 * EF102
      POPULATIONFILE = 'MZHR3B.pop' [9]
    ENDPOPULATIONTABLE
    POPULATIONTABLE
      TABLE = EF5 * EF23 * EF100
      POPULATIONFILE = 'MZHR3B.pop' [9]
    ENDPOPULATIONTABLE
    POPULATIONTABLE
      TABLE = EF13
      POPULATIONFILE = 'MZHR3B.pop' [9]
    ENDPOPULATIONTABLE
     OUTPOPULATIONFILE = 'MZHR3B.pop' [9]
  ENDPOPULATIONTABLES
  WEIGHTSETTINGS
    MODELTERMS = EF106
                 EF103 * EF104,
                 EF23 * EF100 * EF101,
                 EF23 * EF102,
                 EF5 * EF23 * EF100,
                 EF13
```

[9] Complete file path

```
    MINIMALCELLCOUNT = 10
    CONSTANT = YES
    WEIGHTINGMETHOD = LINEARCONSISTENT     {in Batches 1+2: LINEAR}
    BOUNDING
      LOWERBOUND = 0.001
      UPPERBOUND = 5                       {in Batches 1+2: UPPERBOUND = 3}
      MAXSTEP = 10      {A maximum of 10 iterations are possible in the batch process.}
    ENDBOUNDING
    WEIGHTFILE = 'Batch3.wga' 9
  ENDWEIGHTSETTINGS
  SETTINGS
    DIFFTOLERANCE = 50                     {in Batches 1+2: DIFFTOLERANCE = 1}
    ESTCHECKTOLERANCE = 0.001
  ENDSETTINGS
ENDWEIGHT

MANIPULATE
  Res := WeightID
```

Posters

User-friendly Web Surveys
Hayo Bethlehem (The Netherlands)

Using the Blaise Component Pack in the .NET environment
Rob Groeneveld
(Statistics Netherlands)

A Calculator DLL
Roberto Picha
(U.S. Census Bureau, USA)

Programming for Navigation in the Blaise System
Mark Pierzchala (Mathematica Policy Research, USA.)

The Selection of Strata in Nonresponse Adjustment
Barry Schouten
(Statistics Netherlands)

CAPI System at Central Statistical Bureau of Latvia
Norberts Talers & Palvels Onutrijevs
(Central Statistical Bureau of Latvia)

Multiple Researchers working at the Blaise Benchmark Services for the Disabled Act (WVG)
Carlo Vreugde & Mark Gremmen
(SGBO, The Netherlands)