



CHINTEX

Working Paper #20

Work-package 2

Date: 31 January 2004

Hartmut Minkel
(Statistisches Bundesamt)

Report on Data conversion methodology

CHINTEX - The Change from Input Harmonisation to Ex-post Harmonisation in National Samples of the European Community Household Panel – Implications on Data Quality

Financed by the European Commission under contract number IST-1999-11101

Content

1	SURVEY CONVERSION IN THE ECHP FRAMEWORK	4
1.1	SURVEY AND VARIABLE CONVERSION	4
1.2	SURVEY CONVERSION APPLIED TO THE ECHP	5
1.3	EXPERIENCES OF ECHP SURVEY CONVERSION	8
2	GENERATION OF A FRAMEWORK FOR CONVERSION OF SURVEYS.....	10
2.1	EXPLOITATION OF EXISTING INFORMATION OVERLAPS	10
2.2	STRATEGY OF USING EXPLOITATION METHODS.....	11
2.3	GENERAL FRAMEWORK FOR EXPLOITATION OF INFORMATION	13
3	DEMONSTRATION OF EMPIRICAL EXAMPLE ON STRUCTURAL PERSISTENCY	19
3.1	DESCRIPTION OF DATA SITUATION	19
3.2	DISCUSSION OF CITIZENSHIP PERSISTENCY	24
3.3	CONVERSION PROCESS	31
4	CONCLUSIONS.....	38
	REFERENCES	40

FIGURES

Figure 1: Survey conversion as part of ex-post harmonisation in the ECHP	6
Figure 2: Paths of survey conversion within the Data Production Process.....	7
Figure 3: Conversion of a non-harmonised value Y to Y*	12
Figure 4: General framework for the specification of panel data	14
Figure 5: Information overlap within a panel survey.....	15
Figure 6: Dimensions for cross-panel information transfer	17
Figure 7: Weighted shares for <i>Citizenship (Abbrev)</i> in ECHP and aggregated BHPS.....	24
Figure 8: Conditional transition probabilities for <i>Citizenship</i> given the subsequent value	30
Figure 9: Conditional transition probabilities for <i>Citizenship</i> given the previous value	31
Figure 10: Bias of BHPS <i>Citizenship</i> distributions compared to ECHP for 1996	33
Figure 11: Bias of BHPS <i>Citizenship</i> distributions compared to ECHP for 1995	35
Figure 12: Respondents without UK citizenship in BHPS and ECHP waves.....	36

Tables

Table 1: Information transfer for the variable <i>Citizenship</i>	19
Table 2: Observed frequencies for <i>Citizenship (Abbrev)</i>	22
Table 3: Weighted frequencies for <i>Citizenship (Abbrev)</i>	23
Table 4: Transitions between categories of <i>Citizenship</i> in ECHP from 1994 to 1995	25
Table 5: Transitions between categories of <i>Citizenship</i> in ECHP from 1995 to 1996	27
Table 6: Transitions between categories of <i>Citizenship</i> in ECHP from 1994 to 1996	28
Table 7: Transfer, adjustment and evaluation of <i>Citizenship</i> values from BHPS	32

1 Survey conversion in the ECHP framework

1.1 Survey and variable conversion

There are different notions about the meaning of data harmonisation and data conversion.¹ Within the CHINTEX project, the term *conversion* is understood as ex-post output harmonisation of micro-data. This interpretation highlights the two common and most important aspects of a data conversion:

- the issue is to produce micro-data, which is a value of a characteristic (variable) of a unit of observation such as a household, an individual or an enterprise.
- the value is to be harmonised, meaning it suffices given standards of comparability, but it is obtained from a non-harmonised data source. All efforts of harmonising are done ex-post, that is after designing the non-harmonised data source and after collecting its information.

Conversion can be seen in a strict sense – with respect only to a given variable of interest for which a harmonised value for each unit of observation of a survey is sought. We understand this as *variable conversion*. Variable conversion is our notion of the transformation of a non-harmonised data value to a harmonised value. The aspects in which the input value differs from the target value sought can be numerous: differences in the definitions of the underlying concepts or in the definitions of the variables, deviations in the scales of measurements and so on.² It shall be noted that variable conversion differs from the general problem of missing data insofar as we have a value at hand, bearing information very close to what we seek, but which is not sufficiently harmonised, meaning that it has from a statistical point of view a different distribution.

In a broader sense, variable conversion can be extended to apply to a set of variables. In the logical consequence, it can apply to the whole data base of a

¹ See report CHINTEX WP 1, chapter A.

² See report WP 1, Chapter D for types of problems of conversion.

survey. We refer to this as *survey conversion*: the generation of a harmonised survey data set on the basis of a non-harmonised data set which will produce estimates similar to those of an input harmonised survey. But survey conversion in our sense will comprise more issues than the amount of variables. It also involves aspects of a survey apart from the variables' distribution, namely issues of the treatment of non-response (item and unit non-response), weighting and other aspects of estimation. Survey conversion aims at the ex-post production of a harmonised micro data set, that is fully comparable to equivalent micro data sets obtained from input-harmonised surveys in terms of the distribution of its variables of analysis and in terms of estimation results. It therefore also includes the harmonisation of issues related to estimation.

1.2 *Survey conversion applied to the ECHP*

Three countries, Germany, Luxembourg and United Kingdom, obtain the data they are required to deliver for use in the data base of the ECHP via survey conversion.³ These countries take a national academic panel survey which provides information similar to that of the ECHP as the basis for conversion into the ECHP standards and formats. The objective is to clone the ECHP data as complete in the number of variables and as close with respect to concepts as possible. This amounts to the task of producing yearly about 640 variables of the ECHP cross-sectional data base.

With respect to the format of the target data base this survey conversion shows certain special aspects, which are important to note. The countries perform the survey conversion of their national panel into the format of the so-called Production data base of the ECHP (PDB).⁴ The PDB has a 1:1 correspondence to the blue-print questionnaire. Each question is assigned to exactly one PDB-variable. The PDB is the starting point of Eurostat's complex production process which yields the data

³ See report WP 1, Chapter B and C for details.

⁴ Eurostat, ECHP – 1994, Wave 1 variable list, Luxembourg February 1994 (DOC PAN 15/1994); Eurostat, ECHP – 1995, Wave 2 variable list, Luxembourg February 1995 (DOC PAN 30/1995); Eurostat, ECHP – 1996, Wave 3 variable list, Luxembourg March 1996 (DOC PAN 65/1996). Cf. Eurostat, ECHP UDB manual, Waves 1 to 5, survey years 1994 to 1998, Luxembourg December 2001 (DOC PAN 168/2001-12).

base for cross-country analysis, the User's Data Base (UDB).⁵ This process is comprised of several steps:

- editing of micro-data,
- imputation of missing data,⁶
- derivation of variables of analysis, e.g. aggregated income variables,
- computation of sample weights, including estimation of pattern of attrition of sample units and calibration to margins obtained from external data sources.⁷

Each step has non-negligible effects on survey estimates. Figure 1 displays the process of production of ECHP estimates.

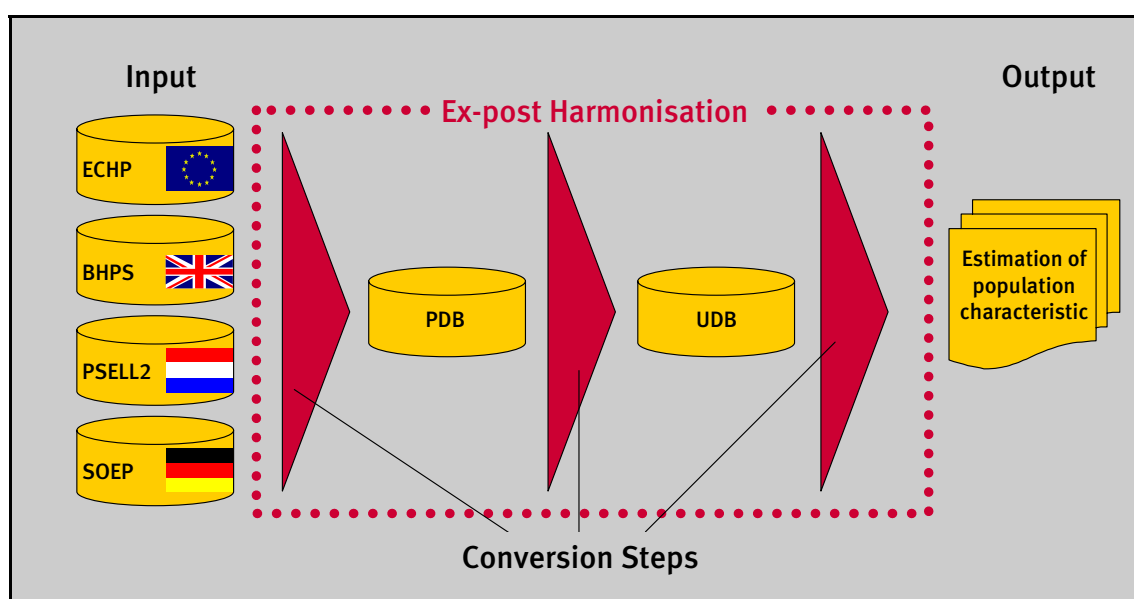


Figure 1: Survey conversion as part of ex-post harmonisation in the ECHP

⁵ Cf. Eurostat, ECHP UDB, construction of variables, from ECHP questions to UDB variables, Luxembourg May 2001 (DOC PAN 167/2001) and Eurostat, ECHP UDB description of variables, codebook and differences between countries and waves, Luxembourg December 2001 (DOC PAN 166/2001-12).

⁶ Cf. Eurostat, construction of weights in the ECHP, Luxembourg 2002 (DOC PAN 165/2002-12).

⁷ Cf. Eurostat, imputation of income in the ECHP, Luxembourg December 2002 (DOC PAN 164/2002-12). See report CHINTEX WP 7.

In principle, there are three phases where a higher degree of harmonisation can be reached: in the production of the PDB, in the transformation to the UDB and in the phase of estimation of population statistics. Measures of survey conversion, i.e. conversion steps on the way to harmonised estimates, can be built in to all phases. Figure 2 unfolds the framework in greater detail and provides a system of possible methods. It displays the different paths we can differentiate in general. Each path stands for a set of methods we can apply in survey conversion.

