Statistics Netherlands

Discussion Paper

Sampling and estimation techniques for household panels

2013 | 15

Jan van den Brakel

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress: Statistics Netherlands, Grafimedia
Design: Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information

60083 2013 X-42

Sampling and estimation techniques for household panels

Jan van den Brakel

A problem with using households as sampling units in the sample design of panels is the instability of these sampling units over time. Changes in the household composition affect the inclusion probabilities required for design-based and model-assisted inference procedures. The required information to derive correct inclusion probabilities is often not available. This problem can be circumvented by sampling persons which are followed over time. At each period the household members of these sampled persons are included in the sample. This comes down to sampling with probabilities proportional to household size where households can be selected more than once but with a maximum equal to the number of household members. In this paper properties of this sample design are described and applied to the Dutch Regional Income Survey.

Keywords: probabilities proportional to size, sampling with replacement, consistent weighting of persons and households, Regional Income Survey.

1. Introduction

Households are often considered as the sampling units in panels conducted to collect information on the level of households and persons, see e.g. Lynn (2009). Using households as sampling unit in a panel design has, however, some major disadvantages due to their instability over time. As time proceeds, households might disintegrate, join or split, new members might enter the households and other members might leave the households for all kind of reasons. As explained in the next paragraph, these changes can affect the selection probabilities of the households in the sample. Reconstruction of the correct inclusion probabilities of the sampling units is essential to derive correct weights for analysis purposes, but might become very complicated or even impossible, depending on the applied sampling design.

Consider a panel where households are selected by means of simple random sampling, say at time *t*. In many panels, people that enter a sampled household at later stage are also included in the panel. Consider for example household A, which is selected in the sample when the panel started. If, after some period of time, this household merges with another household B, which was initially not selected for the panel at time *t*, then the selection probability of this new household is the sum of the selection probabilities of household A and B at time *t*. The required information, to reconstruct the correct inclusion probabilities of the households observed in the panel is often not available. If time proceeds, more and more information about the history of all households in the target population is required to derive inclusion probabilities. If time proceeds, larger households will tend to be overrepresented in the aforementioned sample design. Not correcting for this overrepresentation through correct inclusion probabilities leads to biased inference.

Statistics Netherlands conducts two important sample surveys to describe the income and wealth situation of the Dutch population. First the Dutch Regional Income Survey (RIS) provides a global description of the income and wealth situation, being accurate at a very detailed regional level. This is accomplished by publishing accurate income distributions for persons and households at a level of neighbourhoods on a yearly basis. Second the Income Panel Survey (IPS) publishes yearly a precise detailed overview of income and wealth characteristics of the Dutch population on a global regional level.

The RIS and the IPS are both based on a panel and are conducted to collect information on the level of households and persons. To avoid the problems with panels using households as sampling units, an alternative design is developed. Instead of households, so-called core persons are drawn, which are followed over time. All household members belonging to the household of a core person at each particular period are included in the sample. This results in a sample design where households are drawn proportionally to the household size and households can be selected more than once, but with a maximum that is equal to the household size. The major advantage of this design is that the problems with reconstructing selection probabilities, as pointed out in the preceding paragraph, are circumvented.

The purpose of this paper is to describe a sample design with an estimation technique that is useful for panels that collect information on person and household level. This methodology is of general interest since the proposed sample design avoids the instability problems if households are used as sampling units. Particularly the use of web panels is frequently considered as a cost effective data collection mode to collect information on various social demographic themes. The proposed methodology is employed in the RIS and this application is used throughout the paper for illustration purposes.

The paper starts in Section 2, with a description of the sample design. In Section 3 the concept of inclusion expectations is introduced as a convenient practical alternative for inclusion probabilities. Subsequently first and second order inclusion expectations are derived for the proposed sampling design. The key target variables for the RIS are estimated income distributions. In Section 4 formulas for the minimum required sample size are derived based on a precision measure for estimated income distributions. Since households can be selected more than once, an expression for the expected number of unique households is derived in Section 4. Some additional remarks about the use of this sample design for panels are made in Section 5.

The estimation procedure of the RIS and the IPS is based on linear weighting and is described in Section 6. The starting point is the π-estimator or Horvitz-Thompson estimator, developed by Narain (1951), and Horvitz and Thompson (1952) for unequal probability sampling from finite populations without replacement. The observations are weighted with the inverse of the inclusion expectations, derived in Section 3, and account for the overrepresentation of large households. The precision of the Horvitz-Thompson estimator can be improved by taking advantage of available auxiliary information about the target population using the general regression estimator developed by Särndal et al. (1992). Finally the method of Lemaître and Dufour (1987) is considered to obtain consistent estimates for person and household based estimates. In Section 7 variance approximations for the general regression estimator under the proposed sample design are derived. An application to the RIS is provided in Section 8. The paper concludes with a discussion in Section 9.

2. Sampling design

The target population of the RIS are all natural persons aged 15 years or older residing in the Netherlands. The sample frame is a register containing all natural persons residing in the Netherlands as far as they are known to the Tax Office. From this register a stratified simple random sample of so-called core persons is drawn with a sample fraction of 0.16. Neighbourhoods are used as the stratification variable. At each period, all household members of the core persons are also included in the sample. Persons that leave the household of a core person also leave the panel. New persons entering the household of the core person are followed in the panel as long as this person stays in the household of a core person. As a result, a sample of households is obtained where the households are selected with probabilities proportional to the number of persons aged 15 years or older belonging to a household. Households can be selected more than once, but with a maximum that equals the number of household members, aged 15 year or older. In this paper the term core persons is used to refer to the persons that are initially included in the sample and are followed over time in the panel. The term persons is used to refer to the sample obtained if also all the household members at a particular period are included in the sample.

For each person that is included in the sample, the necessary information for the RIS variables is obtained from the registers of the Tax Office. Problems encountered with data collection where sampling units are asked to complete a questionnaire, for example nonresponse, do not occur. Of course other types of measurement errors are encountered with a survey that is based on registrations, see for example Wallgren and Wallgren (2007). It is assumed that all the required information about income to estimate the target parameters of the RIS, are available in these registers. Since all the required information is available in a register, a complete enumeration of the population is possible. In the past, however, the IT infrastructure was insufficient to produce timely regional income statistics based on a complete enumeration of the Dutch population. Therefore the RIS was traditionally based on a large sample with a fraction of 0.16 core persons.

3. First and second order inclusion expectations

The Horvitz-Thompson estimator, widely applied in design-based inference of probability sampling, is based on the concept of expanding the observations in the sample with the inverse of the inclusion probabilities. In the case of sampling with replacement, or partially with and partially without replacement as in the case of the proposed sampling design, it is convenient to generalise this concept to inclusion expectations, Bethlehem (2009), Chapter 2. Let a_k denote the number of times that unit *k* is selected in the sample for each element in the population. For the moment, a unit can be both a household or a person. In the proposed sample design $a_k \in [0,1,\ldots, g_k]$, with g_k the maximum number of times that unit *k* can be selected in the sample, i.e. g_k is the size of household *k* (if the units are households) or the household size of the household where person *k* belongs to (if the units are persons). Let E(.) denote the expectation with respect to the sample design. Now $\pi_k = \mathrm{E}(a_k)$ denotes the inclusion expectation of sampling unit *k*. Since a_k can be larger than one, π_k can also take values larges than one and can therefore no longer be interpreted as an inclusion probability. It can, however, be interpreted as an expectation.

