

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/311402845>

# Longitudinal statistical matching: transferring consumption expenditure from HBS to SILC panel survey

Working Paper · November 2016

CITATIONS

0

READS

15

2 authors:



[Baris Ucar](#)

Turkish Statistical Institute

6 PUBLICATIONS 1 CITATION

[SEE PROFILE](#)



[Gianni Betti](#)

Università degli Studi di Siena

87 PUBLICATIONS 630 CITATIONS

[SEE PROFILE](#)



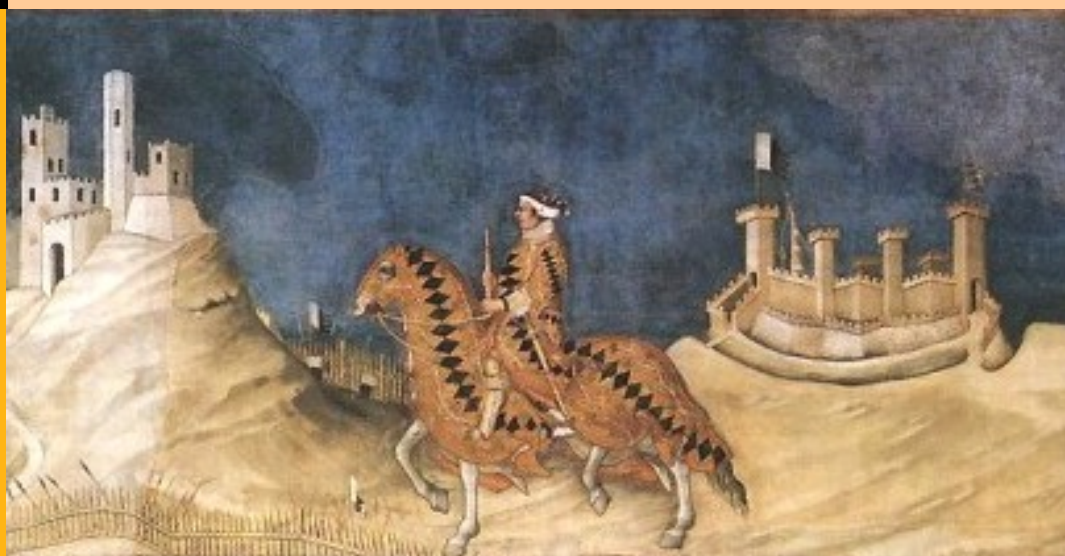
UNIVERSITÀ  
DI SIENA  
1240

**QUADERNI DEL DIPARTIMENTO  
DI ECONOMIA POLITICA E STATISTICA**

**Baris Ucar  
Gianni Betti**

Longitudinal statistical matching: transferring consumption  
expenditure from HBS to SILC panel survey

**n. 739 – Novembre 2016**



# Longitudinal statistical matching: transferring consumption expenditure from HBS to SILC panel survey

Baris Ucar and Gianni Betti

University of Siena

## Abstract

The aim of this study is to look for an appropriate procedure to conduct statistical matching for longitudinal data sets. To our knowledge, among the studies, which are associated to statistical matching with longitudinal data, no such issue has been specifically covered or identified in literature. The longitudinal data set, which is used in this study, involves a longitudinal weight at individual level, which requires further procedures before the matching application. The study will discuss and propose ways to deal with the statistical matching issue for such data sets.

In the process, a four-year longitudinal data set is used and data from each year is matched with a cross-sectional data set for the corresponding year. The matching procedure is comprised of two steps, respectively the Renssen (1998) method followed by nearest neighbor distance hot deck matching, proposed by D’Orazio (2016) and Donatiello et al. (2015).

The application is undertaken on Turkish data to impute consumption expenditure variable from Household Budget Survey (HBS) to Statistics on Income, and Living Conditions (SILC) Survey. A synthetic longitudinal data set is created by using these two survey data sets. The two data sets have many variables in common including income variable, which enables Conditional Independence Assumption (CIA) to be more likely.

---

Baris Ucar, Institute of Population Studies, Hacettepe University; email: [ucarbaris1@hotmail.com](mailto:ucarbaris1@hotmail.com);

Gianni Betti, Department of Economics and Statistics, University of Siena; email: [gianni.betti@unisi.it](mailto:gianni.betti@unisi.it);

The study also carries out validation analyses to determine the quality of the matching procedures and the findings achieved with the proposed itinerary, indicate that the distribution of consumption expenditure estimate in synthetic data set is well preserved with all estimates. The poverty indicators in general and at household level is also looked for and the results indicate a good quality match.

**Keywords:** Statistical matching, Consumption Expenditure, SILC, Household Budget Survey,

**JEL Classification:** C19, C83, E21

## **1. Introduction:**

Statistical matching (SM), in broad terms, is the name of the procedure for merging two or more different data sets, in order to make use of the variables, which are not simultaneously available in either data set. It enables exploiting more from the available data sets to produce more information for inference. Instead of implementing a survey or a census in which all required variables are available, a statistical procedure is put to work which is less costly and more feasible. The data sets to be fused should refer to the same population (D’Orazio, 2016) and the files should be combined in such a way that the distributions of the related variables stay unchanged as much as possible (Kum and Masterson, 2008).

In statistical matching method, the idea is to fuse variables in two data sets by making use of a common set of variables, which are available in both data sets.  $X$  denotes variables available in both data sets, and  $Y$  and  $Z$  denote variables that are only available in one of the data set respectively. The aim is to obtain a data set including  $X$ ,  $Y$  and  $Z$ . Data fusion, data combination, micro data set merging, synthetic matching are other names given to this method (van der Putten et al., 2002; Kum and Masterson, 2008; Leulescu and Agafitei, 2013). In general, one of the data sets is the **recipient** and the other one the **donor**. The matching is realized by transferring variables from the donor to the recipient by making use of the matching variables.

Record linkage should not be confused with statistical matching while at some aspects and at some implementations they have similar prospects. The difference is identified with regard to the units in question. Record linkage is used in case of overlapping units where one-to-one direct matching can be realized. Similar units are the subject of statistical matching. On the other hand, identical units are the subject of record linkage. (Leulescu and Agafitei, 2013)

This paper came out for the need of longitudinal consumption expenditure data in Turkey to be used in research with regard to the relationship between childbirth and poverty.

In Turkey, there is a longitudinal data set of a four-year span from SILC survey, which comprises information on family formation throughout time that includes the presentation of a newborn in the household, as well as other information with regard to household characteristics. Unfortunately, this data set lacks information on household consumption expenditure, which is available in Household Budget Survey (HBS). Since no

such data is available, creating a synthetic data set by fusing available data sets demonstrates itself as a feasible solution. Therefore, this necessity of an ad hoc data set lead to the efforts provided in this paper.

The incorporation of the two surveys will be executed by using **statistical matching method**. Further specifications, requirements and formation of the new data set will be discussed in the following sections in detail.

As pointed out earlier, the main target of this study is to create a data set, which will be used in further research on the relationship between childbirth and poverty in Turkey, for which there is need for longitudinal consumption data. The literature is scarce with regard to studies that target statistical matching with longitudinal data. This study has the property of being one of the few in such effort. The longitudinal data set which is used in this study involves a longitudinal weight at individual level which requires further procedures before the matching application. Among the studies which are associated to statistical matching with longitudinal data, no such issue has been specifically covered or identified in literature which can be considered as a novelty which this study deals with.

This paper is organized as follows. After the brief introduction, which also comprises the main objective of the study, literature review on statistical matching will be put forward in section 2, which comprises of general literature on statistical matching as well as literature specific to the matching of cross-sectional and longitudinal data and literature specific to Turkey in the subsections. Then the methodology will be put forth in section 3. In section 4, the data sets will be defined briefly and descriptive tables will be presented. In section 5 how data sets are prepared for matching will be explained. In section 6, the statistical matching procedures gone through will be explained step by step with full detail. In section 7, the results will be presented and discussed. The quality of the match will be evaluated in section 8 and section 9 will conclude the paper.

## **2. Literature Review on Statistical Matching:**

### **2.1. General Literature on Statistical Matching**

Statistical matching is relatively a new area of research. Okner (1972) is regarded to be the first one to produce academic research in this regard. He merged two files namely the 1967 Survey of Economic Opportunity and 1966 Tax File in order to obtain income distribution with regard to demographic characteristics, which was not available in any available data sets. He used “equivalence classes” which is defined as comparable characteristics available in both files. [Okner \(1972\)](#) defines the procedure as costly and time-consuming, when computer technology was at a much lower level. All the same, he thinks the effort is worth it. In this era, where computers are much more powerful and capable to carry out procedures much faster, the value of statistical matching outshines compared to its alternatives such as conducting a new survey. Because of the advantages it presents today with the help of computers, statistical matching is being used widely all over the world. Many researches have been conducted in this regard. Varieties of new techniques have been developed and researches have been conducted on every detail of the matching process.

Kum and Masterson (2008) indicate the use of statistical matching in different areas such as medical research and economics giving Little and Rubin, 2000; [Rubin and Thomas, 1992, 1996](#); [Rosenbaum and Rubin, 1983](#) as examples of use of statistical matching in medical literature. In the area of economics, they mention numerous examples cited by Rässler (2002), as well as studies by Radner (1981), [Wolff \(2000\)](#), [Greenwood \(1983, 1987\)](#), [Wagner \(2001\)](#), Brodaty, Crépon, and Fougère (2001), and [Keister and Moller \(2000, 2003\)](#), along with many others.

Statistical matching methods have been classified in various ways by some researchers, according to the different characteristics they hold. D’Orazio et al. (2006) have summarized the classifications of these approaches as macro and micro; and parametric, nonparametric and mixed methods. Statistical matching could be realized to obtain joint distributions of variables from two different data sources. In this case, a macro level matching would be in question. If the aim is to obtain a new synthetic data set via the fusion of the data sets then it is a micro level matching. Parametric models could be used for the matching as well as nonparametric methods. There are also cases when both are used in the same process, which is the so-called mixed method.

The matching at micro level could be realized with distance functions, predictive mean matching or by making use of propensity scores (Kum and Masterson, 2008). Methods using distances include nearest neighbour, random or rank hot deck procedures (D’Orazio, 2016).

The procedure can be realized with a constrained statistical matching (CM) where each item can be matched only for once or with an unconstrained statistical matching (USM) where the matching is more disengaged. In USM, a distance function is used for finding the nearest neighbour. When this method is used, it is possible that there are multiple selections or no selection of records from donor data set. The result of this could be different marginal distributions of X (matching variables) or joint distributions of X and Y (variables in the recipient data set), in the statistically matched file compared with those in the original donor file (Kum and Masterson, 2008). On the other hand, CSM does not allow for multiple selection. The records are matched with regard to their rank. The disadvantage of the CSM approach is that matches are possible even with unacceptably large distances. However, in the final synthetic data set all marginal distributions are the same as they are in the original files.

One main issue when dealing with the statistical matching problem is the so-called Conditional Independence Assumption (CIA). When matching is realized by using relationship of X and Y; and X and Z respectively, to obtain an estimate for the relationship between Y and Z, it is assumed that Y and Z are independent conditional on X. This situation, which is not true in most of cases, is called the CIA. Unless there is additional information from another source, this cannot be tested and acceptance of this assumption is one of the weaknesses of SM procedures.

When there is auxiliary information from another source, this assumption can be relaxed and this information can be used to obtain higher quality SM results.

## **2.2. Statistical Matching for Consumption Expenditure:**

Recently, there are statistical matching applications of HBS and SILC data sets where consumption expenditure variable in HBS is imputed into SILC. Donatiello et al. (2014) has carried out this task with Italian HBS and SILC data sets. The consumption expenditure was categorized in this study. Data set of SILC with reference year 2011 for income and 2012 for other variables was matched with HBS 2011. The reason for this choice was to enable comparative analysis of expenditure and income in the synthetic data



file. In this study, use of auxiliary information in order to avoid Conditional Independence Assumption (CIA), which is a critical issue in SM, was evaluated. The auxiliary information relied on the monthly household income, which was derived from HBS.

Another study for this kind of matching is by Baldini et al. (2015). This study also concentrates on imputing expenditure information in HBS into SILC data set in Italy. This time all expenditure items are imputed with a two-stage procedure by making use of expenditure-income relationship which is derived from another data set (Survey on Household Income and Wealth (SHIW)) where joint information on both expenditure and income variables is available where all data sets correspond to 2012. The method in this study consists of three steps. In the first step, in SHIW data set expenditure is regressed on income with other variables, which are also available in SILC. Then, in the second step, estimates of expenditure is obtained in SILC data set. In an intermediate step, households are sorted by per centiles of imputed expenditure in SILC and sorted by original overall expenditure in HBS. Finally, in the last step distance function matching is applied.