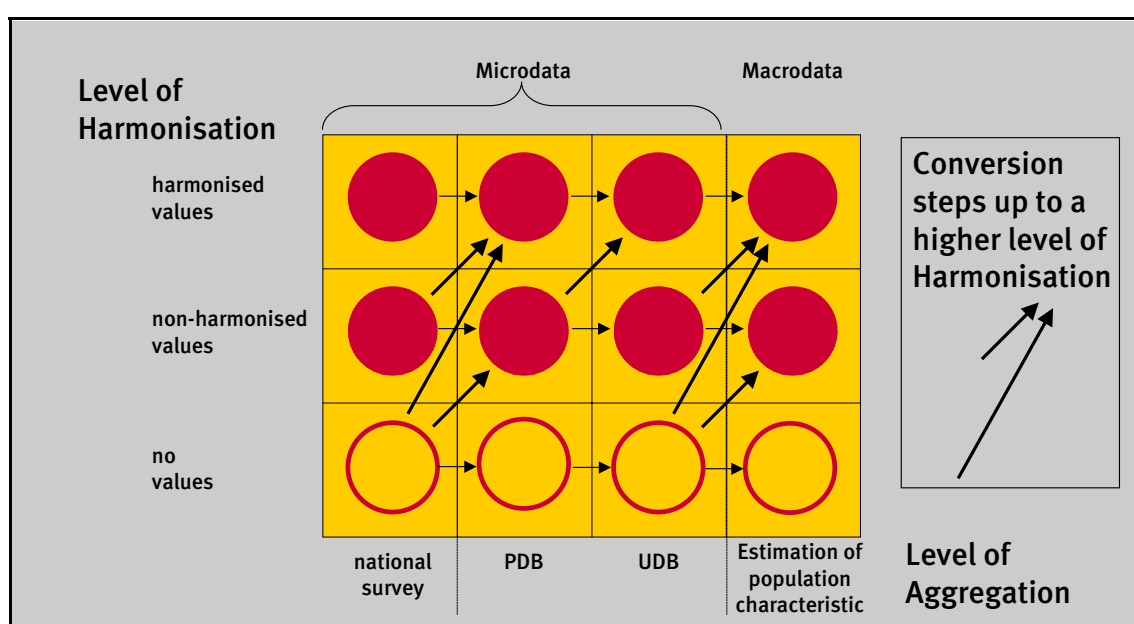


Figure 2: Paths of survey conversion within the Data Production Process

The table shows the two dimensions of the production process of ECHP estimates during a conversion. The horizontal dimension reflects the ECHP production process starting from a low level of aggregation of information (PDB micro-data) and leading to a high level of aggregation (estimates). Since this is general to ECHP data, the path of countries which follow input harmonisation is part of the figure. These countries move along the top arrows in the row of harmonised values. The vertical dimension is applicable for the survey conversion only. It reflects the general levels of harmonisation reached during survey conversion. The bold, sloping arrows display the different paths which can be taken during the steps of

survey conversion. They subdivide into two groups: the left group which ends at a higher level of harmonised micro-data, and the right group which ends at a higher level of harmonised estimates (the macro-data).

The latter group comprises measures of survey conversion at the stage of estimation, such as bias correction. But also weighting can be seen as part of this group. Let us assume, for instance, the weights produced during the production of the UDB suffered from severe deviations from the common standards. Say, some crucial variables could not be taken into account in the attrition analyses for the converted PDB, since they were simply not available. Then this deviation might yield less harmonised weights than the use of the original weights of the national non-harmonised survey. The use of these “external” weights might be the better choice with respect to harmonised estimates and therefore it is a useful method of survey conversion.

The first group of conversion paths in the area of micro-data can also be split into two groups. Movements from no values to a harmonised or non-harmonised value reflect the standard statistical approach of imputation of missing values. In contrast, methods of variable conversion are displayed by the paths from non-harmonised values to harmonised values. So both, methods of imputation and of variable conversion are tools for survey conversion but they can be differentiated with respect to the extent in which they use non-harmonised existing information.

1.3 Experiences of ECHP survey conversion

The experiences made during the survey conversion of the ECHP have been summarised within work-package 1 of the CHINTEX project.⁸ This descriptive analysis shows, that across countries common problems of variable conversion and common approaches used to overcome these problems exist. Experience shows that the set of methods of survey conversion actually made use of, is rather limited:

⁸ See report WP 1, Chapters D and E.

- Most methods applied are located at the conversion step of production of PDB variables. This is in line with the objective to supply fully harmonised data at PDB level, which was agreed in the contracts with Eurostat. But this approach limits the set of possible methods from the beginning.
- Hence, there are only few cases with methods located at the conversion step of production of UDB variables or at the conversion step of estimation of population statistics.
- The particular methods applied at the conversion step of production of PDB variables are of the type of variable conversion. They are mainly based on ad-hoc decisions about recoding, combining or collapsing of existing, non-harmonised variables. There is a tendency to change values from the source questionnaires as slightly as possible when taking them over to the target data set.
- Estimation techniques, which make use of knowledge about empirical distributions and statistical dependencies to predict harmonised values are almost never used. This applies to methods of variable conversion but in particular to imputation methods.
- Information external to the source survey, e.g. information on distributions of target or explanatory variables, was never used.

Therefore it was concluded, that there is clear evidence for the need to enlarge and systematise the available toolkit and to show how to use it.

2 Generation of a framework for conversion of surveys

2.1 *Exploitation of existing information overlaps*

In general it is the goal of a survey to provide information about unknown population characteristics, e.g. household income, in form of a set of data. This information is gained with the help of a survey specific observation instrument like the ECHP which covers the necessary attributes of the population elements and defines the variables and also the measures to assign values to these variables. The aim of converting survey data in view of ex-post harmonisation is utilising existing information in order to generate data in such a way like it was available if data would have been gained by an input harmonised survey. Since, in the non-trivial case, the state of existing and desired information are not equal, converting corresponds to modifying available data in order to exploit the necessary information. When it comes to specific attributes, conversion either has to be applied to existing but differently defined or measured variables or, depending on the above-mentioned level of harmonisation, variables not asked at all have to be imputed. In latter cases there is no variable which resembles exactly the desired information. We can only attain it, if we exploit existing information overlaps, i.e. we have to assume that the information additionally needed can be found elsewhere. In principle, sources for such information can be exclusively parts of the survey to be converted or exploitation can also use information beyond this, e.g. from another survey. But regardless of the source referring to, every gained information is experiencing a modifying extraction out of its original context. In the first case additional information in explaining variables within the survey to be converted is transformed into the desired variable – we call it transformation of an attribute. The second basic possibility for exploitation of additional information is the use of external survey data. Assuming that the same variable used in another survey enables us to draw some conclusions about information hidden in correlated variables and transfer this information to the survey to be converted,

leads to the second possibility for gaining information: the transfer of an attribute. Both basic procedures can be defined as follows:

Definition 1: Transformation of an attribute

Projection of one assignment on another assignment in the same sample.

Definition 2: Transfer of an attribute

Projection of an assignment from sample A on sample B.

2.2 *Strategy of using exploitation methods*

Both, transformation and transfer of information, cause a modification of already existing information. In the first case it leads to a modification of the observation instrument, i.e. the transformation of values related to the variable to be converted. Very often the function by which values are modified will base on the redundancy of information within the range of variables on the one hand and the similarity of the population elements on the other hand. A transformation function is exploiting this information overlap.

The second modification of already existing information is its transfer from one sample context to another one, assuming that the way the assignments work in the different samples and also the representativeness of the population elements, respectively the sampling design, are transferable ⁹. If these assumptions hold, data can be modified in such way that it can be used as if it was observed in the sample context of the survey to be converted.

Applying both methods alternately, we end up with a concatenation of two or even more samples using one or more variables to explain the target variable Y_A^* . In the scheme of Figure 3 a conversion of a single variable Y_A in sample A is shown, using information about a transformation function f_B from sample B. A possible statistical method for this transformation could be item response conversion ¹⁰. This method

⁹ For the general conditions for an information transfer between different samples see Gabler (1997)

¹⁰ For item response conversion see TNO report 2001.097: Response conversion: A new technology for comparing existing health information, Leiden, 2001

would imply that all information necessary for transformation could be extracted from a single variable and no additional explaining information was available in the data sets of the samples A and B, i.e. Y^*_A and Y^*_B are conditional independent of each, X_A and X_B , given the explaining variables Y_A and Y_B . Likewise this assumption for transformation would be both, necessary and sufficient, if Y_A and Y_B were no single variables but vectors.

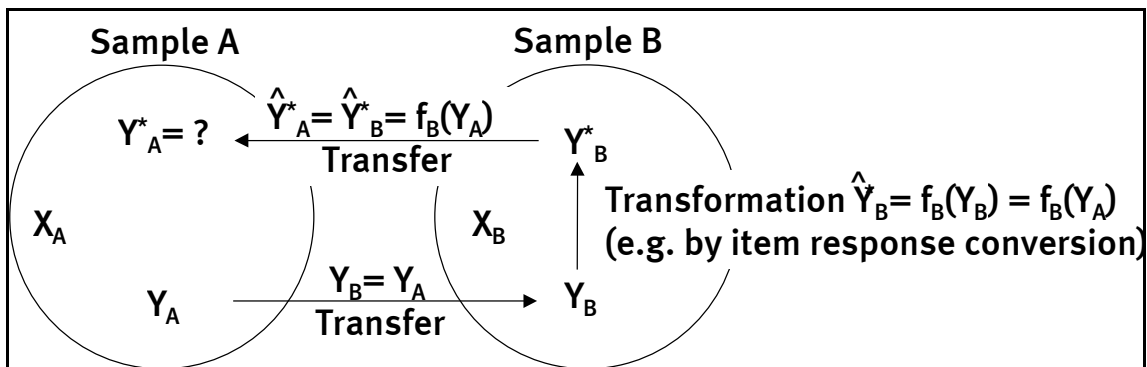


Figure 3: Conversion of a non-harmonised value Y to Y^*

With regard to the observation instruments applied to the different samples it is important whether these are the same or different, equipped with differing variables. If they are the same, the whole conversion process seems to be very close to a standard *imputation problem* where those population elements with an existing Y^* are allocated to sample B and those without Y^* belong to sample A. Things are more complicated if X_A and X_B do not cover the same variables. In this case we have two samples with two partially different sets of variables. If Y^*_B can be explained by the joint vector of variables Y_B and this information is transferred to sample A, we end up with a fusion of Y^* with X_A . Since X_A and X_B are at least partially different this combination of variables is new and the conversion procedure seems to be similar to the strategy applied for *data fusion*, just as their prerequisites are. For the scheme mentioned-above this means that only that information will be available for the construction of Y^* in sample A which could be extracted from the data gained with the help of the observation instrument applied on sample B. If Y^*_B

can not be fully explained by the joint set of variables Y , then it is possible that further information for explanation of Y^* can be found in those variables of X_A which are not covered by X_B . Therefore the correlation between Y^*_A and X_A is not entirely transferable from sample B. The whole conversion procedure can only lead to results we also would have obtained if Y^*_A could have been observed, if Y^*_A is conditional independent from X_A , given the joint explaining variables Y_A ¹¹.