The HT estimator for a population total can be defined as

$$\hat{t}_y = \sum_{k=1}^{N} \frac{a_k y_k}{\pi_k},$$

with *N* the total number of units in the population. Since $\mathrm{E}(a_k) = \pi_k$, it follows that this HT estimator is design unbiased. Let $\pi_{kk'}$ denote the inclusion expectation of units *k* and *k'* , i.e. $\pi_{kk'} = \mathrm{E}(a_k a_{k'})$. The variance of the HT estimator is by definition equal to

$$\begin{aligned}\mathrm{V}(\hat{t}_y) &= \sum_{k=1}^{N}\sum_{k'=1}^{N} \mathrm{Cov}(a_k a_{k'}) \frac{y_k}{\pi_k}\frac{y_{k'}}{\pi_{k'}} = \sum_{k=1}^{N}\sum_{k'=1}^{N} [\mathrm{E}(a_k a_{k'}) - \mathrm{E}(a_k)\mathrm{E}(a_{k'})] \frac{y_k}{\pi_k}\frac{y_{k'}}{\pi_{k'}} \\ &= \sum_{k=1}^{N}\sum_{k'=1}^{N} (\pi_{kk'} - \pi_k \pi_{k'}) \frac{y_k}{\pi_k}\frac{y_{k'}}{\pi_{k'}}.\end{aligned}$$

Note that in the case of sampling without replacement a_k is a dummy taking values zero or one indicating whether unit *k* is selected in the sample. In this case π_k and $\pi_{kk'}$ are the usual first and second order inclusion probabilities. This illustrates that the standard HT estimator, based on inclusion probabilities, can be extended easily to inclusion expectations. In the case of sample designs where units can be selected more than once, it is more convenient to work with inclusion expectations, since they are derived relatively easy. In the remainder of this section, first and second order inclusion expectations for the sample design described in Section 2, are derived.

Result 3.1: *Consider a sample design where so called core persons are drawn by means of stratified simple random sampling. Let N_h denote the number of persons in the population of stratum h, n_h the number of core persons selected in the sample from stratum h and g_{kh} the number of persons, belonging to household k from stratum h. All household members of the sampled core persons are*

included in the sample. First and second order inclusion expectations for households in this sample design are given by

$$\pi_{kh} = g_{kh}\frac{n_h}{N_h},$$

$$\pi_{kk'h} = g_{kh}g_{k'h}\frac{n_h(n_h-1)}{N_h(N_h-1)},\quad \pi_{kkh} = g_{kh}(g_{kh}-1)\frac{n_h(n_h-1)}{N_h(N_h-1)} + g_{kh}\frac{n_h}{N_h},$$

$$\pi_{kk'hh'} = g_{kh}g_{k'h'}\frac{n_h n_{h'}}{N_h N_{h'}}.$$

Proof: The first order inclusion expectation of the *k*-th household equals

$$\pi_{kh} = \mathrm{E}(a_{kh}) = \sum_{i=1}^{g_{kh}} i\mathrm{P}(a_{kh}=i) = \sum_{i=1}^{g_{kh}} i\frac{\binom{g_{kh}}{i}\binom{N_h-g_{kh}}{n_h-i}}{\binom{N_h}{n_h}}, \tag{3.1}$$

with a_{kh} the number of times that household *k* from stratum *h* is selected. The enumerator of the ratio in (3.1) is the number of times that *i* persons from a household of size g_{kh} and $n_h - i$ persons can be drawn from the remaining population of size $N_h - g_{kh}$. The denominator is the number of times that a sample of n_h persons can be drawn from a population of size N_h. Consequently the ratio is the probability that *i* persons form household *k* of size g_{kh} are drawn from a population of size N_h with a simple random sample of size n_h. Equation (3.1) can be expressed as

$$\pi_{kh} = \sum_{i=1}^{g_{kh}} g_{kh}\frac{\binom{g_{kh}-1}{i-1}\binom{N_h-g_{kh}}{n_h-i}}{\binom{N_h}{n_h}}. \tag{3.2}$$

In Mood, Graybill and Boes (1974, page 531) it is proved that

$$\sum_{j=0}^{n}\binom{a}{j}\binom{b}{n-j} = \binom{a+b}{n}. \tag{3.3}$$

By changing to $j = i-1$ and applying formula (3.3), it follows that (3.2) can be simplified to

$$\pi_{kh} = g_{hk}\sum_{j=0}^{g_{hk}-1}\frac{\binom{g_{hk}-1}{j}\binom{N_h-g_{hk}}{n-j-1}}{\binom{N_h}{n_h}} = g_{kh}\frac{\binom{N_h-1}{n_h-1}}{\binom{N_h}{n_h}} = g_{kh}\frac{n_h}{N_h}. \tag{3.4}$$

Second order inclusion expectations for households k and k' for $k \neq k'$ belonging to the same stratum *h,* equal

$$\pi_{kk'h} = \mathrm{E}(a_{kh}a_{k'h}) = \sum_{i=1}^{g_{kh}}\sum_{i'=1}^{g_{k'h}} ii'\,\mathrm{P}(a_{kh} = i, a_{k'h} = i')$$
$$= \sum_{i=1}^{g_{kh}}\sum_{i'=1}^{g_{k'h}} ii' \frac{\binom{g_{kh}}{i}\binom{g_{k'h}}{i'}\binom{N_h - g_{kh} - g_{k'h}}{n_h - i - i'}}{\binom{N_h}{n_h}} \quad . \tag{3.5}$$

Using similar arguments as specified following equation (3.1), the ratio in (3.5) is the probability that *i* persons form household *k* of size g_{kh} and i' persons form household k' of size $g_{k'h}$, both belonging to the same stratum *h*, are drawn from a population of size N_h with a simple random sample of size n_h. Equation (3.5) can be simplified to

$$\pi_{kk'h} = \sum_{i=1}^{g_{kh}}\sum_{i'=1}^{g_{k'h}} g_{kh}g_{k'h} \frac{\binom{g_{kh}-1}{i-1}\binom{g_{k'h}-1}{i'-1}\binom{N_h - g_{kh} - g_{k'h}}{n_h - i - i'}}{\binom{N_h}{n_h}}. \tag{3.6}$$

By changing to $j = i - 1$ and $j' = i' - 1$ and applying formula (3.3) twice, it follows that (3.6) simplifies to

$$\pi_{kk'h} = \sum_{j=0}^{g_{kh}-1}\sum_{j'=0}^{g_{k'h}-1} g_{kh}g_{k'h} \frac{\binom{g_{kh}-1}{j}\binom{g_{k'h}-1}{j'}\binom{N_h - g_{kh} - g_{k'h}}{n - j - j' - 2}}{\binom{N_h}{n_h}}$$
$$= \sum_{j=0}^{g_{kh}-1} g_{kh}g_{k'h} \frac{\binom{g_{kh}-1}{j}\binom{N_h - g_{kh} - 1}{n - j - 2}}{\binom{N_h}{n_h}} = g_{kh}g_{k'h} \frac{\binom{N_h - 2}{n_h - 2}}{\binom{N_h}{n_h}} = g_{kh}g_{k'h} \frac{n_h(n_h - 1)}{N_h(N_h - 1)}. \tag{3.7}$$