A recent study by Webber and Tonkin (2013) also integrated expenditure data in HBS with SILC for UK, 2005. They used three different methods, namely, parametric, nonparametric and mixed methods and found that the mixed methods were slightly better in the matching. EU-SILC in the UK measures current income unlike to other European countries. In HBS, also current income and expenditure is collected, but the reference period, on other hand, is the 2005/2006 financial year. However, these values are deflated to 2005 for coherence.

### **2.3. Literature on Matching of Longitudinal Data Sets**

Up to the present, the literature on statistical matching of cross-sectional and longitudinal data sets is scarce compared to the vast literature on statistical matching.

Betti (1998) imputed consumption expenditure into British Household Panel Study data set by making use of a consumption model created in Family Expenditure Survey for years 1991 to 1994. The matching method was completely parametric.

[Rasner \(2007\)](#) described the preparatory steps for matching administrative data, Completed Insurance Biographies with German Socio-Economic Panel (SOEP). [Rasner et al. \(2011\)](#) discusses the realization of this matching. Mahalanobis distance matching was used in this study.

One conference paper has considered the issue where a statistical matching was exercised between a cross-sectional and longitudinal data sets using propensity matching method (Thiede et al., 2010). For further analysis of the main issue of this paper, a sample of longitudinal data comprising of insured people with monthly career changes was matched with a cross-sectional data set which contained information on the diagnosis which led to premature retirement. As this diagnosis information was not available in the longitudinal data set and it was required for the analysis, such a statistical matching was considered. The paper lacks the details of the quality of the match.

Simonson et al. (2012) German Aging Survey (DEAS) with a sample of administrative data from Active Pension Accounts (VSKT). In this study, they used Mahalanobis distance for the matching procedure. Both data sources in this study are longitudinal. The matching was realized between the corresponding years.

Zacharias et al. (2014) statistically matched South Korean Time Use Survey (KTUS 2009), and the South Korean Welfare Panel Survey of 2009. The matching is realized for one year of the panel, therefore longitudinal analysis was not targeted in this study.

#### **2.4. Literature Review on Statistical Matching in Turkey:**

Use of statistical matching methods in Turkey is very new. So far, there are only one study explicitly using statistical matching method in Turkey. The second one is "Time Deficits and Poverty" study by Zacharias et al. (2014). In this study, there was need for information on time spent on household production, time spent on employment and household consumption expenditures, but in maximum two of these were available in a single data set. In Household Budget Survey (HBS), time spent on employment and household consumption expenditures were variable, but there was no information on time spent on household production. This variable was available in Time Use Survey (TUS). Therefore, time spent on household production for each individual aged 15 years and older in TUS was fused into HBS data. Moreover, [Masterson \(2013\)](#) analyzed the quality of the match in another article where the match was found to be of high quality.

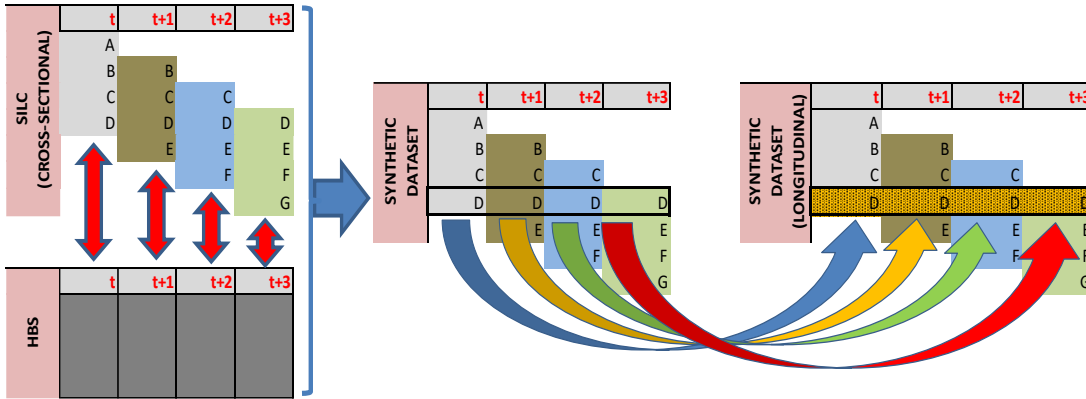
### 3. Methodology

For the creation of a synthetic longitudinal data set, including household consumption expenditure, longitudinal SILC data and cross-sectional HBS data from Turkey will be incorporated using statistical matching method. In the final synthetic data set, there will be variables (Y) from SILC, variables common in both data sets (X) and household consumption expenditure variable (Z) from HBS.

For this study, there will be need for more than one matching application. There is need for a matching for each year of SILC data with the corresponding year of HBS data.

The more information we have, the more will be inferred from it. Having more information available at hand would make the matching process more reliable, consistent and precise. The available information can be maximized by carrying out the statistical matching for the cross-sectional data sets of SILC with HBS (Figure 1) and following this step, the synthetic data set that is created can be matched (not statistical matching, direct record matching) with the corresponding records in the longitudinal data set only to select the longitudinal final data set (Figure 1). This way it will be secured that, a data set which is around four times greater than the section (each data part regarding to one year in the data set) of the longitudinal data set could be used for a better statistical matching implementation.

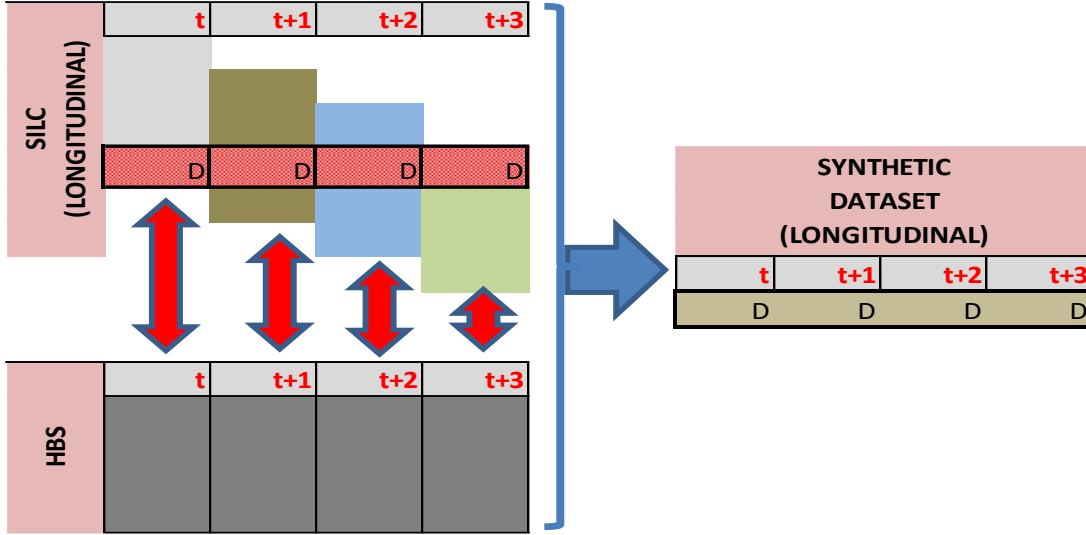
**Figure 1.**



When this is accomplished, a longitudinal data set will be available which is realized by the utmost information available for use. However, unfortunately this cannot be realized for Turkish data, which obstructs such use by concealing the necessary

identification number that would be used to link the records in the cross-sectional and longitudinal data sets of SILC. Therefore, it is only possible to realize the statistical matching procedure only by the corresponding section of the SILC longitudinal data set (Figure 2).

**Figure 2.**



This study will employ the method provided in StatMatch package in R, which is provided by D’Orazio (2016) as a framework. The steps suggested in this study will be followed and these steps will constitute the subsections of the “Procedure Steps” section. After a brief introduction of the data sets to be used, each step of the SM procedure will be put forth in detail. Therefore, further details regarding the methodology are available in the corresponding sections of the study.

### **3.1. Statistical Framework:**

In order to prevent complications that could arise due to differences in the statistical framework and notations that are coming from different studies, the statistical framework and notations will be borrowed from the same source. The statistical framework and notations to be used in this study are based on D’Orazio et al. (2006) and are summarized as follows.

Let  $X, Y, Z$  be a random variable with density  $f(x, y, z)$  and assume that  $A$  and  $B$  are two samples consisting of  $n_A$  and  $n_B$  independent and identically distributed (i.i.d.) observations generated from  $f(x, y, z)$ . Furthermore, let the units in  $A$  have  $Z$  missing and the units  $B$  have  $Y$  missing. When the objective is to gain information on the joint

distribution of (X, Y, Z) from the observed samples of A and B, we are dealing with the statistical matching problem.

#### **4. Data Set Preparation**

The main sources of data are the Household Budget Survey and SILC Survey as mentioned above. Both data sets have stratified and clustered sampling designs. Moreover, SILC includes a rotating structure with regard to its longitudinal design. Twenty five per cent of the sample is replaced with new ones each year.

Household budget survey provides information on socio-economic structures, standards of living, and consumption patterns of the households (Turkstat, 2013a). It is conducted on a yearly basis since 2003. The survey is conducted between January 1 and December 31, where sample households change every month.

Income and Living Conditions (SILC) Survey has been conducted every year since 2006 (Turkstat, 2013b). SILC Survey is carried out yearly by using panel survey technique for displaying the income distribution between individuals and households, measuring the living conditions of the people, social exclusion and poverty with the income dimension. The aim of the survey is to produce comparable data with the EU Countries, on income distribution, relative poverty, living conditions and social exclusion. It is applied according to the European Union Compliance Programme.

Respondents in the sample are monitored for four years in this survey. Panel survey technique is used and field application is carried out regularly every year. Twenty-five per cent of the households are changed every year. The longitudinal weights of SILC are only related to the individuals. Income and Living Conditions Survey is carried out regularly in each year. Data compilation is performed between April-July.

The data sets to be used are 2010-2013 longitudinal data of SILC and data sets of HBS for four corresponding years.

Before everything else, there is need for a step zero for the preparation of the data sets to be matched. At this stage, the framework suggested by van der Laan (2000) will be used as a basis which is also suggested by Leulescu and Agafitei, (2013) and D'Orazio (2016). The data preparation stage requires harmonization between the two data sets to be matched. The two data sets should be in accordance with each other as much as possible to enable a good quality match. The issues suggested by van der Laan (2000) to be

considered, are presented and explained with respect to the statistical matching exercise handled in this study.

a. harmonization of definition of units: are the statistical units defined uniformly in all sources? (special reference to comparability in space and time)

b. harmonization of reference periods: do all data refer to the same period or the same point in time?

c. completion of populations (coverage): do all sources cover the same target population?

d. harmonization of variables: are corresponding variables defined in the same way? (special reference to comparability in space and time);

e. harmonization of classifications: are corresponding variables classified in the same way? (special reference to comparability in space and time);

f. adjusting for measurement errors (accuracy): after harmonizing definitions, do the corresponding variables have the same value?

g. adjusting for missing data (item non-response): do all the variables possess a value?

h. derivation of variables: are all variables derived using the combined information from different sources?

## **4.1. Harmonization of the definition of units**

### **4.1.1. Households and Individuals**

Both surveys use the same household definition, which is an important issue to be considered in a statistical matching application. On the other hand, the panel structure of SILC creates some complications.

In HBS data sets, the issue regarding the relationship between households and individuals is straightforward. Every household and individual in the corresponding data sets have a weight and could be used directly. In SILC, selecting the individuals and households for each year of the panel is a little more complicated, since these are different for each year although we have weights attached to individuals only for the final year. For this purpose, the individuals that have a four-year panel weight are selected and afterwards the households that they belong are selected. And for each year all individuals that are members of these households in the corresponding year are selected and these

individuals and households constitute the population of that year. With this process, some of the households - that are left with no members in the final year, but have an ex-member with a weight who moved out from these households - are neglected. Because these households do not have a member with a weight in the four-year panel, they will not appear in the final synthetic data set, so they are deleted in advance. Number of such households are quite few and they only correspond to 0.5 per cent of all households.