2.3 General framework for exploitation of information

A framework for exploitation of information has to be built up on the basis of a general specification of panel data. With the help of a systematisation of the aspects specifying the data of a panel survey it can be shown, from which assumptions further conversion methods are always starting from. These aspects also condense those problems mentioned in the results of CHINTEX Work Package 1, which occur in the conversion process and influence survey bias.¹² Above all Figure 4 displays the three fundamental dimensions of panel data specification. The first dimension is describing that the population a survey is focussing on has to be defined by the function $d_A(\Omega)$ which separates the population of interest Ω_A from the whole space of all possible population elements Ω . In a second step a sampling function $s_A(\Omega_A)$ is producing a sample with n_A cases of population elements, the panel sample. The specification process for the second dimension, the panel variables, is similar. Within the space of all possible population element characteristics concepts serve as functions which assign attributes to selected characteristics. Measurement functions specify aspects like routing, wording and scales and end up with a set of variables. In panel surveys we have a third dimension resembled by the waves which are a function of specific follow-up rules.

¹¹ For detailed discussion of the conditional independence assumption with regard to data fusion see Rässler and Fleischer (1997)

¹² See report on compiled Information, CHINTEX Work Package 1, Chapter D

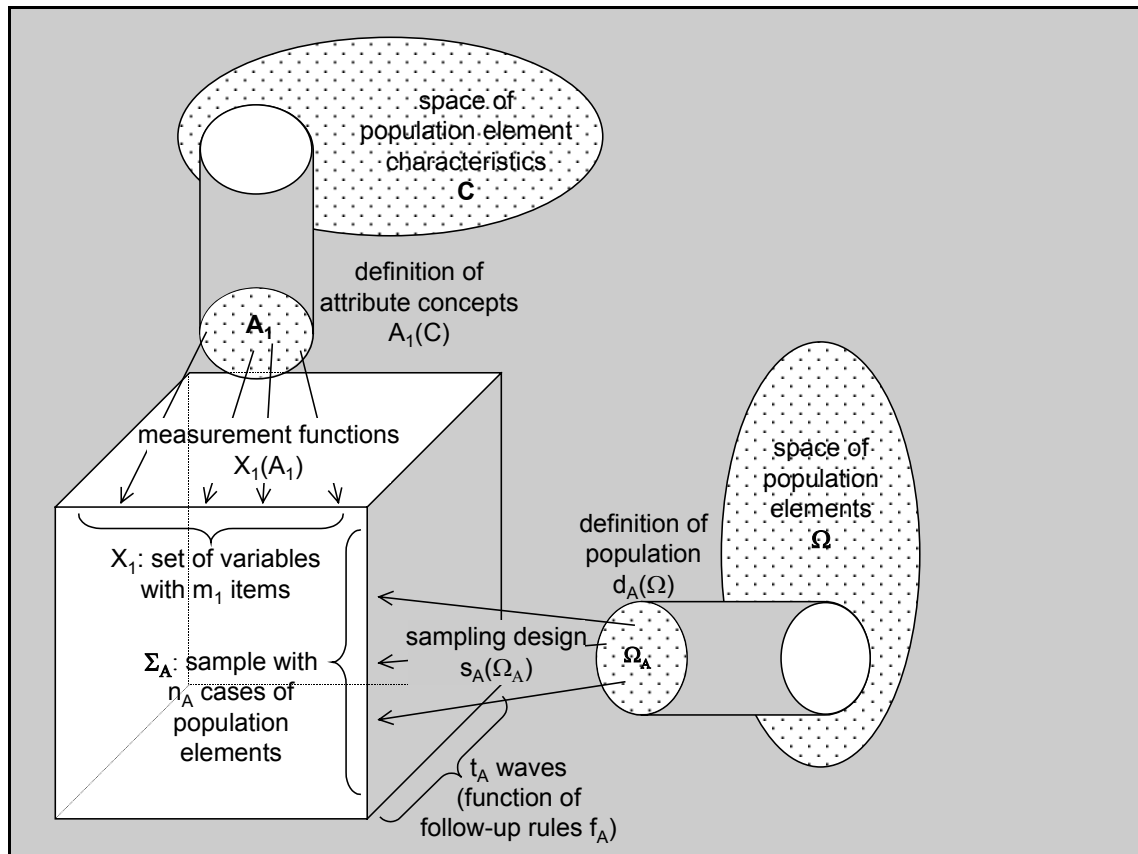


Figure 4: General framework for the specification of panel data

These three dimensions of data specification are also useful to systemise the occurrence of information overlaps within a panel study. In every dimension potentially redundant information does exist and can be exploited. Figure 5 displays that an overlap can be caused at first by *similarity of cases*. This means that within the sample there are elements whose characteristics are very similar or even pretty much the same. If for instance two persons had exactly the same values assigned to a set of variables, the second person would be completely redundant and knowledge only about the first person would be sufficient. In general it can be assumed that the more similar population elements are the more information is overlapping within the population and thus can be exploited from only a part of the sample. If on the other hand population elements have exactly contrary values assigned to the variables, information about attributes of a subset does not help very much with regard to the construction of the other cases.

In the second dimension, extraction of information is based on the *redundancy of items*. In contrast to the similarity of cases, here not elements of population are redundant but those assignments which are applied on the population reflecting its elements' characteristics. Information is overlapping across a group of correlated variables. In such a case values for a certain variable can be exploited from a set of some other variables which, as a whole, ideally fully explain the variable of interest. For a special instance, information can be extracted from a single variable, i.e. attribute concepts are pretty much the same but measurement functions convert information in a different way. Therefore only a single variable has to be converted and the result is variable conversion which was delimited from overall survey conversion in chapter 1.1.

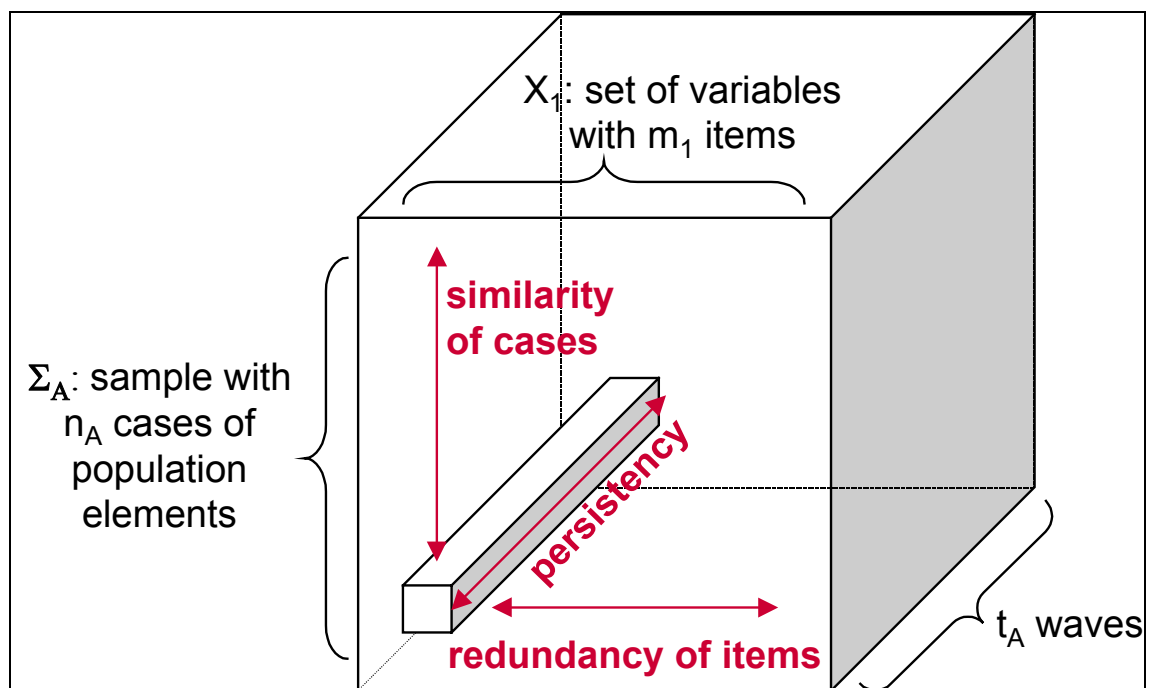


Figure 5: Information overlap within a panel survey

The third dimension is considering the role of time within a panel survey – and in this model especially in contrast to an unique survey. In a panel survey for every single element of the population sample variables are available in the form of time

series covering a number of waves. Depending on the *persistence* of a variable in general there will exist an overlap of information within such a time series. However there are some variables with more and some with less persistence. An absolute persistent variable would be for instance “country of birth”, certainly a little bit less stable is “country of citizenship”. On the other hand a variable like “*current participation in training/courses*” will not be very persistent and for a specific case such values are hardly transferable from one wave to the other.

Corresponding to the information overlap within a single panel survey, we can also use the dimensions mentioned-above for systemisation of information transfer between different surveys. Similarity of cases, redundancy of items and persistence are phenomena which not only occur within a certain panel survey but also in relation to some others. In Figure 6 a framework for a cross-panel information transfer is presented. In correspondence with Figure 5, data from additional panels can also be inserted in a three-dimensional system. In the same way as information can be exploited since information is overlapping across population elements or subsets which are similar with respect to the values assigned to these cases, it is also possible to utilise the similarity of cases in two differently defined populations Ω_A and Ω_B applying the same observation instruments. If we are able to transfer information about correlation of data due to similar structural reasons for this correlation, we call this *structural similarity of population*. But besides definition of population, data sets for each population are also results from the sampling design function. In the scheme shown below this assignment is symbolised by the reduction of the outer (transparent) cubes to the smaller inner ones. Thus, if information transfers use the assumption that both populations have similar cases, this assumption has to be still valid after applying the specific sampling functions.