The second order inclusion expectation for $k = k'$ for households from the same stratum *h*, is given by

$$\pi_{kkh} = \mathrm{E}(a_{kh}a_{kh}) = \mathrm{E}(a_{kh}(a_{kh} - 1)) + \mathrm{E}(a_{kh}). \tag{3.8}$$

An expression for the first order inclusion expectation $\mathrm{E}(a_{kh})$ is already given by (3.4). The first term on the right hand side of (3.8) can be elaborated as follows:

$$\mathrm{E}(a_{kh}(a_{kh}-1)) = \sum_{i=2}^{g_{kh}} i(i-1)\mathrm{P}(a_{kh}=i) = \sum_{i=2}^{g_{kh}} i(i-1)\frac{\binom{g_{kh}}{i}\binom{N_h-g_{kh}}{n_h-i}}{\binom{N_h}{n_h}}$$
$$= \sum_{i=2}^{g_{kh}} g_{kh}(g_{kh}-1)\frac{\binom{g_{kh}-2}{i-2}\binom{N_h-g_{kh}}{n_h-i}}{\binom{N_h}{n_h}}. \qquad (3.9)$$

Changing to $j = i-2$ and applying formula (3.3) gives

$$\mathrm{E}(a_{kh}(a_{kh}-1)) = \sum_{j=0}^{g_{kh}-2} g_{kh}(g_{kh}-1)\frac{\binom{g_{kh}}{j}\binom{N_h-g_{kh}}{n_h-j-2}}{\binom{N_h}{n_h}} = g_{kh}(g_{kh}-1)\frac{\binom{N_h-2}{n_h-2}}{\binom{N_h}{n_h}} \qquad (3.10)$$
$$= g_{kh}(g_{kh}-1)\frac{n_h(n_h-1)}{N_h(N_h-1)}.$$

Inserting the expressions (3.4) and (3.10) into (3.8) gives

$$\pi_{kkh} = g_{kh}(g_{kh}-1)\frac{n_h(n_h-1)}{N_h(N_h-1)} + g_{kh}\frac{n_h}{N_h}. \qquad (3.11)$$

Second order inclusion expectations for households k and k' for $k \neq k'$ belonging to two different strata *h* and h' equal

$$\pi_{kk'hh'} = \mathrm{E}(a_{kh}a_{k'h'}) = \mathrm{E}(a_{kh})\mathrm{E}(a_{k'h'}) = g_{kh}g_{k'h'}\frac{n_h n_{h'}}{N_h N_{h'}}. \qquad (3.12)$$

This result is straightforward, since samples in different strata are drawn independently from each other. Collecting the results obtained in formula's (3.4), (3.8), (3.11), and (3.12), proves result 3.1.

■

Result 3.2: *Since all members of a selected household are included in the sample, it follows for the sample design considered in result 3.1 that:*

1. *The first order inclusion expectations for persons belonging to household k are equal to the first order inclusion expectation of household k, i.e.* π_{kh}.
2. *The second order inclusion expectations for persons from two different households k and* k', *are equal to the second order inclusion expectations of these households, i.e.* $\pi_{kk'h}$ *for two households from the same stratum or* $\pi_{kk'hh'}$ *for two households from two different strata.*
3. *The second order inclusion expectations for persons from the same household, are equal to the second order inclusion expectation for this household, i.e.* π_{kkh}.

4. Sample size determination

The purpose of the RIS is to publish income distributions for households and persons at different geographical levels. The most detailed level is neighbourhoods, which are also used as the stratification variable in the sample design. Income distributions for households for region *r* are defined as

$$P_{lr} = \frac{M_{lr}}{M_{+r}},\ l=1,\ldots,L, \tag{4.1}$$

where M_{lr} denotes the number of households from region *r*, belonging to the *l*-th income category, and $M_{+r} = \sum_l M_{lr}$, the total number of households in region *r*. This income distribution is estimated as

$$\hat{P}_{lr} = \frac{\hat{M}_{lr}}{M_{+r}},\ l=1,\ldots,L, \tag{4.2}$$

where $\hat{M}_{lr}$ denotes an appropriate direct estimator for the total number of households from region *r*, classified to the *l*-th income category. For the moment the Horvitz-Thompson estimator is assumed as an appropriate estimator for M_{lr}, i.e.

$$\hat{M}_{lr} = \sum_{h \in r} \sum_{k=1}^{m_h} \frac{y_{khl}}{\pi_{kh}}, \tag{4.3}$$

where $y_{khl} = 1$ if household *k* from stratum *h* is classified to the *l*-th income class and $y_{khl} = 0$ otherwise and m_h the total number of households selected in stratum *h*. In the RIS *L*=10. Income distributions for persons are defined and estimated accordingly to (4.1), (4.2), and (4.3).

For sample size determination, precision specifications for the estimated income distributions are required. If precision requirements are specified for the separated classes of the income distributions, then the income class with the largest population variance determines the minimum required sample size, resulting in unnecessary large sample sizes. This can be avoided by specifying an alternative precision measure which is defined as the square root of the mean over the variances of the estimated income classes of an income distribution:

$$s = \sqrt{\frac{1}{L} \sum_{l=1}^{L} V(\hat{P}_{lr})}\ . \tag{4.4}$$

In this paragraph an exact expression for *s* will be derived as well as an approximation that can be used to estimate the minimum required sample size which does not require information about income distributions or variances.

Since neighbourhoods are the most detailed regions for which income distributions are published, precision requirements for sample size determination are specified at this regional level. Since

neighbourhoods are used as the stratification variable in the sample design, expressions for *s* can be derived under simple random sampling without replacement of core persons within each region.

Result 4.1: *Consider a sample of* n_h *core persons, drawn by means of simple random sampling without replacement from a finite population of size* N_h*. An expression for the average precision measure* s_h *in (4.4) for an income distribution is given by*

$$s_h = \sqrt{\frac{1}{L}\frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(\frac{N_h}{M_h^2}\sum_{l=1}^{L}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}} - \sum_{l=1}^{L}\left(\frac{M_{lh}}{M_h}\right)^2\right)}.$$

Proof: An expression for the variance of the estimated fraction of households in income class *l* can be derived from the general expression for the variance of the Horvitz-Thompson estimator, Särndal et al. (1992), Section 2.8:

$$V(\hat{P}_{lh}) = \frac{1}{M_h^2}\sum_{k=1}^{M_h}\sum_{k'=1}^{M_h}(\pi_{kk'h} - \pi_{kh}\pi_{k'h})\frac{y_{khl}}{\pi_{kh}}\frac{y_{k'hl}}{\pi_{k'h}}. \qquad (4.5)$$

Inserting first and second order inclusion expectations specified in result 3.1 and taking advantage of the property that $y_{khl} = y_{khl}^2$ since the values of the target variable are restricted to zero or one, it follows after some algebra that (4.5) can be simplified to

$$V(\hat{P}_{lh}) = \frac{N_h - n_h}{n_h}\frac{1}{N_h - n_h}\left(\frac{N_h}{M_h^2}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}} - \left(\frac{M_{lh}}{M_h}\right)^2\right). \qquad (4.6)$$

Result 4.1 is obtained by inserting (4.6) into (4.4). ■

Remark: If $g_{kh} = 1$ for all households in the population of region *h*, then it follows that $M_h = N_h$ and that formula (4.6) simplifies to

$$V(\hat{P}_{lh}) = \frac{N_h - n_h}{n_h}\frac{1}{N_h - n_h}\left(P_{lh}(1 - P_{lh})\right), \qquad (4.7)$$

which can be recognized as the variance of an estimated fraction under simple random sampling without replacement, Cochran (1977), Chapter 3.