#### **4.1.2. Age**

In HBS, like most of the other variables, age corresponds to the completed age at the time of the survey. On the other hand, in SILC, the age corresponds to the month of December preceding the survey year. This creates individuals that are of -1 age for those born between December of preceding year and the survey time. As will be explained in detail later, a weight calibration will be conducted. When such calibration is carried out at Turkstat, those at age -1 are considered in the 0-4 age group, therefore, this will study do the same.


#### **4.2. Harmonization of the reference periods**

While almost all variables in both data sets are pertinent to the survey year  $t$ ,  $t+1$ , etc., the income variable in SILC corresponds to year  $t-1$  for survey year  $t$ , to year  $t$  for survey year  $t+1$ , etc. (Figure 3) However, the incompatibility with regard to income and other variables in SILC requires some touch for harmonization, since it is a basic variable in targeted data set.

One way to overcome this could be to follow Donatiello et al. (2014). The matching of the SILC data set of year  $t$  could be carried out with HBS of  $t-1$ . This will harmonize income and consumption expenditure, but they will not be in accordance with other variables.


**Figure 3.**

SILC					
	<b>t</b>	<b>t+1</b>	<b>t+2</b>	<b>t+3</b>	<b>t+4</b>
	$X_t$	$X_{t+1}$	$X_{t+2}$	$X_{t+3}$	$X_{t+4}$ (na)
	$Y_t$	$Y_{t+1}$	$Y_{t+2}$	$Y_{t+3}$	$Y_{t+4}$ (na)
	$INC_{t-1}$	$INC_t$	$INC_{t+1}$	$INC_{t+2}$	$INC_{t+3}$ (na)
HBS	$X_t$	$X_{t+1}$	$X_{t+2}$	$X_{t+3}$	$X_{t+4}$ (na)
	$Z_t$	$Z_{t+1}$	$Z_{t+2}$	$Z_{t+3}$	$Z_{t+4}$ (na)



This inconsistency of periods could be better overcome by relating each year's SILC data set with the following year's income (Figure 4). When this is accomplished the income variable for the year  $t+3$  cannot be obtained because it is unavailable in the longitudinal data of the range  $(t - t+3)$ . It is only obtainable from the year  $t+4$  which is not available (na).

**Figure 4.**

SILC					
	<b>t</b>	<b>t+1</b>	<b>t+2</b>	<b>t+3</b>	<b>t+4</b>
	$X_t$	$X_{t+1}$	$X_{t+2}$	$X_{t+3}$	$X_{t+4}$ (na)
	$Y_t$	$Y_{t+1}$	$Y_{t+2}$	$Y_{t+3}$	$Y_{t+4}$ (na)
	$INC_t$	$INC_{t+1}$	$INC_{t+2}$	$INC_{t+3}$ (na)	
HBS	$X_t$	$X_{t+1}$	$X_{t+2}$	$X_{t+3}$	$X_{t+4}$ (na)
	$Z_t$	$Z_{t+1}$	$Z_{t+2}$	$Z_{t+3}$	$Z_{t+4}$ (na)

In this case, in the final data set, it is only possible to have a panel data set of three years instead of four years (Figure 5). This limits the data available for further research, but it seems to be one of the few ways of overcoming the existing problem.

**Figure 5.**

SYNTHETIC DATA SET			
	<b>t</b>	<b>t+1</b>	<b>t+2</b>
	$X_t$	$X_{t+1}$	$X_{t+2}$
	$Y_t$	$Y_{t+1}$	$Y_{t+2}$
	$INC_t$	$INC_{t+1}$	$INC_{t+2}$
	$Z_t$	$Z_{t+1}$	$Z_{t+2}$

There is one problem with this approach though. There are entries to and exits from the household, so the income information of year  $t$  does not exactly correspond to the household composition in year  $t$  because it is collected in year  $t+1$  and represents the household composition in year  $t+1$ .

Another alternative option could be to bring the income of year  $t-1$  to year  $t$  by using consumer price index (CPI). This could reduce the number of complications that are faced. This method also provides another very important advantage by enabling use of



four-year panel instead of three, by preventing the loss of one year of panel that would arise as a result of procedures that are followed in the other method.

### **4.3. Completion of populations (coverage)**

#### **4.3.1. Weight Calibration**

Since the overall population and its distribution with regard to age, sex and household size are different for the two data sets at hand; first, there is need for reconciliation. The longitudinal data set has a major divergence from cross-sectional data set populations due to its panel structure. Unlike the cross-sectional data set, the weights are only given at individual level in the longitudinal data set and because there are exits and entries to the households between two waves, neither the population total, nor its age and sex distribution is comparable to that of cross-sectional SILC and HBS data set.

At this point, the adjustment is realized by calibrating the sections of SILC panel data set populations to the corresponding year of HBS data set. The calibration will be realized towards HBS data set, which will enable similar populations with regard to age, sex and household size between the two matching data sets.

The problem here is that a simple calibration will provide an individual data set and when matching is realized at individual level there will be individuals with different consumption expenditure values in the same household. It can be considered that, in the panel it is individuals who are followed and not the households, so this could be ignored since the persons have their own weights and although having variables with regard to household characteristics attached to them, they can be analyzed independently. However, for some individuals that are considered to be in the same household, while having all other variables, including income, of the same value; having different values for consumption would be a problematic issue.

An option to overcome this situation, could be to use integrated calibration technique to attach the same weights to each household member in a household. However, in SILC survey individuals are followed, not households. Individuals with a weight in the data set do not necessarily have the same household compositions throughout the panel. In this respect, there is no straightforward method for such calibration.

Another way to deal with this problem could be to select the households that did not have a change in their composition throughout the panel. This way an integrated

calibration could be realized and a weight could be attached to the household. This solution also comes with problems. First, there will be a definite limitation in the available data set and maybe more importantly this will lead to a biased data set where dynamic households will be ignored.

As mentioned above, there is no straightforward method for overcoming this issue, but, a method which enables the use of integrated calibration seems to be the reasonable solution. Nevertheless, to start an integrated calibration for the panel sections there is need for a household weight to calibrate. In this case, this is acquired by the mean value of total household individual weights with regard to household size. For each year, those that are a member of the household are taken into consideration disregarding whether they have a four-year weight or not.

The original panel data set (2010-2013) for individual registers is split into four. In addition, four data sets are created for each year. First of all, the data set for 2013 is created, because the weights are attached to the last year. For year 2013, the individuals which have a four-year panel weight are selected and then the households they belong to are determined. These households constitute the basis of the study. For each year, these households are selected and then individuals who are a member of these households in the respective year are selected and these individuals constitute the population in the associated year

Household distribution and structure is unique to each year. The result of this procedure is an age, sex and household size distribution in SILC section, which is the same with those of HBS. After the calibration, there is need for one more step to obtain the final weights. The resulting total population is different from HBS; the calibration enables to get the desired age, sex and household size distribution, but not the desired population. Therefore, a final rescaling step is used to adjust the matching data sets.

#### **4.4. Harmonization and other issues**

After this step, other steps suggested by van der Laan (2000) were realized. Variables and classifications were harmonized and some new variables were derived from existing variables.. Because there are no known measurement errors in the data sets, no action was taken in this regard. The variables to be used in this study did not have any missing items.

## **5. Procedure Steps**

In this section, the implemented procedures are presented. The steps are formed mainly as suggested in D’Orazio (2016). The first step involves choice of the variables (Y, Z) that are distinctly available in SILC (A) and HBS (B), in the second step all common variables (X) are identified, the choice is made among X, matching framework is decided, application is implemented and finally in the sixth step results are evaluated.

For the targeted data set, only household consumption expenditure is used as a distinct variable from HBS. Thus, Z consists of only one variable. Y will be income variable in SILC, so it also consists of one variable.

Both data sets were analyzed and a total of 40 variables were selected and created that could serve as matching variables. The categories of the variables were recoded to have compatible categories. The continuous variables were categorized and some of the variables with several categories were recategorized in order to have less categories to increase the similarities between the two data sets. The recoding is realized in a way that does not allow any missing values for any of the variables. Some variables are derived from the same variables, so are definitely exposed to multicollinearity, but they will be filtered in the following phases according to their explanatory power in the models.

The first step to choose among the common variables is to analyze the distributions in both data sets. It is important to have matching variables with similar distributions. A simple comparative analysis of the distributions could be performed (see the appendix for the distributions), but such a comparison is better performed with a distance function, which enables to detect the similarities and differences between the distributions of variables. In this study Hellinger Distance (HD) is used as suggested by (Donatiello, 2014).

$$HD (P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^k (\sqrt{p_i} - \sqrt{q_i})^2} ,$$

where p and q are the respective percentages of frequencies for each category. The indicator takes a value between 0 and 1. 0 indicates a perfect similarity and 1 indicates exact dissimilarity.

There is also need for a selection criterion of a good fit. In literature, 5 per cent is mostly used as a cutoff line (Donatiello, 2014). Variables with HD that are less than 5 per cent are considered to have very similar distributions.

Hellinger Distance is a simple tool to be used for detection of similarities of variables between two data sets; on the other hand, it does not consider the sampling design. For this purpose, as indicated by Leulescu and Agafitei (2013) other tests, such as, Chi square, Kolmogorov Smirnov, Rao-Scott, Wald-Wolfowitz tests could be used, but because these tests require relevant variables with regard to sampling design, they cannot be used in this study.

The analyses with HD demonstrates that four variables have a Hellinger distances that are higher than 5 per cent for all years of the panel. These variables are the reference person's number of weekly working hours, reference person's economic activity at work, number of self-employed persons in the household and the lowest monthly income considered by the household to make ends meet. Among these, the reference person's economic activity at work is recoded into four categories as agriculture, manufacturing construction and service sectors instead of 18 main categories and this way the Hellinger distance is smaller than 5 per cent for all years and therefore could be used in the further stages. The other three variables have different distributions probably due to poor data collection in either or both data sets, and are not considered as matching variables.

In addition, two other variables have a value greater than 5 per cent for two years. These variables are the tenure status of the household and rent (either paid or imputed). The tenure status is recoded into two main categories as owning the dwelling that is lived in or not; and the rent is recoded into two categories instead of four, by collapsing the last three categories into one category. These recoding operations allow the distributions for both variables to have Hellinger distances smaller than 5 per cent, so these variables are eligible for further analysis. After the HD analyses three variables are omitted and there are a total of thirty-seven variables remaining that are considered to have similar distributions.

There should be further selection among the common variables to perform the statistical matching. In this regard, `spearman2` function in `Hmisc` package (Harrell, 2016) in R is used to observe the pairwise correlations between the response variables (income in SILC and consumption expenditure in HBS) and the common variables, and selection is made among the common variables.

Afterwards, dummy variables are created for the previously selected 11 variables and regression is run for all on the log of consumption expenditure in HBS. The total explanatory power of these variables is  $\text{adj-R}^2=0.5138$  (see appendix for the tables). After deselecting ownership of dishwasher and internet, heat system and dwelling type from the model, explanatory power only decreases to 0.5052. Therefore, these four variables can be omitted from the model. Similar results are obtained for all years and also when income (adjusted with CPI) is regressed against the same regressors in SILC.

When adjusted disposable income classes are introduced in the consumption expenditure model in HBS, for 2010, the explanatory power increases up to 0.6273 with all other regressors included. Even when only computer, car ownerships, rent categories and hot water availability in the dwelling are kept as regressors with the income categories the adjusted  $R^2$  is 0.6034.

Having less number of matching variables is preferable for the quality of the match since as the number variables are higher the procedure is exposed to complications more (Kum and Masterson, 2008). Among these imputed rent, which is used to form rent categories, is known to be calculated with a model, so this is also dropped from the model and besides disposable income categories, computer, car ownerships and hot water availability are kept as final regressors. The adjusted  $R^2$  in this model is 0.5950. Thus, without losing from the explanatory power it was possible to omit rent categories.

The first choice regarding the matching framework is to choose between micro and macro level matching. The micro approach enables to obtain a synthetic data set, which is the case in this study. The macro approach only allows for certain contingency tables and correlations between the variables.

Also, a choice should be made between parametric, nonparametric and mixed methods. The parametric method enables use of a model where relationships can be estimated among variables with parametrical indicators. On the other hand, model misspecification would cause further problems. In addition to that, the nonparametric method allows for use of live values. In this regard, use of a mixed method, which involves both parametric and nonparametric methods, sustain the advantages of these approaches concurrently (D'Orazio, 2016). As mentioned before the findings of Webber and Tonkin (2013) also suggest that use of a mixed method yields better results.