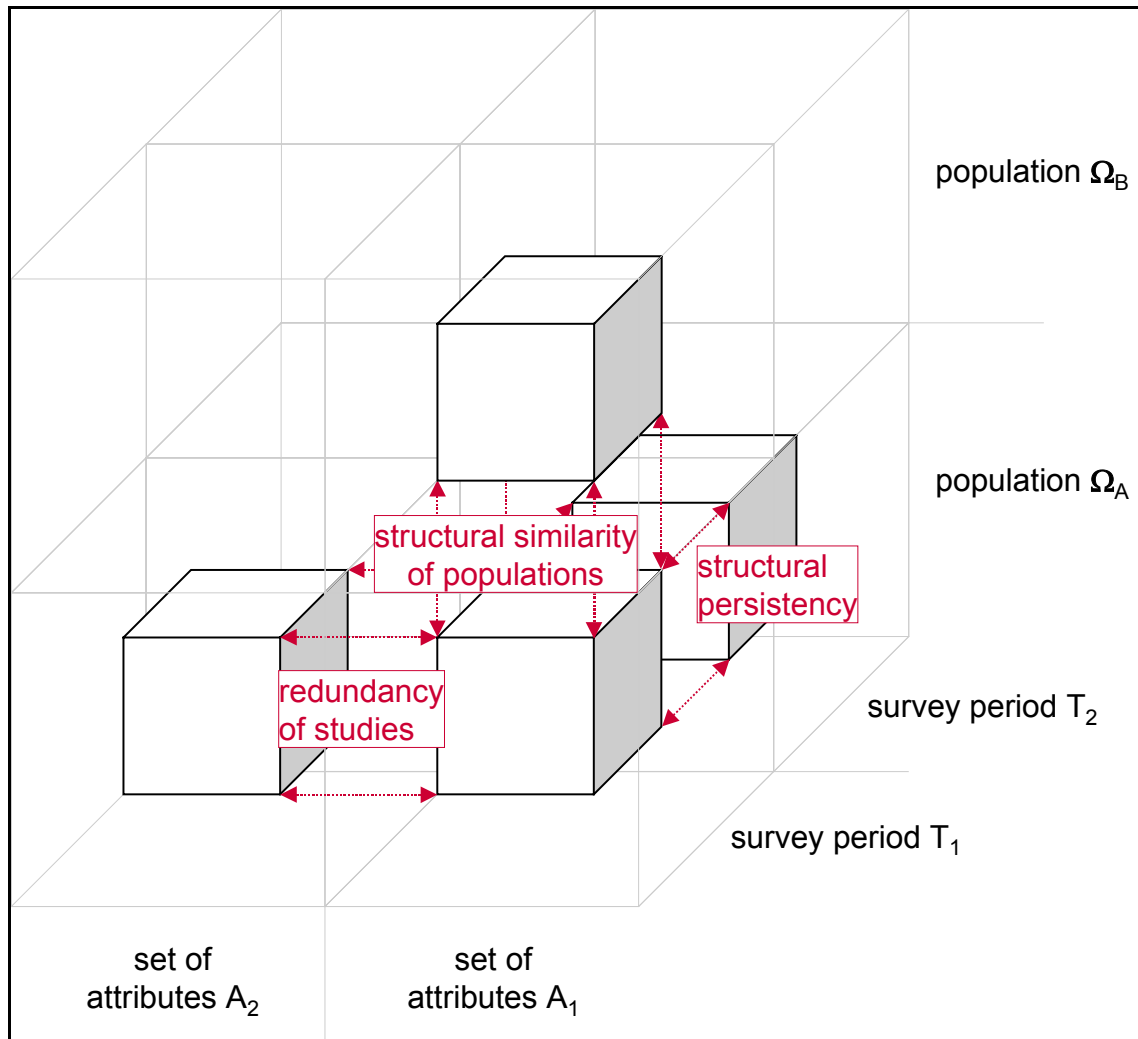


Figure 6: Dimensions for cross-panel information transfer

The second way to extract information is having two different observation instruments (studies) applied on the same population, given that the sampling design, i.e. the assignment of the population, is known for both of them. Here information overlap is dependent on the *redundancy of studies*, i.e. within the observation instruments there must be a set of correlated variables which exists in both surveys. By extracting the knowledge about correlation between these variables a key for transformation of information is transferred from one data set to the other. Thus this kind of information transfer is based on having the same information observed twice in the same population. Again, the reduction of the (transparent) outer cube to the inner one symbolises the application of specific

measurement functions to the set of attributes. If information about attributes shall be transferred, a bias due to different measurement functions has to be avoided and therefore these must be known and have to be made comparable, e.g. scales must be converted ¹³.

The third basis for exploiting information is to transfer correlation between variables over time assuming that the structure which causes such dependencies is persistent. In contrast to the persistency mentioned in Figure 5 it is not necessary to have continuous data, i.e. time series, for the individual cases. Ideally, if the other two dimensions stay unchanged, there is need for survey data observed with the same observation instrument in the same population, just at another point in time. This was the case for instance in a panel survey with rotation group design. Over time several samples of the same population are created and the same questionnaires are used. If the population-specific explanation for a variable to be converted does not change over time, information about this explanation is transferable from a latter survey to a previous one.

¹³ For item response conversion see TNO report 2001.097: Response conversion: A new technology for comparing existing health information, Leiden, 2001

3 Demonstration of empirical example on Structural Persistency

3.1 Description of data situation

As an empirical example for an application of the framework for conversion methodology, item persistency was chosen in order to show difficulties accompanying the conversion process on the basis of the probably easiest information transfer dimension. For variable transfer across the waves of a panel, resulting bias is supposed to be less distinct than for transfer between different surveys or populations. It can be assumed that influence by temporal effects within a panel is smaller than by variations caused by different sampling designs or measurement functions. Therefore findings of this chapter are even more relevant for those latter ways of information transfer.

CHARACTERISTICS	SET OF ATTRIBUTES (STUDY)	POPULATION	SURVEY PERIOD
	SET OF VARIABLES	SAMPLE	FOLLOW-UP RULES
Structural Persistency			
Citizenship	BHPS	UK (without Northern Ireland)	1997 (09/97 – 04/98)
	Present Citizenship: 1 st mention	BHPSw7	Temporary sample members who have a child with an original sample member but no longer reside with him/her will be followed
	BHPS	UK (without Northern Ireland)	1996 (09/96 – 04/97)
		BHPSw6	
	UK-ECHP	UK	1996 (01/96 – 12/96)
	Present Citizenship (Abbreviated)	ECHPw3	

Table 1: Information transfer for the variable *Citizenship*

Citizenship is defined in the course of Eurostat’s “harmonisation of core variables”¹⁴ “as the particular legal bond between an individual and his/her state acquired by birth or naturalisation, whether by declaration, option, marriage or other means according to national legislation”. The importance of this variable is founded in several legal purposes. Eurostat census information and Eurostat’s migration statistics are reference sources for information on this variable but it is also recorded in the ECHP. Countries should be coded according to ISO 3166-1.¹⁵ Considering the Census recommendations, persons with dual or multiple citizenship should declare all but for the purposes of household surveys, this seems unnecessary. Therefore, only the citizenship first mentioned is used as an example for conversion.

A problem with citizenship is that it can be changed, easily in some countries like the Nordic and the Netherlands, more difficult in Germany and Italy and others. This makes it more difficult to produce comparable data.

In the ECHP questionnaire citizenship is asked and coded by following questions and PDB variables:

Is (one of) your present citizenship(s) the citizenship of the country where the survey is conducted? (P01321)

Other citizenship, please specify. Code from Country List (see Annex III). (P01323)

Second citizenship, please specify. Code from Country List (see Annex III). (P01324)

Up to two citizenships are recorded in detail according to the coding from the annexed country list. For the ECHP UDB the variable “*citizenship (abbreviated)*” (PM008) is constructed by recoding of the PDB variable. Without missing, four categories are possible for this :

- *nationals (i.e. UK)*
- *another EU citizenship*
- *other citizenship (Extra-EU)*
- *not national, but citizenship unknown.*

¹⁴ see Eurostat: Harmonisation of core variables, (Doc. Eurostat/E0/00/DSS/2/6/EN)

The sub-sample of the ECHP's third wave covered the whole population of UK in the year 1996. According to the follow-up rules temporary sample members were not followed anymore if they did not longer reside with an original sample member.

In the BHPS questions about citizenship were introduced with the seventh wave i.e. data are available only for the years 1997 and after:

What is your present citizenship? If you have a dual citizenship, please tell me both.

Two citizenships can be given by the respondent. Variables are constructed with the help of a BHPS specific coding frame.

At the beginning the BHPS sample was covering the British population with domestic residence in England, Wales or Scotland south of the Caledonian Canal. Northern Ireland was not considered until 1997. Field work for the 1997 wave was done during the period from September 1997 till April 1998. Temporary sample members who have a child with an original sample member but no longer reside with him/her were followed.

In order to be able to compare both coding frames, the one from ECHP and the one from BHPS, at first the BHPS coding frame was aggregated to the ECHP UDB variable "*Citizenship (abbreviated)*". For respondents in BHPS who were interviewed by proxy or telephone, citizenship was still not asked in the years 1997 to 1999. These cases were weighted zero. The large rise of the number of respondents in BHPS from 1998 to 1999 is due to the integration of two additional samples in Scotland and Wales (see Table 2).

After the variable "*citizenship*" was introduced into BHPS in 1997, in the following waves this variable was only asked if the respondent was never interviewed before. Otherwise the value was taken from the previous wave. Therefore no transition between the different states of citizenship can be observed in BHPS. Neither changing values due to a real change in citizenship, nor the change from or to wrong or incomplete answers (e.g. "*not national, but citizenship unknown*" or "*missing*"). In contrast, the variable was observed in a different way in ECHP. In this

¹⁵ see: International Standard, ISO 3166-1: 1997: Codes for the representation of Names of Countries

panel citizenship was asked every year again. On the one hand, asking such quite persistent questions in every wave means a higher burden for respondent and interviewer than assuming that these values just stay the same. On the other hand only in ECHP it is possible to record real changes of citizenship or to give respondents the opportunity to revise their previous answers what can be noticed as a positive panel effect. Therefore, up to this point answers of both panel surveys are, due to the different methods of routing, only partially comparable.

		ECHP			BHPS		
		1994	1995	1996	1997	1998	1999
UK	%	96.93	97.73	98.08	97.63	97.63	98.01
	N	10194	8196	6807	10571	10307	14813
another EU citizenship	%	1.15	1.16	0.99	1.21	1.15	0.97
	N	121	97	69	131	121	147
other citizenship (Extra-EU)	%	1.84	1.01	0.86	0.90	0.92	0.69
	N	194	85	60	97	97	105
not national, but citizenship unknown	%	0.01	0.02	0.03	0.17	0.17	0.15
	N	1	2	2	18	18	22
missing	%	0.07	0.07	0.03	0.10	0.13	0.18
	N	7	6	2	11	14	27
Total	N	10517	8386	6940	10828	10557	15114

Table 2: Observed frequencies for *Citizenship (Abbrev)*

For examination of citizenship in the two panel surveys data was weighted (see Table 3). In ECHP the normalised base weight for interviewed persons (*PG003*) was used. For BHPS analysis the Individual Respondent Weight (*wXRWGHT*) was taken for the waves up to year 1996. Since the ECHP sub-sample which entered the survey in 1997 had to be considered, data in the years from 1997 to 1999 were weighted by the new introduced respondent weight *wXRWGHTTE*. The new extension samples from Scotland and Wales in 1999 were not considered in computations (i.e. weighted zero), thus for this year unweighted and weighted values in Table 2 and Table 3 have a large difference.