Result 4.2: *The average precision measure* s_h *for an income distribution, specified in result 4.1 can be approximated by*

$$s_h \le \sqrt{\frac{1}{L}\frac{N_h - n_h}{n_h}\frac{1}{N_h - 1}\left(\frac{N_h}{M_h^2}\sum_{t=1}^{T}\frac{M_{th}}{t} - \frac{1}{L}\right)},$$

with M_{th} *the number of households of size t in region h, and t the size of an household.*

Proof: The *population* of households in region *h* can be divided in *T* subpopulations of equally sized households. Let M_{th} denote the number of households of size *t* in region *h*. Now it follows for the double summation between brackets for the expression of *s* in result 4.1 that

$$\sum_{l=1}^{L}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}}=\sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{k=1}^{M_{th}}\frac{y_{khl}}{t}=\sum_{t=1}^{T}\frac{M_{th}}{t}\,. \tag{4.8}$$

According to the Chauchy-Schwartz inequality (Cochran, 1977, Section 5.5) it follows for the single summation between brackets for the expression of s_h in result 4.1 that

$$\sum_{l=1}^{L}\left(\frac{M_{lh}}{M_h}\right)^2=\sum_{l=1}^{L}P_{lh}^2\geq\frac{1}{L}\,. \tag{4.9}$$

Result 4.2 is obtained by inserting (4.8) and (4.9) in the expression for *s* in result 4.1. ■

Remark: If $g_{kh}=1$ for all households in the population of region *h* and the number of classes of the income distribution *L*=2, then it follows that the approximation for the average precision measure s_h in result 4.2 can be simplified to

$$s_h\leq\sqrt{\frac{N_h-n_h}{n_h}\frac{1}{(N_h-1)}\frac{1}{4}}\,, \tag{4.10}$$

which equals the square root of the maximum variance of an estimated fraction at $\hat{P}=0.5$ under simple random sampling. This illustrates that the approximation for the average precision measure in result 4.2 can be interpreted as a generalization of the approximation of the maximum variance of an estimated fraction at $\hat{P}=0.5$, often used in sample size determination. The average precision measure has its maximum value in the case of an equal distribution of the households over the income categories, i.e. $\hat{P}_{lh}=1/L$ for *l*=1, …, *L*. In this situation the approximation for s_h is exact, which follows directly from equation (4.9).

Remark: Equating the expression for s_h in result 4.2 to a pre-specified maximum value, say Δ_h, results in the following expression for the minimum sample size of core persons

$$n_h\geq\frac{\left(\frac{N_h}{M_h}\right)^2\sum_{t=1}^{T}\frac{M_{th}}{t}-\frac{N_h}{L}}{(N_h-1)L\Delta_h^2+\frac{N_h}{M_h^2}\sum_{t=1}^{T}\frac{M_{th}}{t}-\frac{1}{L}}\,. \tag{4.11}$$

The information required to estimate the minimum sample size is the total number of persons and the total number of equally sized households for neighbourhoods. No information about the expected income distribution or its variance is required. More precise estimates for the minimum sample size can be obtained with the expression in result 4.1, but require sample information from, for example, previous periods about the income distributions.

Expression (4.11) gives the minimum sample size for core persons. Subsequently all household members of each core person are included in the sample. As a result, households can be included in the sample more than once and the sample size in terms of unique households and unique persons is random. To plan a survey and control survey costs, it is necessary to know the expected number of unique households and unique persons if a sample of core persons of size n_h is drawn.

Result 4.3: *The expected number of unique households in a sample of n_h core persons, drawn by means of simple random sampling without replacement from a finite population of size N_h is given by*

$$D_h = \sum_{t=1}^{T} M_{th} \left(1 - \frac{\prod_{i=0}^{t-1} (N_h - n_h - i)}{\prod_{i=0}^{t-1} (N_h - i)} \right).$$

Proof: Let $\tilde{\pi}_{tkh}$ denote the inclusion probability for household *k* from stratum *h* of size *t*. Since equally sized households share the same first order probabilities, it follows that $\tilde{\pi}_{tkh} = \tilde{\pi}_{tk'h} \equiv \tilde{\pi}_{th}$. Let I_{tkh} denote an indicator variable, taking value 1 if household *k* from stratum *h* of size *t* is included in the sample and zero otherwise. The expected number of unique households can be derived as

$$D_h = \mathrm{E}\left(\sum_{t=1}^{T} \sum_{k=1}^{M_{th}} I_{tkh}\right) = \sum_{t=1}^{T} M_{th} \tilde{\pi}_{th} = \sum_{t=1}^{T} M_{th} \left(1 - \frac{\binom{N_h - t}{n_h}}{\binom{N_h}{n_h}} \right) .$$

$$= \sum_{t=1}^{T} M_{th} \left(1 - \frac{(N_h - n_h)(N_h - n_h - 1)....(N_h - n_h - t + 1)}{N_h (N_h - 1)...(N_h - t + 1)} \right)$$

■

Result 4.4: *The expected number of unique persons in a sample of n_h core persons, drawn by means of simple random sampling without replacement from a finite population of size N_h follows directly from result 4.3 and is given by*

$$D_h^{[p]} = \sum_{t=1}^{T} tM_{th} \left(1 - \frac{\prod_{i=0}^{t-1} (N_h - n_h - i)}{\prod_{i=0}^{t-1} (N_h - i)} \right).$$

Sample size calculations are conducted at the level of neighbourhoods, which have an average population size of about 5.000 persons. It was finally decided to select core persons with a sampling fraction of 1/6. With this sample size, the maximum value for the average precision measure s_h at the level of neighbourhoods amounts about 0.01 for the estimated household income distributions. With a total population of about 12.000.000 persons, this resulted in a sample size of about 2.100.000 core persons and an expected sample size of about 4.600.000 unique persons. This sample was drawn in 1994, which was the start of the panel for the Dutch RIS.

5. Panel design

The RIS is since 1994 conducted as a panel. Each year, it is determined which part of the population enters the target population of the RIS through birth and immigration. From this subpopulation a stratified simple random sample of core persons with a sample fraction of 1/6 is selected. These core persons are added to the panel of the RIS, with the purpose to maintain a representative sample.