The mixed method is comprised of two consequent steps. In the first step a model is fit in each data set and in the second step, by making use of parameters from the first step, the two data sets are matched with nonparametric matching methods.

A choice should be made between the data sets to determine which one will be the recipient and which one will be the donor file. In this study, the main target is to match consumption expenditure into longitudinal SILC data set. Therefore the recipient is the SILC data set and the donor is HBS.

When dealing with the SM problem, one of the very first issues is the Conditional Independence Assumption (CIA). The mutually exclusive variables in the two data sets are assumed to be independent since no information on this issue can be deduced from the matching data sets. It is a strong assumption and rarely holds in reality as also suggested by D'Orazio (2016). This assumption can be relaxed and a SM of better quality could be realized if there is any information available in another data set where the variables in question can be found simultaneously or there is any other source of information is available suggesting the correlation of the variables. In our case, there is existing information on the correlation of income and consumption expenditure from HBS data where these two variables are available at the same time. Then, this information will be used in the StatMatch package and CIA will be relaxed.

Both of the data sets that are used in this study for SM have complex sampling designs. D'Orazio (2016) suggests two approaches for dealing with the complex survey design issue. One way to deal with the issue is the naïve approach where in principle the sampling design and the weights are ignored, and the other one is explicitly taking into account the complex survey design and the survey weights.

There are mainly three methods for overcoming the complex design issue when the decision is to explicitly take the complex design into consideration. Renssen's calibrations based approach (Renssen, 1998), Rubin's file concatenation ([Rubin, 1986](#)), and Wu's approach based on empirical likelihood methods (Wu, 2004). Renssen's (1998) method is employed in in this study. The method is based on calibration technique to obtain consistency between population totals with regard to the variables at hand. This method requires use of only a number of continuous joint variables (X). It also allows one of the mutually exclusive variables (in Y or Z) to be continuous.

Donatiello et al. (2015) studied the extension of the use of Renssen's method to continuous variables. They used a two-step procedure. In the first step, they predicted consumption in Italian SILC (IT-SILC) by applying a linear model taking into consideration the survey harmonized weights and in the same way, they predicted consumption in HBS. In the second step, they performed a nearest neighbor distance hot

deck procedure on these predictions and imputed the “observed” values for consumption into IT-SILC.

The advantages of Renssen method are several. First, it starts from available data and weights and harmonizes marginal and joint distributions of the matching variables. It provides a synthetic data set that preserves the marginal distribution of the imputed variable and its joint distributions with the matching variables. It also allows introducing auxiliary data sources easily. (Donatiello et al., 2015)

On the other hand, Renssen’s method has a few weaknesses. First, there is a probability that the calibration fails. Another issue in this regard is that heteroskedasticity and residuals are not normally distributed.

## **6. Procedure and Results**

Before the above-mentioned two steps, there is need for harmonizing the matching variables. In this case, the harmonization is realized for the joint distributions of four variables. The harmonization actually means calibrating the two data sets jointly with respect to the four matching variables by adjusting the weights of the data sets accordingly. It is similar to the procedure conducted previously in order to make the two data sets coherent with respect to age, sex and household size of the population. It is true that this second procedure will dislocate the first one, but necessary operations were carried out to select the matching variables between the two procedures. Since there is need for adjusting the four variables and adding others (age, sex, and household size) would be too demanding and would result much poorer results. In addition, there is no need for age, sex and household size to be recalibrated because they will not be used in the matching process. Even if recalibration with all seven variables were conducted, because of too many constraints, the weights would have extreme values, which would inflate the variation of any estimation (Kish, 1965). In order to avoid further complications the procedure will be carried out for the four common variables as suggested by D’Orazio (2016).

The first analysis in Table 6 for 2010 data shows the overlapping of the two data sets with respect to the four common variables by making use of various indicators. The indicators suggest that the data sets are already in good harmony with respect to those four variables.

**Table 6.**

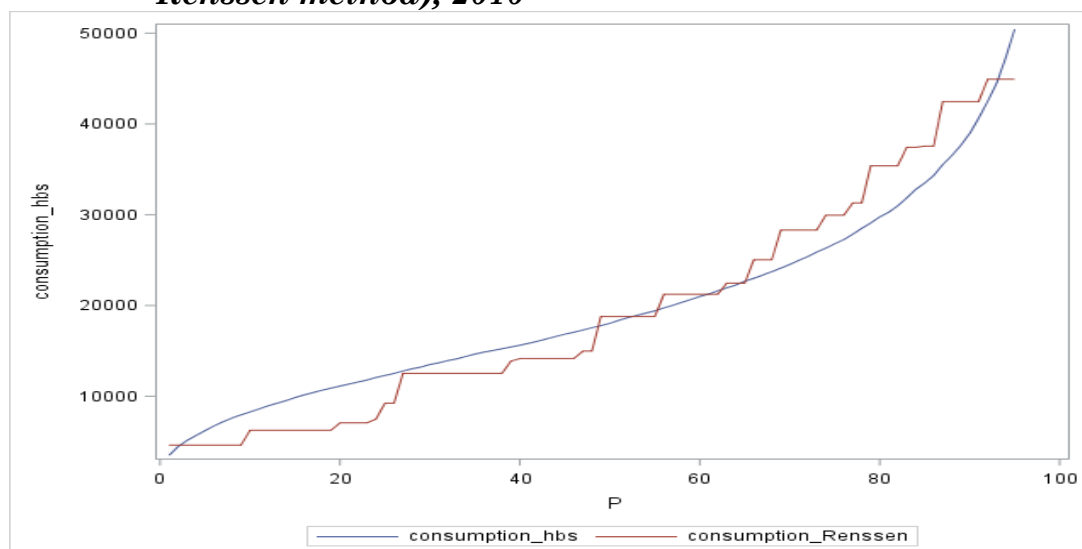
<b>tv</b>	<b>d overlap</b>	<b>Bhatt</b>	<b>Hell</b>
0.059	0.941	0.996	0.064

All the same, further harmonization will be looked for by using “harmonization” function in StatMatch package. After the matching variables are harmonized to the highest possible proximity by “harmonization” function, Renssen (1998) method is applied. In this case, the survey weights are calibrated in order to reach the targeted aim.

In the function, there is an opportunity for using auxiliary information if available. In our case because there is available information in HBS data set regarding the relationship between consumption expenditure and income as well as other common variables we make use of it. All variables including income categories are set as common variables in SILC and in HBS. Only consumption expenditure is the extra one. Among the methods, Synthetic Two-Way Stratification is the one that is suitable to the data sets at hand, so this is used.

After the results are obtained with the Renssen method, a comparison of cumulative density functions of consumption expenditure is made between the two data sets, original HBS data set and SILC after Renssen method is applied. A glance at the figure shows the overall distributions in the two data sets are similar although the distribution acquired by Renssen method does not demonstrate a smooth graph and the observations are gathered at specific values.

**Figure 9. Cumulative distribution of consumption expenditure (HBS and Renssen method), 2010**



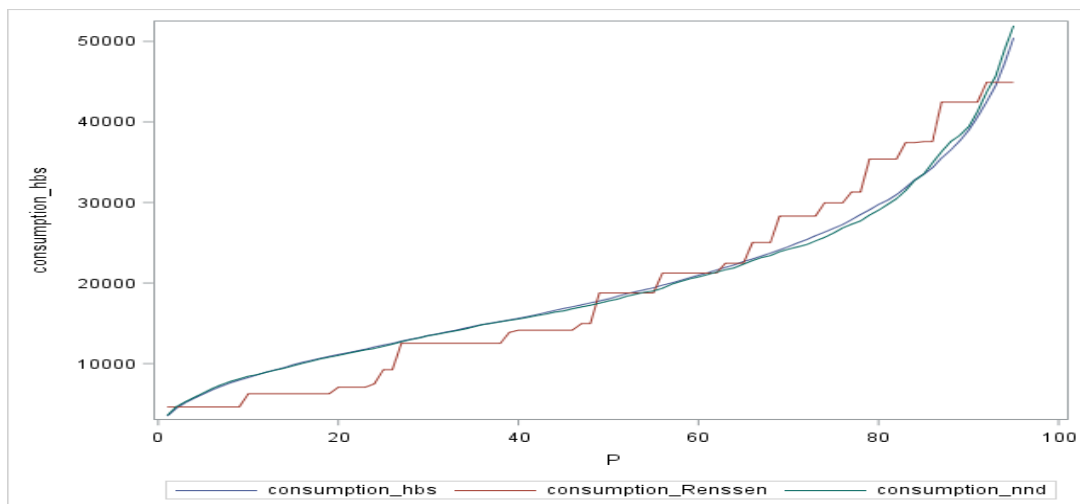


For a more approximate distribution of the matched variable (consumption expenditure), one other step is required as suggested by Donatiello et al. (2015). In this step Donatiello et al. (2015) suggests use of estimates from the Renssen method application. In the Renssen application, consumption expenditure is estimated in both SILC and HBS, by making use of common variables and auxiliary information. Then these estimates are used in this step by nearest neighbor distance function in order to get more approximate distributions among the data sets. In our case because we already have the live values of consumption expenditure from HBS, in addition to all other variables that are existent in the SILC data set to be matched, it is preferable to use these live values instead of estimates.

First, the function was applied with constrained model. In this case, a register from the donor data set can be used only for once. The results indicated huge differences between the estimate in the recipient file and the matched value. This time, unconstrained model was applied. In this case, a register from the donor file could be matched to the recipient file more than once.

Figure 10 shows the comparison of cumulative distributions of consumption expenditure in HBS, after the application of Renssen method and after the finalization with NND method with unconstrained matching. The figure shows that the distribution in the final synthetic data set is quite similar to the original distribution (also, the numbers of those having consumption expenditure higher than their disposable income are very close). The results are similar for all years.

**Figure 10. Cumulative distribution of consumption expenditure  
(HBS, Renssen method and NND), 2010**



## **7. Validation (Quality control)**

The quality of the synthetic data file obtained by the SM procedure determines whether the goal of the effort is achieved. The quality depends on two main conditions. The marginal and joint distributions in the synthetic file should be as close as possible to the respective distributions in the original files (Kum and Masterson, 2008). Special attention is given to the relationship between income and consumption expenditure.

Main indicators created by consumption expenditure, such as poverty measures are also used to validate the quality of the matching procedures. These indicators are created in both HBS and the synthetic data files and compared according to different household characteristics.

Rässler (2002) proposes a framework for the evaluation of quality in a statistical matching procedure. She establishes four levels of validity for a matching procedure: *(1) the marginal and joint distributions of variables in the donor sample are preserved in the statistical matching file; (2) the correlation structure and higher moments of the variables are preserved after statistical matching; (3) the true joint distribution of all variables is reflected in the statistical matching file; (4) the true but unknown values of the Z variable of the recipient units are reproduced.*

The quality control is realized in three main steps. In the first step, marginal distributions of matching variables are compared. In the second step, joint distributions matching variables and consumption groups are compared. And in the final step, poverty head count ratios are compared at household size level.

As will be seen at the end of this section, quality control results referred to a change in the model. With regard to marginal and joint distributions of variables and with regard to overall poverty indicators there weren't significant differences between the synthetic file and HBS. On the other hand, poverty head count ratios at household size level were significantly different from each other. Therefore the model was changed and the statistical matching procedure was repeated. The results provided below refer to the finalized data set. The differences between the two models are insignificant for the first two analyses. On the other hand, there are major differences in the third analysis, therefore a comparison of the two matching models are provided for this analysis.

### 1. Marginal distributions of matching variables:

**Table 9. Availability of car, computer and hot water; and disposable income categories**

		Synthetic Data set				HBS			
		2010	2011	2012	2013	2010	2011	2012	2013
Car	1	31.9	33.5	36.5	38.5	32.0	33.4	36.4	38.7
	2	68.1	66.5	63.5	61.5	68.1	66.6	63.6	61.4
Hot water	1	82.1	83.6	85.4	86.9	82.8	83.6	85.0	87.6
	2	17.9	16.4	14.6	13.1	17.2	16.4	15.0	12.4
Comp	1	42.4	45.6	49.3	49.5	42.1	45.0	49.1	49.3
	2	57.6	54.4	50.7	50.5	58.0	55.0	50.9	50.7
Dis. inc. cat.	1	24.2	18.4	13.8	10.6	24.1	18.2	13.7	10.9
	2	23.5	20.9	19.1	17.4	23.4	21.0	19.0	16.7
	3	17.4	17.4	17.5	17.0	17.5	17.2	17.4	16.6
	4	11.2	13.2	14.0	13.9	11.4	13.3	13.9	13.6
	5	8.2	9.3	10.4	10.7	8.3	9.4	10.5	11.0
	6	15.5	20.8	25.3	30.4	15.3	21.0	25.6	31.2

### 2. Joint distributions of matching variables with consumption categories

For all years, the joint distributions are similar between the synthetic data set and HBS. Only the results for 2010 are presented for demonstration as no significant differences were observed between years.