		ECHP			BHPS		
		1994	1995	1996	1997	1998	1999
UK	%	96.90	97.68	98.02	97.52	97.49	97.77
	N	10190.5	8191.6	6802.7	10559.2	10291.6	10064.6
another EU citizenship	%	1.12	1.06	1.09	1.24	1.24	1.11
	N	117.6	89.3	75.3	134.5	130.4	114.1
other citizenship (Extra-EU)	%	1.91	1.13	0.87	0.99	0.95	0.79
	N	201.3	94.4	60.3	106.8	100.4	81.6
not national, but citizenship unknown	%	0.01	0.04	0.02	0.16	0.20	0.17
	N	1.3	3.1	1.4	17.3	20.8	17.8
missing	%	0.05	0.09	0.00	0.09	0.13	0.16
	N	5.4	7.7	0.3	10.2	13.3	16.6
Total	N	10516.1	8386.0	6940.0	10828.0	10556.5	10294.7

Table 3: Weighted frequencies for *Citizenship (Abbrev)*

Figure 7 shows the weighted frequencies of the three waves of interest of each, ECHP and BHPS. The weighted shares of the five possible categories for citizenship have very different scales and therefore these are displayed logarithmically. Thus deviations between the several waves of the same relative magnitude also make the same difference of columns heights for a certain category of citizenship. In general, percentages of the different waves and panels are very close in their values. Especially for “UK” the absolute range is only 1.12% (i.e. relatively also about 1%) and for “another EU citizenship” 0.15% (relative 13%). The recording of these two categories is obviously very similar by these two different panels. In contrast to “another EU citizenship” the category “other citizenship (Extra-EU)”, although from similar size, shows a very strong decrease in the ECHP data by more than a half for the three years 1994 to 1996 of its UK subsample. Whereas the BHPS data from 1997 to 1999, at this time already the seventh to ninth wave, are pretty much stable in their shares. Their size is close to the last ECHP data from 1996. For the remaining two categories “not national, but citizenship unknown” and “missing” there are even larger deviations between the ECHP waves. Here scale of shares is of one to two decimal powers smaller than for the previous categories, but

again, BHPS data is more stable, whereas ECHP shares show a large inconsistency. For the not quite precise phrase “*not national, but citizenship unknown*” it has to be assumed that in general wording and routing lead to different results in the two panels. Therefore shares in BHPS are always a lot higher than in ECHP.

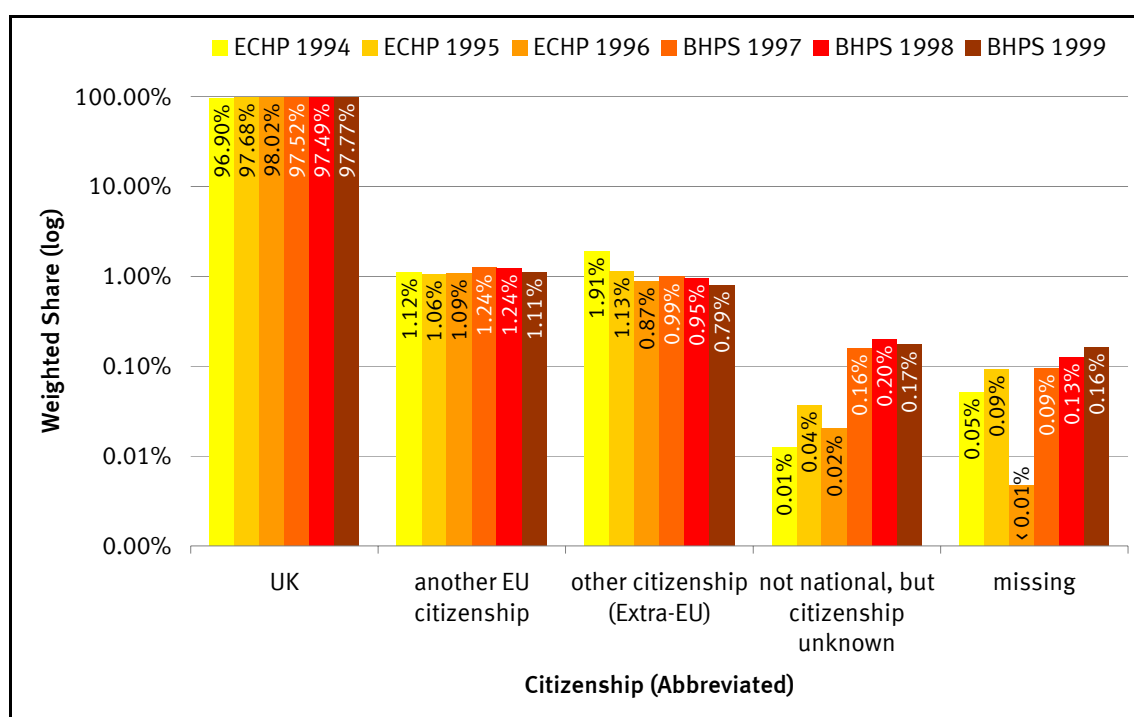


Figure 7: Weighted shares for *Citizenship (Abbrev)* in ECHP and aggregated BHPS

3.2 Discussion of citizenship persistency

In order to be able to convert BHPS data to ECHP as if data was observed with ECHP questionnaires, transitions between the ECHP values across the waves has to be made transparent. These transitions may also explain why some values are very inconsistent across the ECHP waves. The transitions from the first wave (94) to the second (95) are displayed in Table 4. First matter which is remarkable is the high attrition rate after the initial wave from ECHP. In total this is about 24%. Across the several categories of citizenship this rate is varying a lot. Of course, for the dominating category “*UK*” this rate is close to the average. But from those who answered “*another EU citizenship*” almost 37% left the panel. In addition, almost

10% switched their answer to “UK” in the following year. Since there are even more transitions in the opposite direction it has to be left open whether this percentage reflects real changes in citizenship or some of these respondents just gave wrong answers.

Citizenship 94 (Abbrev)		Citizenship 95 (Abbrev)					
		UK	another EU citizenship	other citizenship (Extra-EU)	not national, but citizenship unknown	missing	attrition
UK	N	7759.4	19.3	5.6	.	5.3	2400.9
	% (1994)	76.14	0.19	0.05	.	0.05	23.56
	% (1995)	99.49	21.37	6.97	.	100.00	94.56
another EU citizenship	N	11.5	60.4	2.3	.	.	43.4
	% (1994)	9.79	51.32	1.94	.	.	36.94
	% (1995)	0.15	66.89	2.85	.	.	1.71
other citizenship (Extra-EU)	N	26.7	10.6	70.2	1.9	.	92.0
	% (1994)	13.26	5.26	34.86	0.93	.	45.69
	% (1995)	0.34	11.74	87.48	100.00	.	3.62
not national, but citizenship unknown	N	.	.	1.3	.	.	.
	% (1994)	.	.	100.00	.	.	.
	% (1995)	.	.	1.63	.	.	.
missing	N	1.9	.	0.9	.	.	2.7
	% (1994)	34.12	.	15.66	.	.	50.23
	% (1995)	0.02	.	1.06	.	.	0.11
Total	N	7799.4	90.2	80.2	1.9	5.3	2539.1
	% (1995)	74.17	0.86	0.76	0.02	0.05	24.14

Table 4: Transitions between categories of *Citizenship* in ECHP from 1994 to 1995

However, in total only a bit more than half of the respondents who answered in 1994 with *“another EU citizenship”* kept up with this answer in 1995. For *“other citizenships (Extra-EU)”* this share is even only about one third. Here we observe an attrition rate which is almost one half (46%). And since only a very few respondents switched to this answer in the second wave whereas one out of seven left this group towards an UK citizenship, this leads to an explanation for the strong decrease of this categories share. With regard to the remaining two categories *“not national, but citizenship unknown”* and *“missing”* the transition table underlines the low persistency of such answers. There are anyway only a few respondents and their answering behaviour does not stay the same from one year to the other.

In Table 5 transition frequencies from the second to the third ECHP wave are displayed. These are very similar to those in Table 4. Overall attrition rate is now a little bit higher than 20% and therefore lower than from the first wave to the second. Attrition rates for the other two citizenship categories are much higher but they are also lower than the year before. For all these three citizenship categories persistency of its values is increasing. Whereas about 76% of the *“UK”*-respondents from 1994 kept their answer in 1995, one year later almost 80% did not change their statement. For *“another EU citizenship”* this share increased from 51% to 57% and for *“other citizenship (Extra-EU)”* from only 35% up to 55%. These changes underline the hypothesis from Work Package 6, that attrition behaviour changes with the duration of the panel. Furthermore we have also seen in Figure 7 that the different attrition rates across the citizenship categories led to a share of *“other citizenship (Extra-EU)”* of 0.87% in 1996 which was less than a half of the one in 1994 (1.91%). It has to be assumed that by this effect representativeness is influenced negatively especially with respect to those respondents with an origin outside the EU. Nevertheless after two waves the ECHP panel obviously reached a share for this category which is close to the percentages in the BHPS of the years from 1997 to 1999.

Citizenship 95 (Abbrev)		Citizenship 96 (Abbrev)					
		UK	another EU citizenship	other citizenship (Extra-EU)	not national, but citizenship unknown	missing	attrition
UK	N	6525.1	10.3	3.5	0.3	0.4	1651.9
	% (1995)	79.66	0.13	0.04	<.01	<.01	20.17
	% (1996)	99.63	16.35	6.10	18.45	100.00	96.41
Another EU citizenship	N	13.7	51.0	1.2	1.1	.	22.2
	% (1995)	15.34	57.16	1.36	1.27	.	24.88
	% (1996)	0.21	80.68	2.09	81.55	.	1.30
other citizenship (Extra-EU)	N	7.0	1.9	52.2	.	.	33.2
	% (1995)	7.47	1.99	55.31	.	.	35.23
	% (1996)	0.11	2.97	89.81	.	.	1.94
not national, but citizenship unknown	N	1.4	1.6
	% (1995)	46.30	53.70
	% (1996)	0.02	0.10
Missing	N	2.2	.	1.2	.	.	4.3
	% (1995)	28.62	.	15.14	.	.	56.24
	% (1996)	0.03	.	2.00	.	.	0.25
Total	N	6549.5	63.3	58.1	1.4	0.4	1713.4
	% (1996)	78.10	0.75	0.69	0.02	<.01	20.43

Table 5: Transitions between categories of *Citizenship* in ECHP from 1995 to 1996

Table 6 contains figures for transition in ECHP across two waves, i.e. from 1994 to 1996. Due to repeated attrition, item persistency is falling for “UK” down to 61%, for “another EU citizenship” to 43% and from those who stated in the first wave an “other citizenship (Extra-EU)” only 24% still had this value for the variable in the third wave. That means conversion by using item persistency is more difficult for EU-citizens than for those from UK and even harder for citizens from outside the EU

than for those from within. Especially for the vast majority of the latter group of respondents, due to attrition no values will be available from later waves.