As already explained in the introduction, households are inappropriate units to be used in a panel due to its relatively instable nature. As an alternative the core persons included in the sample are followed in the panel. Each year the household composition of the core persons is derived and the relevant information about income and wealth is gathered from registrations. The advantage of a self-weighted sample design for the core persons is that inclusion expectations for households and persons can be derived in a straightforward manner from the observed household composition at each point in time, since the inclusion expectations are proportional to the household size to which a core person in a particular year belongs.

Complications arise if different sampling fractions are applied in different strata. Consider a sample design where sample fractions vary over the strata. If new persons enter the household of a core person that originates from strata where different sample fractions are applied, then the inclusion expectation for this household is the sum over the inclusion expectations of the household members. This implies that the correct derivation of the inclusion expectations, at each point in time, requires information about the selection expectations for the entire population at the moment that the sample for the panel was drawn.

The situation becomes more complicated if households are used as sampling units, even in the case of self-weighted sampling designs. Integration and disintegration of households over time affects the inclusion probabilities of the households observed in the panel at particular time periods. In this case the correct reconstruction requires historical information about the household compositions in the population.

6. Linear weighting

For household surveys, estimates are required for person characteristics as well as household characteristics. Let t_y denote the total of a target variable *y*. With linear weighting, an estimator for a person based target variable is defined as:

$$\hat{t}_y = \sum_{h=1}^{H} \sum_{k \in 1}^{m_h} \sum_{i \in k} w_{ikh} y_{ikh} \ , \tag{6.1}$$

with w_{ikh} a weight for person *i* belonging to household *k* and stratum *h* and y_{ikh} the value of the target variable for person (*i,k,h*). An estimator for a household based target variable is given by:

$$\hat{t}_y = \sum_{h=1}^{H} \sum_{k=1}^{m_h} w_{kh} y_{kh} \ , \tag{6.2}$$

with w_{kh} a weight for household *k* from stratum *h* and y_{kh} the corresponding value of the target variable.

The weights can be obtained by means of the general regression estimator, Särndal et al. (1992). The regression based method uses auxiliary variables which are observed in the sample and for which the population totals are known from other sources. Consequently, the weights reflect the (unequal) inclusion expectations of the sampling units and an adjustment such that for auxiliary variables the weighted observations sum to the known population totals. Often categorical variables like gender, age, marital status or region are used as auxiliary variables. Due to the fact that the values of auxiliary variables differ from person to person within the same household, different weights can be derived for the same household. Therefore, it is relevant to apply a weighting method which yields one unique household weight for all its household members and still fulfil the additional requirement that the weighted auxiliary variables of the sampling units sum to the known population totals for persons. If the weights for persons within a household are the same, then household and person based estimates of the same target variables are consistent with each other (for example the total income estimated from households and that from persons).

Lemaître and Dufour (1987) proposed a method where it is forced that the weights for persons within a household are the same. They applied the linear weighting method at a person's level to obtain person weights. The original auxiliary variables defined at the person level are replaced by the corresponding household mean. Since members of the same household have the same inclusion expectation and share the same auxiliary information, the resulting weights based on the regression estimator are forced to be the same. In this paper, a slightly more general and direct approach is presented by applying the linear weighting method at the household level, where the auxiliary information of person based characteristics is aggregated at the household level. The method proposed by Lemaître and Dufour (1987) is a special case of this approach. The generalisation provides an interpretation of Lemaître and Dufour's method, since it explains under which variance structure the approach is efficient.

Let $\mathbf{x}_{kh}$ denote a *q* vector containing *q* auxiliary variables for household *k* from stratum *h*. Person based characteristics are aggregated to household totals. The general regression estimator is derived from a linear regression model that specifies the relationship between the target variable and the available auxiliary variables for which population totals are known, and is defined as

$$y_{kh} = \mathbf{x}_{kh}^{t}\boldsymbol{\beta} + e_{kh} \text{ , with} \qquad (6.3)$$
$$\mathrm{E}_m(e_{kh}) = 0 \text{ , } \mathrm{V}_m(e_{kh}) = \sigma_{kh}^2 .$$

In (6.3) $\boldsymbol{\beta}$ denotes a vector containing the *q* regression coefficients of the regression of y_{kh} on $\mathbf{x}_{kh}$ and e_{kh} the residuals and E_m and V_m denote the expectation and variance with respect to the regression model. It is required that all σ_{kh}^2 be known up to a common scale factor, that is $\sigma_{kh}^2 = \sigma^2\omega_{kh}$, with ω_{kh} known.

The general regression estimator for the population total of *y* is defined as (Särndal et al., 1992, Ch.6)

$$\hat{t}_y = \hat{t}_{y\pi} + \hat{\mathbf{b}}^t(\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x}\pi}) , \qquad (6.4)$$

with $\hat{t}_{y\pi}$ the Horvitz-Thompson estimator for t_y , $\mathbf{t}_\mathbf{x}$ a q vector containing the known population totals of the auxiliary variables $\mathbf{x}$, $\hat{\mathbf{t}}_{\mathbf{x}\pi}$ the Horvitz-Thompson estimator for $\mathbf{t}_\mathbf{x}$, $\hat{\mathbf{b}} = (\mathbf{X}^t\boldsymbol{\Pi}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\Pi}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{y}$ a design-based estimator for the regression coefficients in the population, i.e. $\mathbf{b} = (\mathbf{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{y}$ and therefore also for $\boldsymbol{\beta}$ in (6.3). Furthermore, $\mathbf{X}$ denotes an *m*×*q* matrix containing the auxiliary variables for the households observed in the sample, $\mathbf{y}$ an *m* vector containing the values of the target variables of the households observed in the sample, $\boldsymbol{\Sigma}$ an *m*×*m* diagonal matrix with the residual variances σ_{kh}^2 of all households in the sample and $\boldsymbol{\Pi}$ an *m*×*m* diagonal matrix containing the first order inclusion expectations.

An alternative expression for the general regression estimator is given by (Särndal et al. 1992, Ch. 6)

$$\hat{t}_y = \mathbf{y}^t\mathbf{w} , \qquad (6.5)$$

with

$$\mathbf{w} = \boldsymbol{\Pi}^{-1}(\mathbf{j} + \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^t\boldsymbol{\Pi}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x}\pi})) \qquad (6.6)$$

an *m* vector containing the weights w_{kh} obtained with the general regression estimator for the households observed in the sample and **j** an *m* vector with each element equal to one. The weights are calculated at the household level and can be used for weighting person based characteristics of the corresponding household members, using formula (6.1) since $w_{ikh} = w_{kh}$ for all persons belonging to the same household *k*.

The role of the linear regression model (6.3) is to describe the finite population in order to derive an estimator for the target variable. If the linear model explains the variation of the target parameter in the finite population reasonably well, then this will result in a reduction of the design variance of the Horvitz-Thompson estimator and decrease the bias due to selective non-response, Särndal and Swenson (1987), Bethlehem (1988), and Särndal and Lundström (2005). Another important property is that the general regression estimator ensures that for the auxiliary variables the weighted

observations sum to the known population totals. This property is used to enforce consistency between the marginal totals of different publication tables.