**Table 10. Matching variables by consumption categories, 2010**

		<u>Synthetic Data set</u>			<u>HBS</u>		
		<u>Consumption categories</u>			<u>Consumption categories</u>		
		<u>&gt;=1500</u>	<u>1500-3000</u>	<u>&gt;3000</u>	<u>&gt;=1500</u>	<u>1500-3000</u>	<u>&gt;3000</u>
<b>Car</b>	<b>1</b>	14.4	41.4	77.8	14.6	41.5	72.5
	<b>2</b>	85.6	58.6	22.2	85.5	58.5	27.5
<b>Hot water</b>	<b>1</b>	72.3	90.8	96.9	72.4	92.1	96.3
	<b>2</b>	27.7	9.2	3.1	27.6	7.9	3.7
<b>Comp</b>	<b>1</b>	22.9	57.6	78.5	21.8	56.6	79.0
	<b>2</b>	77.1	42.4	21.5	78.2	43.4	21.0
<b>Dis. inc. cat.</b>	<b>1</b>	43.5	5.1	1.0	44.0	5.3	1.0
	<b>2</b>	32.5	17.1	4.8	33.4	16.7	4.2
	<b>3</b>	14.8	24.2	7.4	14.3	25.1	7.4
	<b>4</b>	5.1	19.9	9.7	4.6	20.5	11.3
	<b>5</b>	2.3	15.0	12.5	2.4	14.5	13.2
	<b>6</b>	1.8	18.8	64.5	1.3	18.0	63.0

### **3. Poverty measures by household size**

Overall poverty head count ratios are not very different from each other. On the other hand, especially for single households the differences are substantial. For households with 1 and 2 members the poverty ratios are underestimated demonstrating overestimation for such households. On the other hand, households with 3 or more members generally have higher poverty ratios in the synthetic file compared to HBS, which indicates an underestimation for such households.

Main reason for this substantial divergence at household breakdown could be the lack of household size among the matching variables. For this purpose, household size was included among the matching variables instead of availability of hot water and estimation was repeated for the data. Availability of hot water was chosen because it was the least related to income and consumption compared to other matching variables. Although household size demonstrated even lower relationship in the pairwise analyses, after the divergence of poverty head count ratio between HBS and the synthetic file was observed at household size breakdown, it was decided to include it in the matching variables. This condition appears as a necessity when the target is to carry out further research on poverty

measures based on the matched consumption expenditure, especially at household size breakdown.

The results with the new model generated good results with respect to poverty head count ratio comparison between HBS and the synthetic file. Now the estimates are more reliable, and could be used for further analyses with more confidence.

**Table 17. Poverty head count ratios by household size, 2010-2013**

Year	HHS	hbs	cons_M1	cons_M2
2010	1	14.8	6.9	16.2
2010	2	11.3	7.4	9.2
2010	3	10.5	9.9	11.6
2010	4	15.5	19.9	16.7
2010	5	39.5	45.0	42.5
2010	Total	19.9	21.2	20.9
2011	1	14.6	5.6	15.4
2011	2	8.9	5.9	7.0
2011	3	8.6	12.2	9.1
2011	4	15.2	14.9	16.0
2011	5	35.0	37.6	36.0
	Total	17.6	17.8	17.8
2012	1	14.1	4.6	13.9
2012	2	11.3	8.5	10.1
2012	3	9.2	9.3	11.2
2012	4	11.5	15.9	13.8
2012	5	34.2	37.7	32.5
	Total	16.6	17.4	17.0
2013	1	13.0	7.1	12.6
2013	2	10.3	6.9	7.3
2013	3	7.8	10.6	8.3
2013	4	11.9	14.7	14.1
2013	5	32.4	39.9	33.7
2013	Total	15.5	17.6	15.9

## **8. Conclusion**

This study aims to create a consumption expenditure variable in longitudinal SILC survey data set via statistical matching of SILC and HBS. Conditional independence assumption is considered to be confirmed by use of auxiliary information. Income variable which is also available in HBS serves in this regard.

The study used the approach by D'Orazio (2016) and its extension by Donatiello et al. (2015) where the procedure is extended to continuous variables. StatMatch R package is used for the matching procedure. The matching procedure actually consists of two main steps. In the first one statistical matching is realized with Renssen (1998) methodology. In the second step, nearest neighbor distance function is applied to the results achieved in the first step to get the final results.

The first results indicated a good match at aggregated levels. On the other hand, poverty head count ratios were substantially divergent at household size breakdown. In this regard, household size was substituted into the matching variables instead of hot water availability. The results improved to a considerable extent.

This showed that even if household size is not selected in the first place as a matching variable because it is not one of the best predictors of the response variables, it should definitely be added among the matching variables, if the target is to pursue further study at disaggregated level.

For further research on the statistical matching of consumption expenditure, another option could be to use equivalized measures of consumption expenditure and income. Hereby household size will be intrinsic in the matched variables. This way an extra variable could be added to increase the quality of the match.

## **References**

Baldini M., Pacifico, D., Termini, F. (2015). "Imputation of missing expenditure information in standard household income surveys, "Department of Economics 0049, University of Modena and Reggio E., Faculty of Economics "Marco Biagi".

Betti, G. (1998), "Intertemporal equivalence scales and cost of children using BHPS", ERSC Research Centre on Micro-social Change Working Papers. Paper 11/98, Colchester: University of Essex.

Brodaty, T., Crépon, B., Fougère D. (2001). "Using Matching Estimators to Evaluate Alternative Youth Employment Programs: Evidence from France, 1986–1988." In *Econometric Evaluation of Labour Market Policies*, Michael Lechner and Friedhelm Pfeiffer (eds.). Heidelberg: Physica-Verlag.

Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M., Spaziani, M. (2014) "[Statistical Matching of Income and Consumption Expenditures](#)", *International Journal of Economic Sciences*, Vol. III, pp. 50-65.

Donatiello, G., Frattarola, D., Rizzi, A., Spaziani, M. (2015), The Role of the Available Information in Statistical Matching IT-SILC and HBS, EU-SILC best practice workshop, London, 16-17 September 2015

D'Orazio, M. (2016) "Statistical Matching and Imputation of Survey Data with StatMatch", R package vignette, [http://cran.rstudio.com/web/packages/StatMatch/vignettes/Statistical\\_Matching\\_with\\_StatMatch.pdf](http://cran.rstudio.com/web/packages/StatMatch/vignettes/Statistical_Matching_with_StatMatch.pdf)

D'Orazio, M., Di Zio, M., Scanu, M. (2006), *Statistical Matching: Theory and Practice*. John Wiley & Sons, Chichester, ISBN: 0-470-02353-8.

[Greenwood, D. T. \(1983\). "An Estimation of U.S. Family Wealth and Its Distribution from Microdata, 1973." \*Review of Income and Wealth\* 29\(1\): 23–44.](#)

[Greenwood, D. T. \(1987\). "Age, Income, and Household Size: Their Relation to Wealth Distribution in the United States." in \*International Comparisons of the Distribution of Household Wealth\*, Edward N. Wolff \(ed.\). Oxford, New York, Toronto, and Melbourne: Oxford University Press, Clarendon Press.](#)

Harrell, F. (2016), Package "Hmisc", date of access:11.06.2016. <https://cran.rstudio.com/web/packages/Hmisc/Hmisc.pdf>.

Keister, L. A., Moller, S. (2000). "Wealth Inequality in the United States." *Annual Review of Sociology* 26: 63–81.

Keister, L. A., Moller, S. (2003). "Sharing the Wealth: The Effect of Siblings on Adults' Wealth Ownership." Demography 40(3): 521–542.

Kish, L. (1965). Survey Sampling, New York: John Wiley & Sons

Kum and Masterson (2008), Statistical Matching Using Propensity Scores: Theory and Application to the Levy Institute Measure of Economic Well-Being, The Levy Economics Institute of Bard College, Working Paper No:535.

Laan, P. van der. 2000. 'Integrating Administrative Registers and Household Surveys'. Netherlands Official Statistics, Vol. 15 (Summer 2000): Special Issue, Integrating Administrative Registers and Household Surveys, ed. P.G. Al and B.F.M. Bakker, pp. 7-15.

Leulescu, A. and Agafitei, M. (2013). Statistical matching: a model based approach for data integration. Eurostat Methodologies and Working Papers.

Little, R. J., and Rubin, D. B. (2000). "Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches." Annual Review of Public Health 21(1): 121–45.

Masterson, T. (2013), "Quality of Statistical Match and Simulations Used in the Estimation of the Levy Institute Measure of Time and Consumption Poverty (LIMTCP) for Turkey in 2006," Economics Working Paper Archive wp\_769, Levy Economics Institute, [http://www.levyinstitute.org/pubs/wp\\_535.pdf](http://www.levyinstitute.org/pubs/wp_535.pdf).

Okner, B. (1972), "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File", Annals of Economic and Social Measurement 1, pp. 325-342.

Putten, P. van der, Kok, J.N. and Gupta, A. (2002). Data Fusion through Statistical Matching. MIT Sloan School of Management Working Paper No. 4342-02, Cambridge, MA.

Radner, D. B. (1981). "An Example of the Use of Statistical Matching in the Estimation and Analysis of the Size Distribution of Income." Review of Income and Wealth 27(3): 211–42.

Rässler, S. (2002). Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches. New York: Springer.

Rasner, A., J. R. Frick, and M. M. Grabka. 2011. Extending the Empirical Basis for Wealth Inequality Research Using Statistical Matching of Administrative and Survey Data. SOEP papers 359. Berlin: DIW.

Rasner, A., R. K. Himmelreich, M. M. Grabka and J. R. Frick (2007). Best of Both Worlds – Preparatory Steps in Matching Survey Data with Administrative Pension



Records. The Case of the German Socio Economic Panel and the Scientific Use File Completed Insurance Biographies 2004. SOEP papers 70. Berlin, Deutsches Institut für Wirtschaftsforschung: 210.

Renssen, R. H. (1998), “Use of Statistical Matching Techniques in Calibration Estimation”, *Survey Methodology*, 24, 171-183.

Rosenbaum, P. R., Rubin, D. B. (1983). “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70(April): 41–55.

Rubin, D. B. (1986), “Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations”, *Journal of Business and Economic Statistics*, 4, 87-94.

Rubin, D. B., Thomas, N. (1992). “Characterizing the Effect of Matching Using Linear Propensity Score Methods With Normal Distributions.” *Biometrika* 79(4): 797–809.

Rubin, D. B., Thomas, N. (1996). “Matching Using Estimated Propensity Scores: Relating Theory to Practice.” *Biometrics* 52(March): 249–264.

Simonson, J., Romeu Gordo, L., Kelle, N. (2012). Statistical matching of the German Aging Survey and the Sample of Active Pension Accounts as a source for analyzing life courses and old age incomes, *Historical Social Research*, Vol. 37 — 2012 — No. 2, 185-210.

Thiede, M., Weske, M., Mueller U. (2010). “Premature Retirement due to back disorders; Using Propensity Score Matching to combine Cross-Sectional and Longitudinal Data sets - An Example with Data from the German Research Datacenter of the Federal Pension Insurance”, *European Population Conference*, Vienna.

Turkstat, (2013a), *Handbook for Household Budget Survey*, Ankara.

Turkstat, (2013b), *Handbook for Statistics on Income and Living Conditions Survey*, Ankara.

Wagner, J. (2001). “The Causal Effects of Exports on Firm Size and Labor Productivity: First Evidence from a Matching Approach.” *Hamburgisches Welt-Wirtschafts-Archiv Discussion Paper* 155. Hamburg, Germany: Hamburg Institute of International Economics.

Webber, D., Tonkin, R. (2013), “Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation”, *Eurostat Working Paper*.

Wolff, E. (2000). “Recent Trends in Wealth Ownership, 1983–1998.” *Working Paper* 300. Annandale-on-Hudson, NY: The Levy Economics Institute of Bard College.

Wu, C. (2004). "Combining information from multiple surveys through the empirical likelihood method", *The Canadian Journal of Statistics*, 32, 112.