Citizenship 94 (Abbrev)		Citizenship 96 (Abbrev)					
		UK	another EU citizenship	other citizenship (Extra-EU)	not national, but citizenship unknown	missing	attrition
UK	N	6227.2	11.4	3.3	1.2	0.5	3947.0
	% (1994)	61.11	0.11	0.03	0.01	<.01	38.73
	% (1996)	99.46	17.11	6.25	100.00	100.00	95.47
Another EU citizenship	N	6.8	51.1	1.2	.	.	58.5
	% (1994)	5.79	43.47	1.01	.	.	49.73
	% (1996)	0.11	77.03	2.24	.	.	1.41
other citizenship (Extra-EU)	N	24.4	3.9	48.3	.	.	124.7
	% (1994)	12.11	1.93	24.01	.	.	61.94
	% (1996)	0.39	5.86	91.51	.	.	3.02
not national, but citizenship unknown	N	1.3
	% (1994)	100.00
	% (1996)	0.03
Missing	N	2.7	2.7
	% (1994)	49.77	50.23
	% (1996)	0.04	0.07
Total	N	6261.1	66.4	52.8	1.2	0.5	4134.2
	% (1996)	59.54	0.63	0.50	0.01	0.00	39.31

Table 6: Transitions between categories of *Citizenship* in ECHP from 1994 to 1996

Nevertheless, for those who are still in the panel it seems to be plausible that values can be transferred to the previous waves. For UK-citizens this means that almost 99.5% made the same statement two years before. If transition from other

categories to UK-citizenship mainly reflects real changes in citizenship, i.e. naturalisation, this percentage necessarily must be so high since rate of naturalisation in relation to the whole population is very small. Things are a little bit different for EU-citizens. For only about three quarter of them the statement concerning citizenship was the same two years before. Especially those 17% who stated two years before an UK-citizenship would not have got the correct value by having it transferred from 1996.

In contrast to the EU-citizens from 1996, 91.5% from the respondents with an “*other citizenship (Extra-EU)*” made the same statement in 1994. Therefore for this group we also have a fairly good possibility to transfer existing values from later waves. Supposing that these values exist, which is the case for only a minority because these 91.5% are of course only 24% from those who started with this statement in 1994. A remarkable share (12%) of those who stated an “*other citizenship (Extra-EU)*” at the beginning switched their answer till 1996 to a “*UK-citizenship*”. These are not less than one half of those who kept their first statement. A large part of this transition will be based on real changes, i.e. naturalisation. This high percentage, especially if the comparatively low share of persistent values is taken into consideration, corresponds with the British naturalisation laws which designate only a five year stay as necessary to attain the UK-citizenship. This period is only half as long as the necessary time in Germany and Austria ¹⁶.

The transition matrix in Figure 8 displays the one year transition probability given an existing value for the value in the previous wave. Two such probabilities can be computed with ECHP data: for the transition from 1994 to 1995 and for 1995 to 1996. For all categories of citizenship the matrix assigns probabilities to every possible predecessor value. This enables us to make the scale of uncertainty visible which accompanies the transfer of variable values over time assuming persistency of these values. Figures in the diagonal of the matrix are a measure for the quality of such a transfer compared with original data. For the first three categories, those with empirical relevance, both probabilities, each resembled by

¹⁶ see Tomei, Verónica: Überblick zu Staatsangehörigkeitskonzepten in der EU, p. 4

one of the two circles, are very close together in their scale. The percentages given in the matrix are those for the latest transition from 1995 to 1996. Especially due to transitions from “UK” to “another EU citizenship” (16.35%) the largest bias is occurring upon conversion of “another EU citizenship”. A wrong value would be assigned to almost one out of five respondents. For “other citizenship (Extra-EU)” we had a wrong assignment only for one out of ten. Best results of course could be obtained for the “UK- citizens” with more than 99.6% values transferred correctly.

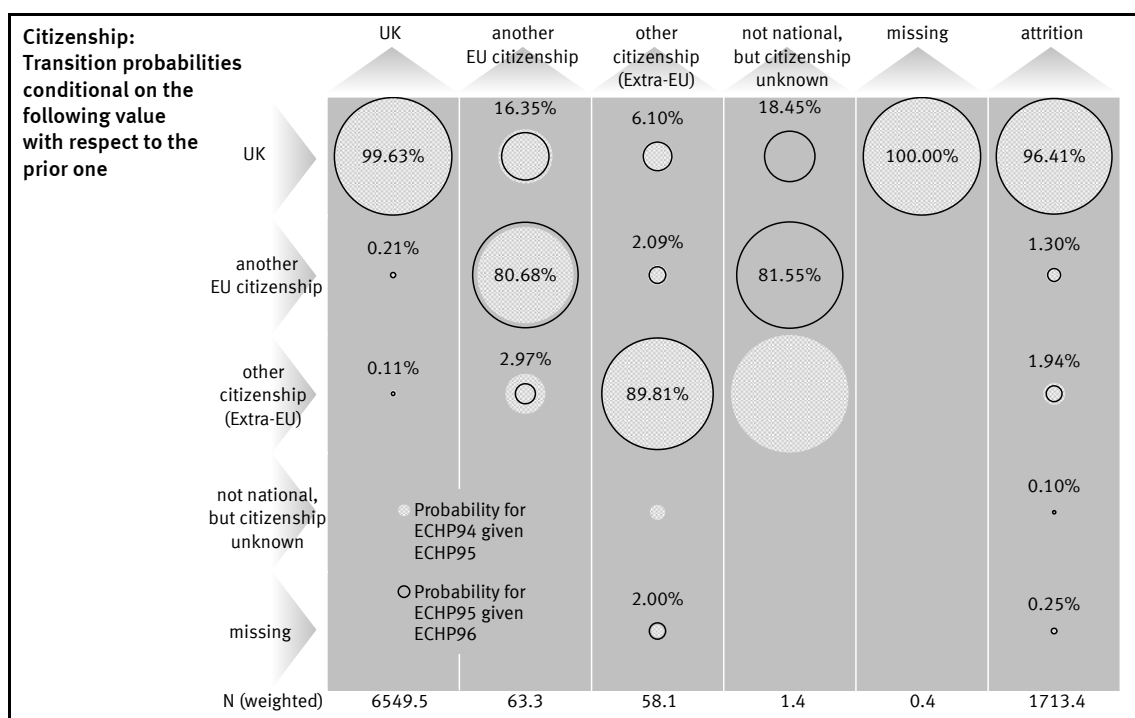


Figure 8: Conditional transition probabilities for *Citizenship* given the subsequent value

These percentages indicate the share of values converted correctly out of those which actually could be converted (from one wave to the prior one). The figures do not say anything about the share of the values which could be converted correctly out of those which should have had converted (from the first wave). Percentages are displayed in Figure 9 in the same manner like above. We see that circles and figures in the diagonal are smaller since for a large part of respondents who left the panel after the first wave there are no values available. These appear in the attrition column on the right. As attrition is getting smaller with age of the panel (see Work

Package 6) it can be assumed that efficiency of this conversion method is higher for older panels than for younger ones. We can also observe that the group with the highest share of attrition is the one with missing values. This corresponds with the findings from Work Package 4 that item non-response is positively associated with attrition at the next wave.

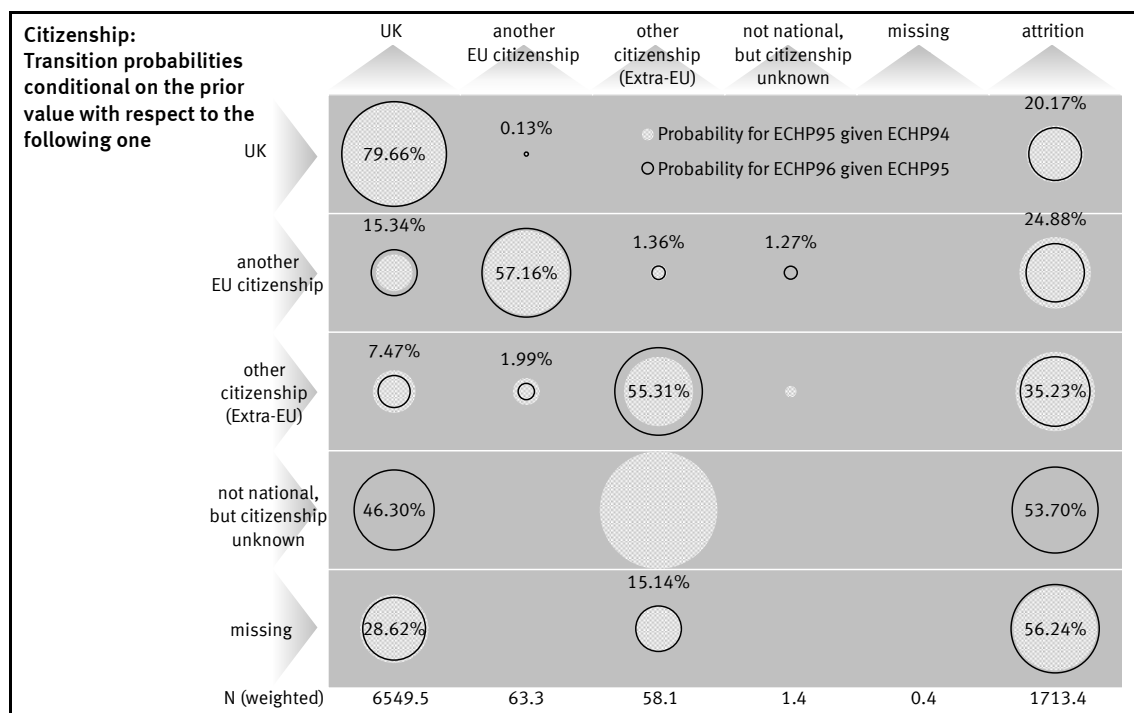


Figure 9: Conditional transition probabilities for *Citizenship* given the previous value

3.3 Conversion process

In BHPS the variable *citizenship* was introduced in 1997. In Table 7 weighted observations and percentages from this year are listed. In a first conversion step we have to try to transfer these existing values to the data sets from the previous years. Transferring the corresponding values in micro-data to the respondents from the years 1996 down to 1994 leads to the figures also displayed in this table. In 1996 for 6% of these respondents no value exists due to attrition. This rate is by far lower

than the one in ECHP (24% after the first wave and 20% after the second). Therefore conversion covers after all 94% of the BHPS sample from 1996. If attrition is disregarded, we end up with comparable (net) percentages for the several categories of citizenship, labeled “%, net” in the table. Considering the low attrition rate, it does not surprise that these are very close to the figures from 1997.