Result 6.1: *The integrated method for weighting persons and households proposed by Lemaître and Dufour (1987) follows as a special case from (6.6) if person based characteristics in* $\mathbf{X}$ *are aggregated to household totals and* $\boldsymbol{\Sigma}$ *is a diagonal matrix containing the household size at the diagonal elements. The latter condition implies that the variance of the residuals in the underlying regression model (6.3) are assumed to be proportional to the household size, i.e.* $\sigma_{kh}^2 = \sigma^2 \omega_{kh}$ *with* ω_{kh} *the size of household (k,h).*

Proof: The weighting procedure proposed by Lemaître and Dufour (1987) is a person based approach, assuming the following regression model for each person in the population; $y_{ikh} = \mathbf{z}_{ikh}^t \boldsymbol{\beta} + e_{ikh}$. The auxiliary information for each person in $\mathbf{z}_{ikh}$ is replaced by its household mean. Furthermore they implicitly assume that the residuals are independently distributed with equal variance, i.e. $\mathrm{V}_m(e_{ikh}) = \sigma^2$. In this case the regression weights are obtained by

$$\mathbf{w}_p = \boldsymbol{\Pi}_p^{-1}(\mathbf{j} + \mathbf{Z}(\mathbf{Z}^t \boldsymbol{\Pi}_p^{-1} \mathbf{Z})^{-1}(\mathbf{t}_z - \hat{\mathbf{t}}_{z\pi})) . \tag{6.7}$$

Let *n* denote the number of persons observed in the sample. Now $\mathbf{w}_p$ is an *n*-vector, containing the regression weights w_{ikh} for the persons observed in the sample, $\boldsymbol{\Pi}_p$ an *n*×*n* diagonal matrix containing the first order inclusion expectations for the persons selected in the sample, and $\mathbf{Z}$ denotes an *n*×*q* matrix containing the auxiliary variables $\mathbf{z}_{ikh}$. For the auxiliary information it follows that $\mathbf{t}_z = \mathbf{t}_x$ and $\hat{\mathbf{t}}_{z\pi} = \hat{\mathbf{t}}_{x\pi}$. Finally **L** is an *n*×*m* matrix, introduced to link person based information to household based information, Elements l_{ik} of this matrix are equal to one if person of the *i*-th row belongs to the household of the *k*-th column. Now we have the following relations:

$$\mathbf{Z} = (\mathbf{L}^t\mathbf{L})^{-1}\mathbf{L}\mathbf{X} \;,\; (\mathbf{L}^t\mathbf{L})^{-1}\mathbf{L}^t\mathbf{w}_p = \mathbf{w} \;,\; \mathbf{L}^t\boldsymbol{\Pi}_p^{-1} = \boldsymbol{\Pi}^{-1}\mathbf{L}^t . \tag{6.8}$$

If both sides of (6.7) are premultiplied with $(\mathbf{L}^t\mathbf{L})^{-1}\mathbf{L}^t$, then it follows from the relations specified in (6.8), that the resulting vector of weights is exactly equal to the vector specified in (6.6), under the condition that $\boldsymbol{\Sigma}$ is a diagonal matrix with the household size as diagonal elements. This proof was initially proposed by Nieuwenbroek (1993). ■

In the light of the model assisted approach of Särndal et al. (1992), result 6.1 provides an additional interpretation of this weighting technique, in particular concerning the choice of the variance structure for the residuals in (6.3). If y_{kh} is a household characteristic derived from summing the individual information of household members, then a suggestion is to assume a variance structure proportional to the household size, as implied by the method of Lemaître and Dufour.

7. Variance estimation

Since general regression estimators are non-linear, variances are obtained from a linearized approximation, obtained by means of a Taylor series expansion of the general regression estimator that is truncated at the first order term. Therefore general regression estimators are approximately design-unbiased, see Särndal et al (1992) for details.

Parameters of the RIS, are estimated as the ratio of two population totals

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_z}, \tag{7.1}$$

where $\hat{t}_y$ and $\hat{t}_z$ are defined by (6.1) or (6.2) in the case of person-based or household-based target variables, respectively.

Result 7.1: *The variance of (7.1) under a sample design where core persons are drawn by means of stratified simple random sampling, and all household members of these core persons are included in the sample is given by*

$$V(\hat{R}) = \frac{1}{t_z^2}\sum_{h=1}^{H} \frac{N_h^2(1-f_h)}{n_h} \frac{1}{N_h - 1}\sum_{k=1}^{N_h}\left(\frac{e_{kh}}{g_{kh}} - \frac{1}{N_h}\sum_{k'=1}^{N_h}\frac{e_{k'h}}{g_{k'h}}\right)^2,$$

where $f_h = n_h / N_h$, $e_{kh} = (y_{kh} - \mathbf{x}_{kh}^t \mathbf{b}_y) - R(z_{kh} - \mathbf{x}_{kh}^t \mathbf{b}_z)$, *and* $\mathbf{b}_y$ *and* $\mathbf{b}_z$ *the finite population regression coefficients of the regression of* y_{kh} *respectively* z_{kh} *on* $\mathbf{x}_{kh}$.

Proof: A general approximation for the variance of the ratio of two general regression estimators is given by (Särndal et al. 1992, Section 7.13):

$$V(\hat{R}) = \frac{1}{t_z^2}\sum_{h=1}^{H}\sum_{k=1}^{N_h}\sum_{h'=1}^{H}\sum_{k'=1}^{N_{h'}} (\pi_{kk'hh'} - \pi_{kh}\pi_{k'h'})\frac{e_{kh}}{\pi_{kh}}\frac{e_{k'h'}}{\pi_{k'h'}}. \tag{7.2}$$

After inserting first and second order inclusion expectations specified in result 3.1, it follows that (7.2) can be simplified to the variance expression defined in result 7.1. ■

Result 7.2: *An estimator for the variance specified in Result 7.1 is given by*

$$\hat{V}(\hat{R}) = \frac{1}{\hat{t}_z^2}\sum_{h=1}^{H}(1-f_h)\frac{n_h}{n_h - 1}\sum_{k=1}^{n_h}\left(w_{kh}\hat{e}_{kh} - \frac{1}{n_h}\sum_{k'=1}^{n_h} w_{k'h}\hat{e}_{k'h}\right)^2,$$

where $\hat{e}_{kh} = (y_{kh} - \mathbf{x}_{kh}^t \hat{\mathbf{b}}_y) - \hat{R}(z_{kh} - \mathbf{x}_{kh}^t \hat{\mathbf{b}}_z)$ *and* $\hat{\mathbf{b}}_y$ *and* $\hat{\mathbf{b}}_z$ *the Horvitz-Thompson type estimators for* $\mathbf{b}_y$ *and* $\mathbf{b}_z$*, defined by (6.5).*