Zacharias, A., Masterson, T., Kim, K. (2014), "The Measurement of Time and Income Poverty in Korea". Economics Working Paper Archive , Levy Economics Institute, [http://www.levyinstitute.org/pubs/rpr\\_8\\_14.pdf](http://www.levyinstitute.org/pubs/rpr_8_14.pdf).

## ***Appendix 1. List of common variables***

<b>#</b>	<b><u>Name of Variable</u></b>	<b><u>Variable</u></b>
1.	hsize	Household size
2.	num_ch	Number of children (0-17) in the household
3.	num_adu	Number of adults (18-64) in the household
4.	num_eld	Number of elderly (65+) in the household
5.	num_wom	Number of women in the household
6.	all_adu	All household members are adults
7.	all_eld	All household members are elderly
8.	all_wom	All household members are women
9.	num_emp	Number of employed people
10.	num_emp_inc	Number of individuals with employee income
11.	num_self_emp_inc	Number of individuals with self-employed income
12.	num_ret_inc	Number of individuals with retired income
13.	ref_sex	Reference person's sex
14.	ref_age	Reference person's age group
15.	ref_mar	Reference person's marital status
16.	ref_edu	Reference person's education
17.	ref_pro	Reference person's professional status
18.	ref_occ	Reference person's occupation
19.	ref_eco	Reference person's economic activity of work
20.	ref_whrs	Reference person's number of weekly working hours
21.	dwe	Dwelling type
22.	tenure	Tenure status
23.	rent_cat	Current rent related to occupied dwelling (including imputed rent)
24.	room_num	Number of rooms (except for kitchen, bathroom and toilet) available to the household
25.	tot_ar	Total space available to the household (m2)
26.	heat_sys	Heating system of the dwelling
27.	bath	Bath or shower in dwelling
28.	toilet	Indoor flushing toilet for sole use of household
29.	piped_wat	Piped water
30.	hot_wat	Hot water
31.	mobile	Mobile
32.	comp	Computer
33.	internet	Internet
34.	wash_m	Washing machine
35.	refrig	Refrigerator
36.	dish_w	Dishwasher
37.	air_con	Air conditioner
38.	car	Car
39.	low_mon_inc	Lowest monthly income to make ends meet
40.	dis_inc_cat	Total disposable household income

## Appendix 2. Distributions of common variables

#	NAME OF VARIABLE	VARIABLE	SILC (%)				HBS (%)			
			2010	2011	2012	2013	2010	2011	2012	2013
1	hsize	Household size								
1			6.1	6.3	6.9	6.9	6.1	6.3	6.9	6.9
2			18.0	19.5	19.6	20.5	18.0	19.5	19.6	20.5
3			23.5	23.1	23.8	23.5	23.5	23.1	23.8	23.5
4			26.0	25.1	25.4	25.7	26.0	25.1	25.4	25.7
5+			26.4	26.0	24.3	23.5	26.4	26.0	24.3	23.5
2	num_ch	Number of children (0-17) in the household								
0			41.0	42.4	42.1	42.7	40.0	41.2	42.5	43.2
1			22.4	22.6	23.4	23.3	24.1	23.9	23.8	23.1
2			22.1	20.9	22.0	21.5	21.7	21.7	20.6	21.2
3			9.3	9.0	7.6	7.9	8.4	8.1	8.1	8.0
4+			5.3	5.1	4.9	4.7	5.7	5.1	5.0	4.5
3	num_adu	Number of adults (18-64) in the household								
0			6.9	6.9	7.5	7.3	6.3	6.5	6.7	6.9
1			8.4	8.8	8.9	9.8	8.9	9.2	9.9	9.4
2			52.2	52.3	53.0	52.1	52.2	52.9	52.5	53.4
3			18.9	18.7	17.3	18.1	18.6	17.7	18.1	17.5
4+			13.5	13.5	13.3	12.6	14.0	13.8	12.9	12.8
4	num_eld	Number of elderly (65+) in the household								
0			79.6	79.6	79.9	80.0	79.5	79.6	79.6	80.0
1			14.3	14.2	13.8	13.7	14.5	14.1	14.5	13.8
2+			6.2	6.2	6.3	6.4	6.1	6.3	5.9	6.3
5	num_wom	Number of women in the household								
0			2.5	2.7	3.0	3.1	2.6	3.4	3.3	3.3
1			41.1	42.2	42.7	43.0	42.2	41.6	43.5	44.4
2			32.9	32.3	33.2	33.4	31.5	31.8	31.4	31.1
3			15.2	14.5	13.9	13.4	14.9	14.9	14.1	13.2
4+			8.3	8.3	7.3	7.1	8.9	8.4	7.7	8.0
6	all_adu	All household members are adults								
No			73.4	72.2	72.4	71.9	73.8	72.7	72.3	71.2
Yes			26.6	27.8	27.6	28.1	26.2	27.3	27.7	28.8
7	all_eld	All household members are elderly								
No			93.2	93.2	92.6	92.8	93.8	93.7	93.4	93.2
Yes			6.8	6.8	7.4	7.2	6.2	6.3	6.7	6.8
8	all_wom	All household members are women								
			94.0	93.8	93.4	93.4	93.8	93.7	93.6	93.6
			6.0	6.2	6.6	6.6	6.2	6.3	6.4	6.4

#	NAME OF VARIABLE	VARIABLE	SILC (%)				HBS (%)			
			2010	2011	2012	2013	2010	2011	2012	2013
<b>9</b>	<b>num_emp</b>	<b>Number of employed people</b>								
	0		20.2	19.7	19.0	20.2	18.5	17.4	17.2	17.4
	1		44.8	44.9	44.9	45.7	44.3	43.4	43.7	44.1
	2		25.4	25.7	26.7	25.8	27.2	28.3	28.7	28.1
	3		6.5	6.6	6.5	5.6	6.9	7.5	7.6	7.2
	4+		3.2	3.0	2.9	2.7	3.2	3.4	2.8	3.2
<b>10</b>	<b>num_emp_inc</b>	<b>Number of individuals with employee income</b>								
	0		35.1	33.5	32.3	31.9	40.7	38.0	31.2	31.7
	1		44.1	41.9	43.0	42.6	40.4	40.7	40.9	40.3
	2		16.6	19.6	19.5	20.3	15.4	17.6	21.7	21.8
	3		3.2	3.7	4.1	4.4	2.8	3.0	4.8	4.7
	4+		0.9	1.4	1.1	0.8	0.8	0.8	1.4	1.4
<b>11</b>	<b>num_self_emp_inc</b>	<b>Number of individuals with self-employed income</b>								
	0		68.3	67.5	69.2	69.3	83.6	84.9	80.0	79.3
	1		30.1	30.4	29.0	28.8	15.7	14.4	18.7	19.2
	2+		1.6	2.2	1.9	1.9	0.7	0.7	1.4	1.4
<b>12</b>	<b>num_ret_inc</b>	<b>Number of individuals with retired income</b>								
	0		66.9	66.7	66.9	66.4	67.9	67.4	67.6	68.1
	1		27.7	28.0	27.7	28.1	28.0	28.7	28.0	27.6
	2+		5.3	5.3	5.4	5.5	4.2	4.0	4.4	4.2
<b>13</b>	<b>ref_sex</b>	<b>Reference person's sex</b>								
	Male		84.5	84.6	83.9	83.3	83.9	84.1	83.6	84.3
	Female		15.5	15.4	16.2	16.7	16.1	15.9	16.4	15.7
<b>14</b>	<b>ref_age</b>	<b>Reference person's age group</b>								
	<25		5.7	5.9	4.9	5.6	5.8	5.2	5.6	5.5
	>=25 and <35		26.3	24.6	24.3	24.0	25.6	25.8	25.2	25.0
	>=35 and <45		25.5	26.1	25.7	24.8	26.1	26.2	25.7	25.5
	>=45 and <65		31.3	32.4	33.4	34.0	31.7	31.9	32.2	32.8
	>=65		11.3	11.1	11.8	11.6	11.0	10.9	11.3	11.2
<b>15</b>	<b>ref_mar</b>	<b>Reference person's marital status</b>								
	Married		81.1	81.5	81.2	80.7	81.8	80.7	80.7	80.6
	Never married		9.1	8.8	8.2	8.8	8.8	9.4	9.1	9.3
	Widow		7.5	7.3	8.0	7.7	7.1	7.0	7.0	6.9
	Divorced		2.2	2.5	2.6	2.8	2.4	3.0	3.2	3.1

#	NAME OF VARIABLE	VARIABLE	SILC (%)				HBS (%)			
			2010	2011	2012	2013	2010	2011	2012	2013
<b>16</b>	<b>ref_edu</b>	<b>Reference person's education</b>								
	No formal education		11.6	11.4	10.9	10.3	11.0	10.8	10.8	10.5
	Less than high school		54.9	54.4	53.4	53.9	55.8	54.6	52.6	53.5
	High school		19.0	19.4	20.0	19.8	19.1	19.2	19.4	18.7
	Higher education		14.5	14.8	15.6	16.0	14.2	15.4	17.2	17.3
<b>17</b>	<b>ref_pro</b>	<b>Reference person's professional employment status</b>								
	Doesn't work		26.3	25.7	25.1	26.9	24.6	22.9	22.9	23.6
	Regular employee		44.2	44.8	46.3	45.9	43.8	44.9	45.6	45.8
	Casual employee		5.5	5.8	5.8	5.5	6.4	7.2	6.6	5.8
	Employer		5.3	5.2	5.4	5.3	4.4	4.3	4.6	4.4
	Own account worker		18.1	17.9	17.0	16.0	20.4	20.5	19.9	20.0
	Unpaid family worker		0.7	0.6	0.4	0.4	0.4	0.3	0.3	0.4
<b>18</b>	<b>ref_occ</b>	<b>Reference person's occupation (ISCO-88)</b>								
	Doesn't work		26.3	25.7	25.1	26.9	24.6	22.9	22.9	23.6
	Legislators, senior, officials and managers		8.9	8.6	5.4	5.1	10.4	10.3	7.2	6.6
	Professionals		6.3	6.2	7.0	7.0	5.3	5.9	6.8	6.8
	Technicians and associate professionals		4.3	4.9	4.3	4.2	4.9	5.4	4.7	4.2
	Clerks		4.0	4.1	4.1	4.2	4.2	4.4	4.1	4.2
	Service workers and shop and market sales workers		8.3	8.3	13.0	13.0	7.6	8.5	12.9	14.0
	Skilled agricultural, and fishery workers		11.7	11.7	11.4	10.5	11.8	12.8	12.0	12.0
	Craft and related trades workers		12.2	12.5	12.7	12.2	12.2	10.7	12.0	11.9
	Plant and machine operators and assemblers		9.2	9.2	8.8	9.3	9.9	9.4	9.5	8.7
	Elementary occupations		8.9	8.9	8.3	7.6	9.2	9.7	7.8	8.0
<b>19</b>	<b>ref_eco</b>	<b>Reference person's economic activity of work (NACE Rev.2)</b>								
	Doesn't work		26.3	25.7	25.1	26.9	24.6	22.9	23.0	24.0
	Agriculture, forestry and fishing (A)		12.9	12.5	12.0	11.3	12.7	13.8	12.9	12.5
	Mining and quarrying (B)		0.6	0.7	0.7	0.6	0.5	0.4	0.6	0.9
	Manufacturing (C)		14.7	14.8	15.1	15.2	15.4	14.2	15.0	14.0
	Electricity, gas, steam, water supply, sewerage etc. (D+E)		1.2	1.0	1.0	1.1	0.8	0.7	0.8	1.1
	Construction (F)		5.5	6.1	6.0	6.4	6.3	7.1	6.9	6.1
	Wholesale and retail trade (G)		11.5	12.0	11.8	10.9	12.2	12.1	10.9	11.5
	Transportation and storage (H)		4.5	4.5	4.9	4.3	3.5	3.8	3.5	4.0
	Accommodation and food service activities (I)		3.4	3.3	3.0	3.1	4.9	5.1	4.6	4.4
	Information and communication (J)		0.7	0.6	0.9	0.9	0.6	0.8	0.6	0.7
	Financial and insurance activities (K)		1.3	1.3	1.2	1.0	0.8	1.0	1.2	1.2
	Real estate activities (L)		0.2	0.6	0.3	0.3	0.2	0.7	0.9	0.6
	Professional, scientific and technical activities (M)		1.1	1.4	1.2	1.1	1.3	1.6	1.5	1.7
	Administrative and support service activities (N)		3.0	2.7	3.5	3.2	2.5	2.4	2.4	2.7
	Public administration and defence (O)		4.4	4.2	4.3	4.2	5.5	5.4	6.0	5.7
	Education (P)		3.8	3.6	4.4	4.3	3.4	3.4	3.9	3.6
	Human health and social work activities (Q)		2.0	2.2	2.3	2.5	2.0	1.9	2.5	2.7
	Arts, entertainment and recreation (R)		0.5	0.5	0.4	0.6	0.3	0.3	0.3	0.4
	Other social, community and personal service activities (S+T+U)		2.6	2.3	2.2	2.1	2.7	2.5	2.5	2.3