BHPS wave	Citizenship (Abbrev)	Observations (weighted)			adjusted by transition probabilities from ECHP		ECHP
		N	%	%, net	N	%	%
1997	UK	10,559.2	97.52				
	another EU citizenship	134.5	1.24				
	other citizenship (Extra-EU)	106.8	0.99				
	not national, but citizenship unknown	17.3	0.16				
	missing	10.2	0.09				
1996	UK	8,445.4	91.87	97.74	8,972.1	97.60	98.02
	another EU citizenship	99.2	1.08	1.15	119.2	1.30	1.09
	other citizenship (Extra-EU)	78.2	0.85	0.90	93.0	1.01	0.87
	not national, but citizenship unknown	15.6	0.17	0.18	2.4	0.03	0.02
	missing	2.2	0.02	0.03	5.8	0.06	<0.01
	no value due to attrition in next wave	551.9	6.00				
1995	UK	7,975.7	88.81	97.80	8,754.6	97.49	97.68
	another EU citizenship	87.4	0.97	1.07	89.4	1.00	1.06
	other citizenship (Extra-EU)	77.4	0.86	0.95	131.9	1.47	1.13
	not national, but citizenship unknown	13.4	0.15	0.16	0.3	0.00	0.04
	missing	1.2	0.01	0.01	4.0	0.04	0.09
	no value due to attrition until 1997	825.1	9.19				
1994	UK	7,724.0	83.96	97.89			96.90
	another EU citizenship	78.1	0.85	0.99			1.12
	other citizenship (Extra-EU)	74.1	0.81	0.94			1.91
	not national, but citizenship unknown	13.4	0.15	0.17			0.01
	missing	0.9	0.01	0.01			0.05
	no value due to attrition until 1997	1,309.1	14.23				

Table 7: Transfer, adjustment and evaluation of *Citizenship* values from BHPS

If BHPS also was our target survey, for those respondents who did not leave the panel this transfer would correspond to the variable construction in the later waves after the variable citizenship was introduced and therefore it would be sufficient for conversion. But it is the aim to convert data from BHPS into the survey context of ECHP. Thus transition between two or more waves has to be assumed like it is

recorded in ECHP. A correct conversion had to take these deviations of complete persistency into consideration. Therefore, the shares of the different categories of citizenship were adjusted by the specific conditional transition probabilities resulting from the last transition between ECHP waves 1995 and 1996. This adjustment may show whether introducing changes in citizenship in BHPS leads to data more comparable with ECHP. Adjusted shares and those from ECHP are listed in Table 7 on the right. All these columns were also computed for the conversion of data from BHPS 1997 to 1995, where the specific conditional two year transition probabilities were used. Surprisingly, attrition from 1995 to 1997 is about 9% and thus not much higher than attrition from 1996 to 1997. From 1994 to 1997 only 14% left the panel.

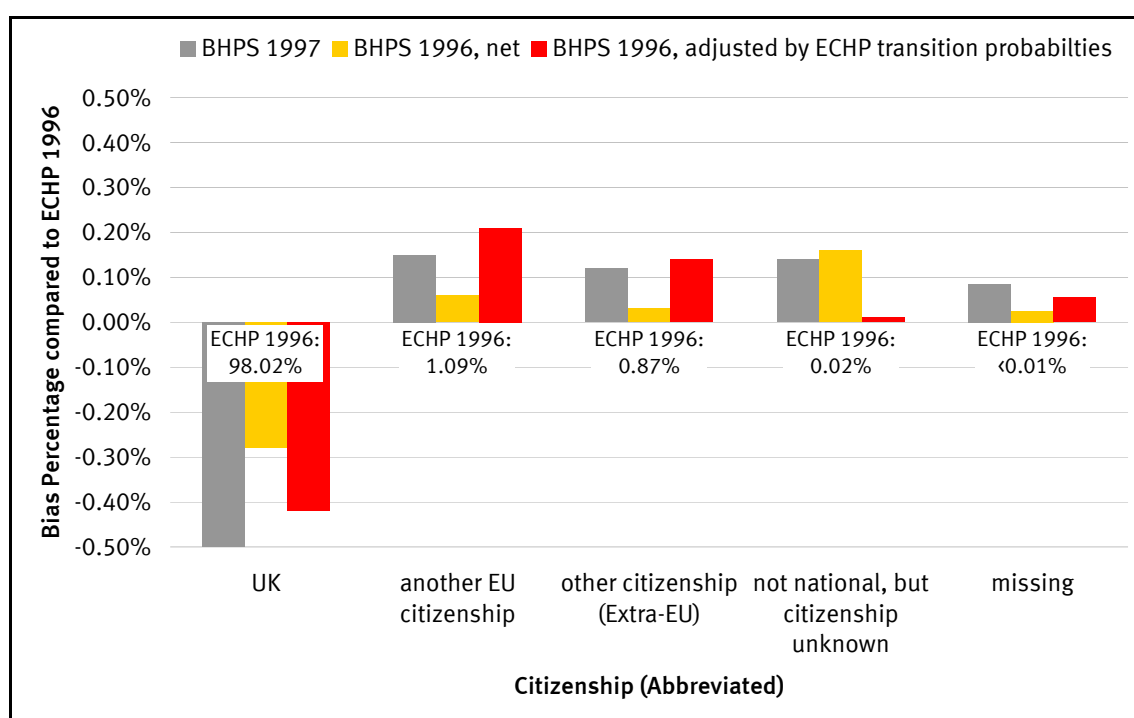


Figure 10: Bias of BHPS *Citizenship* distributions compared to ECHP for 1996

In Figure 10 data is displayed in form of biases in relation to ECHP 1996 for three different distributions. First distribution is the one from BHPS 1997. The second refers to data from BHPS 1996 where citizenship was transferred from the following year neglecting those respondents who left the panel after this wave. Adjustment of

the latter data by ECHP transition probabilities results in the third bias. Graphical representation is reflecting that for all these distributions share of UK citizens is lower and share of the other categories, especially those for the other citizenships is higher than in ECHP 1996. With respect to Figure 9 (Conditional transition probabilities for *Citizenship* given the previous value) and Table 7 (Transfer, adjustment and evaluation of *Citizenship* values from BHPS) it was discussed already that older panels are more successful in preserving its sample size than fresh started panels, like it was also found in Work Package 6. Furthermore it was shown that attrition probabilities in ECHP were higher for alien respondents than for UK citizens. Therefore it has to be assumed that due to attrition in the first waves of a fresh panel, a large bias is produced towards an underestimation of these alien respondents in comparison to an older panel where shares are more stable. This is what we see in Figure 10.

Furthermore the different columns are showing that by transfer of values from BHPS 1997 to the respondents of the 1996 wave the share of UK citizens is increasing and the shares of EU and other citizenships are decreasing. We have to assume that in the group of those who left the panel after 1996 and for whom therefore no values from 1997 are available, respondents with another than an UK citizenship are overrepresented. Therefore this group may have lower shares in 1996 than in 1997. That means that similar to ECHP, attrition is more related with this group than with UK citizenship. The hypothesis from Work Package 4, that item non-response is positively associated with attrition in the next wave, is supported by the BHPS figures as well: since net share in 1996 is only about 0.02% compared to 0.16% in BHPS 1997, a large overrepresentation in the group of attriters may also be assumed for those who have a missing value for citizenship.

For understanding of the third series of columns in Figure 9 we have to remember the path which led to it. In a first step those respondents from BHPS 1996 which still were in the panel in 1997 got assigned the citizenship category from this wave. In that way those who left the panel were neglected when distribution was computed. In a second step these shares were adjusted again in order to reconstruct influence by transition, both between the different citizenship

categories and towards attrition, on the basis of the ECHP transition matrix. Therefore adjustments are made with regard to the transition probabilities in ECHP. For 1996 this adjustment causes a compensation of attrition effects, i.e. share of alien respondents is increased whereas share of UK citizens is decreased. This leads to the conclusion that data quality after conversion is getting better when adjustment by transition probabilities could be used. However this only is possible if transition data does exist and this is not the case for BHPS. Therefore, if data sets like those from the ECHP are not available, respondents who left the panel had to be estimated, considering the attrition bias towards underestimation of alien citizenship. Analogously this is also necessary for reconstruction of missing data. Certainly this does only make sense if data quality is more important than harmonisation. If the main priority is to receive data comparable to ECHP, this adjustment may cause less comparability because the attrition bias is reduced in BHPS whereas it is by far larger in ECHP. Therefore in this case an adjustment would not be useful anyway. This is of course an important systematic problem for conversion, whether best data quality should be reached or just the same as in the target study.