Proof: An estimator for the variance approximation (7.2) is given by (Särndal et al. 1992, Section 7.13):

$$V(\hat{R}) = \frac{1}{\hat{t}_z^2} \sum_{h=1}^{H} \sum_{k=1}^{N_h} \sum_{h'=1}^{H} \sum_{k'=1}^{N_{h'}} \frac{(\pi_{kk'hh'} - \pi_{kh}\pi_{k'h'})}{\pi_{kk'hh'}} \frac{c_{kh}\hat{e}_{kh}}{\pi_{kh}} \frac{c_{k'h'}\hat{e}_{k'h'}}{\pi_{k'h'}}, \qquad (7.3)$$

where $c_{kh} = w_{kh} / \pi_{kh}$ are the correction weights. After inserting first and second order inclusion expectations specified in result 3.1 and some algebra, it follows that (7.3) equals

$$\hat{V}(\hat{R}) = \frac{1}{\hat{t}_z^2} \sum_{h=1}^{H} \frac{N_h^2 (1 - f_h)}{n_h} \frac{1}{n_h - 1} \sum_{k=1}^{n_h} \left(\frac{c_{kh}\hat{e}_{kh}}{g_{kh}} - \frac{1}{n_h} \sum_{k'=1}^{n_h} \frac{c_{k'h}\hat{e}_{k'h}}{g_{k'h}} \right)^2,$$

which is also equal to the estimator defined in result 7.2. ■

Note that the expressions for the variance are equal to the variance of stratified simple random sampling using transformed target variables y_{kh} / g_{kh}, z_{kh} / g_{kh} and transformed auxiliary variables $\mathbf{x}_{kh} / g_{kh}$. Indeed, expression (6.6) can also be expressed as

$$\mathbf{w} = \mathbf{\Gamma}^{-1}[\tilde{\mathbf{\Pi}}^{-1}(\mathbf{j} + \mathbf{\Lambda}^{-1}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^t \tilde{\mathbf{\Pi}}^{-1}\mathbf{\Lambda}^{-1}\tilde{\mathbf{X}})^{-1}(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}\pi}))] = \mathbf{\Gamma}^{-1}\tilde{\mathbf{w}}$$

with $\mathbf{\Lambda} = \mathbf{\Sigma}\mathbf{\Gamma}^{-1}$, $\mathbf{\Gamma}$ an $n \times n$ diagonal matrix with elements g_{kh}, $\tilde{\mathbf{\Pi}}$ a *n×n* diagonal matrix with elements n_h / N_h and $\tilde{\mathbf{X}} = \mathbf{\Gamma}^{-1}\mathbf{X}$, a matrix with the transformed auxiliary information. Inserting $\mathbf{w} = \mathbf{\Gamma}^{-1}\tilde{\mathbf{w}}$ in to (6.1) or (6.2) shows that these general regression estimators can be interpreted as a weighting applied to the transformed target and auxiliary variables under stratified simple random sampling. There is, however, no clear interpretation for the variance structure $\mathbf{\Lambda}$.

8. Application

In the RIS, core persons are selected from the population aged 15 years and older through stratified simple random sampling without replacement with a sample fraction of 0.16. In this application results are presented for a large municipality (Rotterdam), a municipality of intermediate size (Enschede) and a small municipality (Sevenum) for three subsequent years 2006, 2007 and 2008. Population and sample sizes for these three municipalities are summarized in Table 1.

Target variables of interest for the RIS are:

- Income distribution of households in ten classes where the categories are based on ten percentage quintile points of the national distribution (abbreviated as Inc. distr. hh.)
- Mean income households (abbreviated as HHinc)
- Mean income persons (abbreviated as Pinc)

Municipality	Population		Sample		
	Households	Persons 15 and older	Core persons	Unique households	Unique persons
Rotterdam	293400	484000	73000	67600	171400
Enschede	74200	128000	19300	17600	46300
Sevenum	2950	6100	870	750	2500

Table 1: population and sample size RIS for three Dutch municipalities.

Estimates for official publications of the RIS are obtained with the GREG estimator using the method of Lemaître and Dufour (1987). Since this survey does not suffer from nonresponse, auxiliary information is used in the estimation for variance reduction and consistency between the marginal of different publication tables. Inclusion expectations are based on the formulas derived in Section 3. For each municipality the following weighting scheme is applied in the GREG estimator:

Age(7)×Gender + Age(4)×Gender×MaritalStatus(2) + Address(3).

All auxiliary variables are categorical. The number between brackets denote the number of categories. MaritalStatus distinguishes between people who are married and other forms of marital status. Address distinguish between addresses where one person is residing, one family is residing and other types of addresses. Standard errors for these GREG estimates are based on the approximations derived in Section 7. Estimates for the aforementioned target variables with their standard errors based on the HT estimator, the GREG estimator and the GREG estimator with the method of Lemaître and Dufour are given in Tables 2, 3, and 4 for Rotterdam, Enschede and Sevenum respectively.

For each municipality there is a steady increase over time of the mean of the income for households and persons. Also the income distributions for each municipality show a stable pattern over the years. This can be expected if a panel is applied in combination with large sample sizes to estimate phenomena that are not very volatile in time. Differences in precision between the HT estimator and

the GREG estimator are small for large samples like Rotterdam. For smaller samples like Sevenum, the use of auxiliary information through the GREG estimator results in an increase of precision.

Comparing GREG estimates with and without using the method of Lemaître and Dufour shows that standard errors of estimated household parameters are smaller if the method of Lemaître and Dufour is applied. This is particularly visible for the mean household income in the small sample of Sevenum. For estimated person based parameters, on the other hand, the method of Lemaître and Dufour increases the standard error compared with the regular GREG estimator. For the household income distributions, which are defined as the mean over an indicator variable at the household level, the standard errors are more or less equal. This can be explained with the interpretation for the method of Lemaître and Dufour provided in Section 6. Lemaître and Dufour implies a linear regression model with a residual variance proportional to the household size. This assumption is reasonable for household variables that are obtained by summing the individual information from household members, like the mean household income, but less efficient for personal based characteristics. The additional advantage of Lemaître and Dufour is that totals for household and person based income, which can be derived directly from their means, are consistent.

Rotterdam 2006

Variable	HT		GREG		GREG consistent (L&D)	
Inc. distr. hh. 1	0.2380	(0.0019)	0.2233	(0.0016)	0.2260	(0.0016)
2	0.1876	(0.0017)	0.1797	(0.0016)	0.1838	(0.0016)
3	0.1335	(0.0014)	0.1319	(0.0013)	0.1346	(0.0014)
4	0.1022	(0.0012)	0.1026	(0.0012)	0.1043	(0.0012)
5	0.0764	(0.0010)	0.0789	(0.0010)	0.0794	(0.0010)
6	0.0651	(0.0009)	0.0687	(0.0009)	0.0678	(0.0009)
7	0.0574	(0.0008)	0.0617	(0.0008)	0.0596	(0.0008)
8	0.0509	(0.0007)	0.0552	(0.0007)	0.0523	(0.0007)
9	0.0463	(0.0007)	0.0508	(0.0007)	0.0470	(0.0006)
10	0.0424	(0.0006)	0.0469	(0.0006)	0.0449	(0.0006)
HHinc	19790	(83)	20134	(80)	20161	(76)
PPinc	22074	(94)	22219	(84)	22233	(93)