#	NAME OF VARIABLE	VARIABLE	SILC (%)				HBS (%)			
			2010	2011	2012	2013	2010	2011	2012	2013
<b>20</b>	<b>ref_whrs</b>	<b>Reference person's number of weekly working hours</b>								
	Doesn't work		26.3	25.7	25.1	26.9	24.6	22.9	22.9	23.6
	<20		1.7	1.5	1.6	1.4	61.5	63.3	63.5	62.8
	>=20 and <40		19.1	17.7	18.4	17.8	11.9	12.2	11.8	11.9
	>=40 and <60		36.9	37.8	37.4	37.6	2.0	1.6	1.7	1.7
	>=60		16.1	17.3	17.5	16.3	0.1	0.1	0.1	0.0
<b>21</b>	<b>dwe</b>	<b>Dwelling type</b>								
	Detached or semidetached		46.3	44.7	43.4	42.9	44.4	44.1	43.6	44.4
	Apartment		53.7	55.3	56.6	57.1	55.6	55.9	56.5	55.6
<b>22</b>	<b>tenure</b>	<b>Tenure status</b>								
	Owner occupied		60.0	59.9	59.9	59.7	60.0	60.2	57.5	59.8
	Rented		21.1	21.3	20.6	21.4	23.8	23.9	24.9	23.3
	Owned by governmental or private organizations		1.3	1.3	1.1	1.0	2.0	2.0	2.3	2.4
	Not owner occupied, but no rent is paid		17.7	17.5	18.4	18.0	14.2	13.8	15.3	14.5
<b>23</b>	<b>rent_cat</b>	<b>Current rent related to occupied dwelling (including imputed rent)</b>								
	<250		49.5	43.5	38.1	43.1	46.9	43.4	38.6	36.9
	>=250 and <500		36.2	42.0	47.8	44.6	42.8	44.2	44.5	42.8
	>=500 and <750		10.0	10.6	10.7	8.7	7.3	8.2	11.3	13.5
	>=750		4.3	3.9	3.5	3.6	3.0	4.3	5.7	6.8
<b>24</b>	<b>room_num</b>	<b>Number of rooms available to the household</b>								
	1		1.0	1.0	0.9	0.8	1.0	0.9	0.8	0.9
	2		9.1	8.8	8.3	8.1	8.0	8.2	8.8	7.8
	3		42.4	41.8	42.0	42.2	39.8	41.0	38.7	40.4
	4+		47.6	48.4	48.9	48.9	51.2	49.8	51.7	50.9
<b>25</b>	<b>tot_ar</b>	<b>Total space available to the household (m<sup>2</sup>)</b>								
	<=60		7.8	7.6	7.1	6.9	7.5	7.3	7.6	7.7
	>60 and <=80		17.3	16.8	16.9	17.0	15.7	15.8	14.7	14.8
	>80 and <=100		38.9	38.5	37.7	37.2	36.3	36.8	36.2	35.4
	>100 and <=120		20.0	20.2	20.6	21.0	22.1	21.3	21.5	20.3
	>120		16.0	16.9	17.8	17.9	18.3	18.8	20.1	22.0
<b>26</b>	<b>heat_sys</b>	<b>Heating system of the dwelling</b>								
	Stove (coal, gas, natural gas, electricity, etc.)		62.9	59.2	56.7	53.7	65.3	60.1	56.3	55.6
	Central heating for one or more buildings		10.1	10.3	10.8	10.5	9.7	10.6	11.2	10.9
	Central heating for one dwelling		22.9	26.0	29.0	32.3	20.8	25.5	28.1	29.0
	Air conditioner		4.0	4.2	3.3	3.2	3.8	3.5	4.3	4.5
	Other		0.2	0.2	0.2	0.2	0.4	0.3	0.1	0.1
<b>27</b>	<b>bath</b>	<b>Bath or shower in dwelling</b>								
	Yes		95.9	97.5	97.8	97.9	96.5	96.8	97.0	97.3
	No		4.1	2.5	2.2	2.1	3.5	3.2	3.0	2.7

#	NAME OF VARIABLE	VARIABLE	SILC (%)				HBS (%)			
			2010	2011	2012	2013	2010	2011	2012	2013
<b>28 toilet</b>		<b>Indoor flushing toilet for sole use of household</b>								
	Yes		89.2	92.7	92.9	93.3	89.8	90.0	91.1	91.5
	No		10.8	7.3	7.1	6.7	10.2	10.0	8.9	8.5
<b>29 piped_wat</b>		<b>Piped water</b>								
	Yes		98.0	98.4	98.9	99.0	98.7	99.0	98.6	99.5
	No		2.0	1.6	1.1	1.0	1.3	1.0	1.4	0.6
<b>30 hot_wat</b>		<b>Hot water</b>								
	Yes		82.1	83.6	85.5	87.1	82.8	83.6	85.0	87.6
	No		17.9	16.4	14.5	12.9	17.2	16.4	15.0	12.4
<b>31 mobile</b>		<b>Mobile</b>								
	Yes		94.0	94.9	95.4	95.7	93.9	94.1	94.6	95.8
	No		6.0	5.1	4.6	4.3	6.1	5.9	5.4	4.2
<b>32 comp</b>		<b>Computer</b>								
	Yes		43.5	47.9	50.1	50.4	42.1	45.0	49.1	49.3
	No		56.5	52.1	49.9	49.7	58.0	55.0	50.9	50.7
<b>33 internet</b>		<b>Internet</b>								
	Yes		34.2	37.5	39.3	39.9	31.3	33.6	37.0	36.6
	No		65.8	62.5	60.7	60.1	68.8	66.5	63.0	63.4
<b>34 wash_m</b>		<b>Washing machine</b>								
	Yes		92.9	94.8	95.4	96.6	94.1	95.1	95.6	96.4
	No		7.1	5.2	4.7	3.4	5.9	4.9	4.4	3.6
<b>35 refrig</b>		<b>Refrigerator</b>								
	Yes		98.0	98.9	99.0	98.9	98.6	98.7	98.6	98.9
	No		2.0	1.1	1.0	1.2	1.4	1.3	1.4	1.1
<b>36 dish_w</b>		<b>Dishwasher</b>								
	Yes		44.2	49.8	54.4	58.0	42.1	46.3	51.8	56.4
	No		55.8	50.2	45.6	42.0	57.9	53.7	48.2	43.6
<b>37 air_con</b>		<b>Air conditioner</b>								
	Yes		15.8	17.2	18.0	20.9	14.0	15.2	16.6	21.3
	No		84.2	82.8	82.0	79.1	86.0	84.8	83.4	78.7
<b>38 car</b>		<b>Car</b>								
	Yes		31.8	34.0	36.9	37.9	32.0	33.4	36.4	38.7
	No		68.2	66.0	63.1	62.1	68.1	66.6	63.6	61.4



#	NAME OF VARIABLE	VARIABLE	SILC (%)				HBS (%)			
			2010	2011	2012	2013	2010	2011	2012	2013
<b>39</b>	<b>low_mon_inc</b>	<b>Lowest monthly income to make ends meet</b>								
	<=1000		25.7	22.8	15.2	11.8	52.0	47.0	38.8	32.2
	>1000 and <=1500		24.2	25.3	23.9	18.8	24.6	25.1	25.0	26.3
	>1500 and <=2000		22.3	21.9	22.1	23.3	13.9	16.2	19.3	21.3
	>2000 and <=2500		8.7	9.6	9.5	12.2	3.4	3.9	5.5	6.2
	>2500 and <=3000		9.1	9.0	12.6	14.4	3.7	4.5	6.4	8.0
	>3000		10.0	11.4	16.8	19.6	2.4	3.3	5.0	6.0
<b>40</b>	<b>dis_inc_cat</b>	<b>Total disposable household income</b>								
	<=1000		24.7	19.2	14.1	9.7	24.1	18.2	13.7	10.9
	>1000 and <=1500		23.6	20.8	19.6	19.6	23.4	21.0	19.0	16.7
	>1500 and <=2000		17.2	17.9	17.7	18.2	17.5	17.2	17.4	16.6
	>2000 and <=2500		10.1	12.9	14.3	14.9	11.4	13.3	13.9	13.6
	>2500 and <=3000		8.1	9.0	10.1	9.7	8.3	9.4	10.5	11.0
	>3000		16.3	20.2	24.2	27.9	15.3	21.0	25.6	31.2

### Appendix 3. Regression tables

#### HBS, MODEL 1, 2010-2013

	2010		2011		2012		2013	
	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
Intercept	7.63	0.04***	7.82	0.03***	7.84	0.03***	7.88	0.03***
low_mon_inc_1	-0.51	0.03***	-0.59	0.03***	-0.61	0.02***	-0.57	0.02***
low_mon_inc_2	-0.35	0.03***	-0.43	0.03***	-0.47	0.02***	-0.43	0.02***
low_mon_inc_3	-0.24	0.03***	-0.35	0.03***	-0.35	0.02***	-0.35	0.02***
low_mon_inc_4	-0.20	0.04***	-0.31	0.03***	-0.29	0.03***	-0.26	0.03***
low_mon_inc_5	-0.16	0.04***	-0.23	0.03***	-0.23	0.03***	-0.20	0.02***
comp_1	0.08	0.02***	0.09	0.02***	0.10	0.02***	0.13	0.01***
dish_w_1	0.07	0.01***	0.10	0.01***	0.08	0.01***	0.09	0.01***
ref_edu_1	-0.34	0.02***	-0.30	0.02***	-0.33	0.02***	-0.33	0.02***
ref_edu_2	-0.10	0.02***	-0.09	0.02***	-0.11	0.01***	-0.13	0.01***
ref_edu_3	-0.07	0.02***	-0.06	0.02***	-0.08	0.02***	-0.08	0.02***
internet_1	0.12	0.02***	0.11	0.02***	0.11	0.02***	0.07	0.02***
car_1	0.30	0.01***	0.34	0.01**	0.33	0.01***	0.32	0.01***
rent_cat2_1	-0.14	0.01**	-0.16	0.01***	-0.14	0.01***	-0.11	0.01***
heat_sys_2	0.11	0.02***	0.07	0.02***	0.08	0.02***	0.09	0.02***
heat_sys_3	0.09	0.01***	0.11	0.01***	0.09	0.01***	0.14	0.01***
heat_sys_4	0.06	0.02**	0.06	0.03**	0.01	0.02	0.04	0.02*
heat_sys_5	-0.02	0.07	0.10	0.08	-0.02	0.16	-0.38	0.14***
dwe_1	0.01	0.01	0.03	0.01**	0.03	0.01**	0.03	0.01***
hot_wat_1	0.17	0.01***	0.11	0.01***	0.17	0.01***	0.14	0.02***
tot_ar_1	-0.34	0.02***	-0.31	0.02***	-0.30	0.02***	-0.33	0.02***
tot_ar_2	-0.19	0.02***	-0.19	0.02***	-0.19	0.02***	-0.18	0.02***
tot_ar_3	-0.14	0.01***	-0.15	0.01***	-0.16	0.01***	-0.15	0.01***
tot_ar_4	-0.07	0.01***	-0.10	0.01***	-0.12	0.01***	-0.11	0.01***