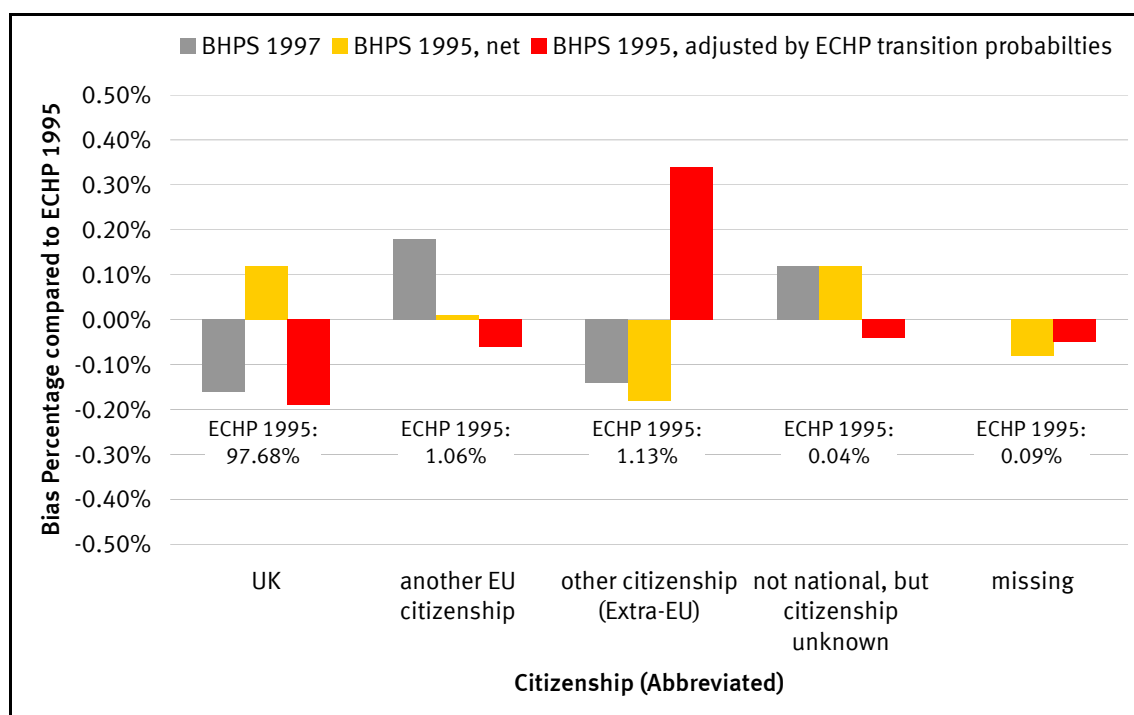


Figure 11: Bias of BHPS *Citizenship* distributions compared to ECHP for 1995

In Figure 11 results of these conversion processes are displayed for variable transfer across two waves from 1997 to 1995. In accordance to the observations made above, we see that for UK-citizens, due to the fresh panel’s biased attrition, difference to BHPS 1997 is still smaller for the second ECHP wave than for the third one. Probably for similar reasons for the other values of citizenship from BHPS 1997 we can not find a general underestimation with respect to ECHP 1995 as we did in comparison to ECHP 1996. In BHPS carrying values backward to the prior wave leads to a biased attrition effect towards an underestimation of alien citizens. Due to this procedure attrition effect is orientated backward. In ECHP things are just the other way round since attrition effects are in the same temporal direction as the panel is.

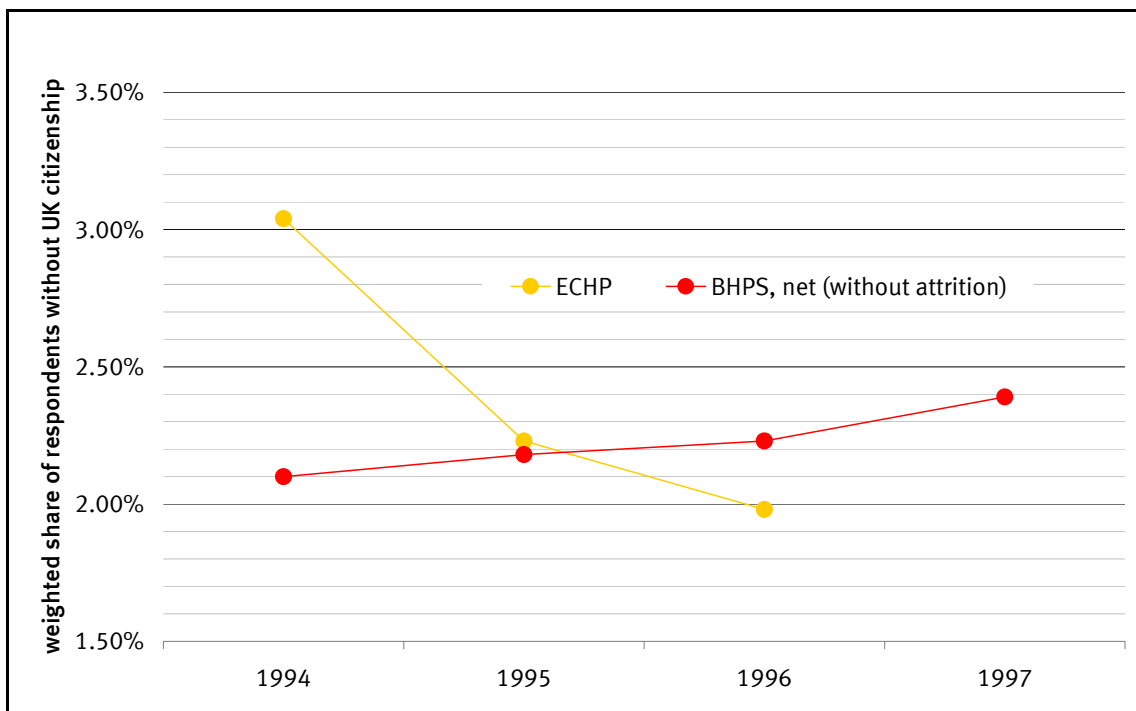


Figure 12: Respondents without UK citizenship in BHPS and ECHP waves

Figure 12 shows that these two effects of BHPS data with transferred citizenship on the one hand and ECHP data on the other hand cross each other. Adding shares of

all categories of citizenship other than UK, both lines meet each other in 1995. This is reflected by the very diverse picture of the columns in Figure 11 in contrast to those of Figure 10. Both attrition effects have an opposite direction and therefore in a certain wave they compensate each other. Whereas in 1996 due to biased attrition the share of respondents without UK citizenship reached a very low level in ECHP, for the transferred BHPS data this share was still high. In 1994 this is not the case anymore. Due to impact of attrition on conversion the computed share is by far lower for BHPS than for the first wave from ECHP. Besides this opposite transition effect we can also observe the large differences in the two curves slope. For each curve the slope indicates the intensity of the attrition bias which affects the sample from one wave to the other.

4 Conclusions

Within the CHINTEX project, conversion was understood as the instrument for ex-post generation of a micro data set on the basis of existing survey data in such a way that the data set is fully comparable to equivalent data sets from other surveys in terms of the distribution of its variables and estimation results. In view of ex-post harmonisation it is the aim of conversion to produce a data set like it was available if data would have been gained by an input harmonised survey.

Since in the non-trivial case of data conversion the states of existing and desired information are not equal, converting means modifying available data in order to exploit the necessary information from survey internal or external sources. If additional information in explaining variables within the survey to be converted is transformed into the desired variable this is a transformation of an attribute. A transfer of an attribute is possible if the same variable is used in another survey which enables us to draw some conclusions about information hidden in correlated variables and transfer this information to the survey to be converted.

Regardless of the source referring to, every gained information is experiencing a modifying extraction out of its original context. Either it leads to a modification of the observation instrument or the sample context is replaced by another one, assuming that the way the assignments work in the different samples and also the representativeness of the population elements, respectively the sampling design, are transferable.

In general, transfer and transformation of information from internal or external sources is possible in a qualified sense:

If redundancy of studies is utilised, target survey and information source have to cover the same population. Sampling design and measurement functions should be known and comparable. Structural similarity of populations means that information transfer depends on a high homogeneity from two populations.

Observation instruments should be same and sampling design should be known and comparable.

For the utilisation of structural persistency we have to ensure that in the information source there has to be a very small share of transitions between the possible values. By means of the variable *citizenship* it was shown that even for quite persistent variables it has to be assumed that transitions especially in terms of attrition exists and this has to be considered. These effects are hard to made out since they vary, probably for most variables, over the other two dimensions of the panel data framework: wave and population. For BHPS and ECHP data we made the observation that attrition is very high for the first wave of a panel and that it is decreasing continuously later on and on the other hand persistency is increasing. Therefore it is harder to transfer individual values to a previous wave within a fresh panel than within an older one in which sample population does not change very much. Transfer of data is accompanied by a bias due to the variation of attrition across several stratum of the population. There is a close correlation between attrition and citizenship. Since respondents without an UK citizenship are more likely to leave the panel, shares of the different categories are very unstable and are changing with the duration of the panel towards smaller shares of this stratum. Considering both effects together we conclude that a transfer bias is smaller within an old panel than within a new one. However, before conversion is applied in general a systematic question has to be answered: whether best data quality should be reached after conversion of a data set or whether harmonisation should also consider biases in input harmonised data sets, e.g. by attrition. If conversion has to be applied, both by utilisation of internal and external data, the framework developed in this part of the project, gives an instrument at hand to ensure that harmonisation aspects are considered with respect to all possible dimensions of data and appendant causes for biases. If conversion problems are located within this model, it should be possible - like it was shown for the example - to examine the relevant determinants of survey data with respect to a harmonised assignment of population characteristics. This work has to be deepened with respect to the increasing need for harmonised data from totally different sources of micro-data.

References

- Eurostat, ECHP UDB manual, Waves 1 to 5, survey years 1994 to 1998, Luxembourg December 2001 (DOC PAN 168/2001-12)
- Eurostat, ECHP – 1994, Wave 1 variable list, Luxembourg February 1994 (DOC PAN 15/1994)
- Eurostat, ECHP – 1995, Wave 2 variable list, Luxembourg February 1995 (DOC PAN 30/1995)
- Eurostat, ECHP – 1996, Wave 3 variable list, Luxembourg March 1996 (DOC PAN 65/1996)
- Eurostat, ECHP UDB, construction of variables, from ECHP questions to UDB variables, Luxembourg May 2001 (DOC PAN 167/2001)
- Eurostat, ECHP UDB description of variables, codebook and differences between countries and waves, Luxembourg December 2001 (DOC PAN 166/2001-12)
- Eurostat, Harmonisation of core variables, Luxembourg June 2000 (Doc. Eurostat/E0/00/DSS/2/6/EN).
- Eurostat, The European Community Household Panel (ECHP): Volume 1 – Survey questionnaires: waves 1-3, Theme 3, series E, Luxembourg 1996.
- Eurostat, The European Community Household Panel (ECHP): Volume 1 – Survey methodology and Implementation, Theme 3, series E, Luxembourg 1996.
- Gabler, Siegfried, Datenfusion, in: ZUMA-Nachrichten 40, Jg. 21, Mai 1997, Mannheim, 81-92
- Günther, Roland, CHINTEX Deliverable No. 3: Report on compiled information, 2002
- Rässler, Susanne, Fleischer, Karlheinz, Aspects Concerning Data Fusion Techniques, Friedrich – Alexander-Universität Erlangen-Nürnberg, Wirtschafts- und Sozialwissenschaftliche Fakultät, Discussion Paper 16/1997
- Tomei, Verónica: Überblick zu Staatsangehörigkeitskonzepten in der EU, efms Paper Nr. 25
- Van Buuren, Stef et al., Response conversion: A new technology for comparing existing health information, TNO report 2001.097, TNO Prevention and Health, Leiden, June 2001