\* significant at 90% level

\*\* significant at 95% level

\*\*\* significant at 99% level

**SILC, MODEL 1, 2010-2013**

	2010		2011		2012		2013	
	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
<b>Intercept</b>	10.39	0.06***	10.43	0.05***	10.45	0.05***	10.57	0.05***
<b>low_mon_inc_1</b>	-0.75	0.04***	-0.85	0.04***	-0.81	0.04***	-0.79	0.04***
<b>low_mon_inc_2</b>	-0.62	0.04***	-0.72	0.04***	-0.66	0.03***	-0.65	0.03***
<b>low_mon_inc_3</b>	-0.47	0.04***	-0.53	0.04***	-0.52	0.03***	-0.52	0.03***
<b>low_mon_inc_4</b>	-0.38	0.04***	-0.39	0.04***	-0.45	0.04***	-0.38	0.03***
<b>low_mon_inc_5</b>	-0.30	0.04***	-0.34	0.04***	-0.32	0.03***	-0.33	0.03***
<b>comp_1</b>	0.09	0.03***	0.02	0.03	0.04	0.03	0.05	0.03**
<b>dish_w_1</b>	0.09	0.02***	0.11	0.02***	0.11	0.02***	0.10	0.02***
<b>ref_edu_1</b>	-0.36	0.04***	-0.30	0.04***	-0.31	0.04***	-0.37	0.04***
<b>ref_edu_2</b>	-0.23	0.03***	-0.18	0.03***	-0.17	0.03***	-0.23	0.03***
<b>ref_edu_3</b>	-0.20	0.03***	-0.19	0.03***	-0.22	0.03***	-0.25	0.03***
<b>internet_1</b>	0.08	0.03**	0.12	0.03***	0.13	0.03***	0.10	0.03***
<b>car_1</b>	0.18	0.02***	0.21	0.02***	0.20	0.02***	0.18	0.02***
<b>rent_cat2_1</b>	-0.11	0.02***	-0.09	0.03***	-0.01	0.03	-0.04	0.03
<b>heat_sys_2</b>	0.19	0.04***	0.14	0.04***	0.18	0.04***	0.20	0.03***
<b>heat_sys_3</b>	0.13	0.03***	0.13	0.03***	0.12	0.03***	0.11	0.02***
<b>heat_sys_4</b>	0.06	0.05	0.10	0.05**	0.06	0.05	0.11	0.05**
<b>heat_sys_5</b>	0.23	0.21	0.43	0.18**	0.56	0.19***	0.73	0.17***
<b>dwe_1</b>	0.04	0.02	0.06	0.03**	0.03	0.03	0.07	0.02***
<b>hot_wat_1</b>	0.23	0.03***	0.20	0.03***	0.15	0.03***	0.11	0.03***
<b>tot_ar_1</b>	-0.36	0.04***	-0.28	0.04***	-0.33	0.04***	-0.29	0.04***
<b>tot_ar_2</b>	-0.29	0.03***	-0.17	0.03***	-0.17	0.03***	-0.19	0.03***
<b>tot_ar_3</b>	-0.22	0.03***	-0.13	0.03***	-0.14	0.03***	-0.14	0.02***
<b>tot_ar_4</b>	-0.12	0.03***	-0.07	0.03***	-0.07	0.03**	-0.10	0.03***

\* significant at 90% level

\*\* significant at 95% level

\*\*\* significant at 99% level

**HBS, MODEL 2, 2010-2013**

	2010		2011		2012		2013	
	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
<b>Intercept</b>	7.78	0.04***	7.98	0.03***	7.98	0.03***	8.03	0.03***
<b>low_mon_inc_1</b>	-0.57	0.03***	-0.63	0.03***	-0.65	0.02***	-0.61	0.02***
<b>low_mon_inc_2</b>	-0.40	0.03***	-0.47	0.03***	-0.51	0.02***	-0.46	0.02***
<b>low_mon_inc_3</b>	-0.28	0.03***	-0.37	0.03***	-0.38	0.02***	-0.37	0.02***
<b>low_mon_inc_4</b>	-0.22	0.04***	-0.32	0.03***	-0.31	0.03***	-0.27	0.03***
<b>low_mon_inc_5</b>	-0.18	0.04***	-0.24	0.03***	-0.24	0.03***	-0.21	0.02***
<b>comp_1</b>	0.18	0.01***	0.19	0.01***	0.20	0.01***	0.19	0.01***
<b>ref_edu_1</b>	-0.38	0.02***	-0.34	0.02***	-0.36	0.02***	-0.37	0.02***
<b>ref_edu_2</b>	-0.13	0.02***	-0.12	0.01***	-0.14	0.01***	-0.16	0.01***
<b>ref_edu_3</b>	-0.09	0.02***	-0.08	0.02***	-0.10	0.02***	-0.10	0.02***
<b>car_1</b>	0.31	0.01***	0.35	0.01***	0.34	0.01***	0.33	0.01***
<b>rent_cat2_1</b>	-0.19	0.01***	-0.21	0.01***	-0.18	0.01***	-0.16	0.01***
<b>hot_wat_1</b>	0.19	0.01***	0.13	0.01***	0.20	0.01***	0.17	0.02***
<b>tot_ar_1</b>	-0.36	0.02***	-0.34	0.02***	-0.31	0.02***	-0.35	0.02***
<b>tot_ar_2</b>	-0.21	0.02***	-0.21	0.02***	-0.20	0.02***	-0.20	0.02***
<b>tot_ar_3</b>	-0.16	0.01***	-0.16	0.01***	-0.17	0.01***	-0.16	0.01***
<b>tot_ar_4</b>	-0.08	0.01***	-0.11	0.01***	-0.12	0.01***	-0.11	0.01***

\* significant at 90% level

\*\* significant at 95% level

\*\*\* significant at 99% level

**SILC, MODEL 2, 2010-2013**

	2010		2011		2012		2013	
	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
<b>Intercept</b>	10.59	0.05***	10.65	0.05***	10.67	0.05***	10.79	0.04***
<b>low_mon_inc_1</b>	-0.79	0.04***	-0.90	0.04***	-0.86	0.04***	-0.82	0.04***
<b>low_mon_inc_2</b>	-0.66	0.04***	-0.77	0.04***	-0.71	0.03***	-0.68	0.03***
<b>low_mon_inc_3</b>	-0.50	0.04***	-0.57	0.04***	-0.56	0.03***	-0.54	0.03***
<b>low_mon_inc_4</b>	-0.40	0.04***	-0.42	0.04***	-0.48	0.04***	-0.40	0.03***
<b>low_mon_inc_5</b>	-0.31	0.04***	-0.36	0.04***	-0.34	0.03***	-0.35	0.03***
<b>comp_1</b>	0.16	0.02***	0.12	0.02***	0.14	0.02***	0.14	0.02***
<b>ref_edu_1</b>	-0.42	0.04***	-0.35	0.04***	-0.37	0.04***	-0.43	0.04***
<b>ref_edu_2</b>	-0.29	0.03***	-0.23	0.03***	-0.22	0.03***	-0.28	0.03***
<b>ref_edu_3</b>	-0.23	0.03***	-0.21	0.03***	-0.25	0.03***	-0.27	0.03***
<b>car_1</b>	0.19	0.02***	0.23	0.02***	0.21	0.02***	0.19	0.02***
<b>rent_cat2_1</b>	-0.16	0.02***	-0.13	0.02***	-0.08	0.02***	-0.09	0.02***
<b>hot_wat_1</b>	0.25	0.03***	0.22	0.03***	0.18	0.03***	0.14	0.03***
<b>tot_ar_1</b>	-0.37	0.04***	-0.30	0.04***	-0.36	0.04***	-0.33	0.04***
<b>tot_ar_2</b>	-0.32	0.03***	-0.21	0.03***	-0.21	0.03***	-0.23	0.03***
<b>tot_ar_3</b>	-0.23	0.03***	-0.15	0.03***	-0.17	0.03***	-0.17	0.02***
<b>tot_ar_4</b>	-0.13	0.03***	-0.08	0.03***	-0.07	0.03***	-0.12	0.03***

\* significant at 90% level

\*\* significant at 95% level

\*\*\* significant at 99% level

**HBS, MODEL 3, 2010-2013**

	2010		2011		2012		2013	
	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
<b>Intercept</b>	7.96	0.03***	8.09	0.03***	8.14	0.03***	8.16	0.03***
<b>dis_inc_cat_1</b>	-0.90	0.02***	-0.85	0.02***	-0.91	0.02***	-0.98	0.02***
<b>dis_inc_cat_2</b>	-0.58	0.02***	-0.54	0.01***	-0.58	0.02***	-0.61	0.01***
<b>dis_inc_cat_3</b>	-0.43	0.02***	-0.40	0.01***	-0.43	0.01***	-0.45	0.01***
<b>dis_inc_cat_4</b>	-0.31	0.02***	-0.30	0.01***	-0.29	0.01***	-0.31	0.01***
<b>dis_inc_cat_5</b>	-0.23	0.02***	-0.19	0.02***	-0.22	0.02***	-0.23	0.01***
<b>low_mon_inc_1</b>	-0.32	0.03**	-0.39	0.02**	-0.41	0.02***	-0.36	0.02***
<b>low_mon_inc_2</b>	-0.24	0.03***	-0.30	0.02***	-0.34	0.02***	-0.28	0.02***
<b>low_mon_inc_3</b>	-0.18	0.03***	-0.28	0.02***	-0.28	0.02***	-0.26	0.02***
<b>low_mon_inc_4</b>	-0.17	0.03***	-0.25	0.03***	-0.24	0.03***	-0.22	0.02***
<b>low_mon_inc_5</b>	-0.14	0.03***	-0.21	0.03***	-0.21	0.02***	-0.18	0.02***
<b>comp_1</b>	0.04	0.01***	0.06	0.01***	0.06	0.01***	0.06	0.01***
<b>dish_w_1</b>	0.02	0.01**	0.05	0.01***	0.03	0.01***	0.03	0.01***
<b>ref_edu_1</b>	-0.16	0.02***	-0.14	0.02***	-0.15	0.02***	-0.14	0.02***
<b>ref_edu_2</b>	0.00	0.01	-0.02	0.01	-0.01	0.01	-0.04	0.01***
<b>ref_edu_3</b>	0.01	0.01	0.00	0.01	-0.02	0.01	-0.03	0.01*
<b>internet_1</b>	0.05	0.01***	0.06	0.01***	0.06	0.01***	0.04	0.01***
<b>car_1</b>	0.20	0.01***	0.24	0.01***	0.23	0.01***	0.23	0.01***
<b>rent_cat2_1</b>	-0.08	0.01***	-0.11	0.01***	-0.08	0.01***	-0.08	0.01***
<b>heat_sys_2</b>	0.05	0.02***	0.04	0.02**	0.03	0.02*	0.05	0.02***
<b>heat_sys_3</b>	0.03	0.01**	0.06	0.01***	0.04	0.01***	0.08	0.01***
<b>heat_sys_4</b>	0.01	0.02	0.03	0.02	-0.01	0.02	0.01	0.02
<b>heat_sys_5</b>	-0.06	0.06	0.10	0.07	-0.01	0.14	-0.27	0.12**
<b>dwe_1</b>	0.00	0.01	0.02	0.01	0.01	0.01	0.03	0.01***
<b>hot_wat_1</b>	0.12	0.01***	0.05	0.01***	0.09	0.01***	0.06	0.01***
<b>tot_ar_1</b>	-0.23	0.02***	-0.18	0.02***	-0.19	0.02***	-0.19	0.02***
<b>tot_ar_2</b>	-0.1	0.01***	-0.11	0.01***	-0.12	0.01***	-0.11	0.01***
<b>tot_ar_3</b>	-0.07	0.01***	-0.08	0.01***	-0.12	0.01***	-0.1	0.01***
<b>tot_ar_4</b>	-0.03	0.01***	-0.08	0.01***	-0.09	0.01***	-0.08	0.01***

\* significant at 90% level

\*\* significant at 95% level

\*\*\* significant at 99% level

**HBS, MODEL 4-5, 2010**

	<b>Model-4</b>		<b>Model-5</b>	
	<b>est.</b>	<b>s.e.</b>	<b>est.</b>	<b>s.e.</b>
<b>Intercept</b>	7.69	0.02***	7.61	0.02***
<b>dis_inc_cat_1</b>	-1.06	0.02***	-1.11	0.02***
<b>dis_inc_cat_2</b>	-0.68	0.01***	-0.71	0.01***
<b>dis_inc_cat_3</b>	-0.51	0.01***	-0.53	0.01***
<b>dis_inc_cat_4</b>	-0.37	0.02***	-0.38	0.02***
<b>dis_inc_cat_5</b>	-0.27	0.02***	-0.27	0.02***
<b>comp_1</b>	0.12	0.01***	0.15	0.01***
<b>car_1</b>	0.23	0.01***	0.23	0.01***
<b>hot_wat_1</b>	0.16	0.01***	0.20	0.01***
<b>rent_cat2_1</b>	-0.14	0.01***	-	-

\* significant at 90% level

\*\* significant at 95% level

\*\*\* significant at 99% level