Rotterdam 2007

Variable	HT		GREG		GREG consistent (L&D)	
Inc. distr. hh. 1	0.2370	(0.0019)	0.2223	(0.0016)	0.2242	(0.0016)
2	0.1911	(0.0017)	0.1832	(0.0016)	0.1878	(0.0016)
3	0.1327	(0.0014)	0.1312	(0.0013)	0.1346	(0.0013)
4	0.1045	(0.0012)	0.1053	(0.0012)	0.1074	(0.0012)
5	0.0770	(0.0010)	0.0797	(0.0010)	0.0798	(0.0010)
6	0.0628	(0.0009)	0.0663	(0.0009)	0.0660	(0.0009)
7	0.0561	(0.0008)	0.0600	(0.0008)	0.0576	(0.0008)
8	0.0503	(0.0007)	0.0546	(0.0007)	0.0514	(0.0007)
9	0.0460	(0.0007)	0.0506	(0.0007)	0.0467	(0.0006)
10	0.04256	(0.0006)	0.04696	(0.0006)	0.0445	(0.0006)
HHinc	22306	(73)	22950	(64)	22866	(64)
PPinc	24094	(82)	24362	(75)	24432	(78)

Rotterdam 2008

Variable	HT		GREG		GREG consistent (L&D)	
Inc. distr. hh. 1	0.2355	(0.0019)	0.2201	(0.0016)	0.2222	(0.0016)
2	0.1887	(0.0017)	0.1807	(0.0016)	0.1851	(0.0016)
3	0.1335	(0.0014)	0.1317	(0.0013)	0.1350	(0.0014)
4	0.1048	(0.0012)	0.1056	(0.0012)	0.1070	(0.0012)
5	0.0760	(0.0010)	0.0788	(0.0010)	0.0792	(0.0010)
6	0.0641	(0.0009)	0.0677	(0.0009)	0.0671	(0.0009)
7	0.0577	(0.0008)	0.0621	(0.0008)	0.0601	(0.0008)
8	0.0510	(0.0007)	0.0557	(0.0007)	0.0526	(0.0007)
9	0.0465	(0.0007)	0.0511	(0.0007)	0.0472	(0.0006)
10	0.0421	(0.0006)	0.0467	(0.0006)	0.0444	(0.0006)
HHinc	23750	(78)	24511	(69)	24410	(68)
PPinc	25325	(84)	25625	(75)	25705	(78)

Table 2: Estimation results RIS for Rotterdam (large Dutch municipality), standard errors between brackets.

Enschede 2006

Variable	HT		GREG		GREG consistent (L&D)	
Inc. distr. hh. 1	0.2572	(0.0038)	0.2360	(0.0030)	0.2398	(0.0029)
2	0.1782	(0.0033)	0.1695	(0.0030)	0.1701	(0.0029)
3	0.1283	(0.0026)	0.1258	(0.0025)	0.1268	(0.0025)
4	0.1024	(0.0022)	0.1041	(0.0022)	0.1050	(0.0021)
5	0.0849	(0.0019)	0.0906	(0.0019)	0.0916	(0.0019)
6	0.0682	(0.0017)	0.0745	(0.0017)	0.0748	(0.0017)
7	0.0587	(0.0015)	0.0644	(0.0015)	0.0630	(0.0015)
8	0.0496	(0.0013)	0.0550	(0.0014)	0.0528	(0.0013)
9	0.0411	(0.0012)	0.0462	(0.0012)	0.0435	(0.0012)
10	0.0314	(0.0011)	0.0341	(0.0011)	0.0327	(0.0010)
HHinc	19810	(128)	20353	(111)	20300	(107)
Pinc	20402	(102)	20608	(92)	20590	(92)

Enschede 2007

Variable	HT		GREG		GREG consistent (L&D)	
Inc. distr. hh. 1	0.2621	(0.0039)	0.2397	(0.0030)	0.2427	(0.0029)
2	0.1728	(0.0033)	0.1647	(0.0030)	0.1658	(0.0029)
3	0.1273	(0.0026)	0.1248	(0.0025)	0.1264	(0.0025)
4	0.1035	(0.0022)	0.1054	(0.0022)	0.1060	(0.0022)
5	0.0845	(0.0019)	0.0899	(0.0019)	0.0909	(0.0019)
6	0.0692	(0.0017)	0.0756	(0.0017)	0.0764	(0.0017)
7	0.0583	(0.0015)	0.0645	(0.0015)	0.0635	(0.0015)
8	0.0502	(0.0014)	0.0555	(0.0014)	0.0527	(0.0013)
9	0.0407	(0.0012)	0.0456	(0.0012)	0.0431	(0.0012)
10	0.0315	(0.0011)	0.0343	(0.0011)	0.0325	(0.0010)
HHinc	20878	(128)	21716	(107)	21753	(105)
Pinc	21387	(115)	21751	(103)	21852	(106)

Enschede 2008

Variable	HT		GREG		GREG consistent (L&D)	
Inc. distr. hh. 1	0.2672	(0.0038)	0.2432	(0.0029)	0.2469	(0.0029)
2	0.1725	(0.0033)	0.1641	(0.0029)	0.1651	(0.0029)
3	0.1264	(0.0026)	0.1240	(0.0025)	0.1252	(0.0025)
4	0.0989	(0.0022)	0.1011	(0.0021)	0.1019	(0.0021)
5	0.0868	(0.0020)	0.0924	(0.0019)	0.0934	(0.0019)
6	0.0686	(0.0016)	0.0759	(0.0017)	0.0765	(0.0017)
7	0.0588	(0.0015)	0.0649	(0.0015)	0.0637	(0.0015)
8	0.0490	(0.0013)	0.0549	(0.0014)	0.0526	(0.0013)
9	0.0408	(0.0012)	0.0453	(0.0012)	0.0422	(0.0012)
10	0.0310	(0.0010)	0.0343	(0.0011)	0.0326	(0.0010)
HHinc	22254	(148)	23235	(125)	23237	(123)
Pinc	22235	(123)	22659	(110)	22724	(114)

Table 3: Estimation results RIS for Enschede (Dutch municipality of intermediate size), standard errors between brackets.

Sevenum 2006

Variable	HT		GREG		GREG consistent (L&D)	
Inc. distr. hh. 1	0.0880	(0.0131)	0.0835	(0.0112)	0.0821	(0.0108)
2	0.1195	(0.0145)	0.1148	(0.0123)	0.1153	(0.0121)
3	0.1079	(0.0125)	0.1013	(0.0111)	0.1043	(0.0111)
4	0.0908	(0.0107)	0.0885	(0.0100)	0.0885	(0.0100)
5	0.0911	(0.0101)	0.0928	(0.0100)	0.1001	(0.0100)
6	0.0900	(0.0094)	0.0968	(0.0092)	0.0980	(0.0093)
7	0.1345	(0.0111)	0.1352	(0.0105)	0.1346	(0.0103)
8	0.1001	(0.0094)	0.1018	(0.0091)	0.0984	(0.0090)
9	0.0829	(0.0082)	0.0859	(0.0081)	0.0841	(0.0081)
10	0.0952	(0.0090)	0.0996	(0.0089)	0.0946	(0.0086)
HHinc	25696	(799)	25698	(734)	25968	(711)
Pinc	21328	(466)	21680	(428)	21712	(428)

Sevenum 2007

Variable	HT		GREG		GREG consistent (L&D)	
Inc. distr. hh. 1	0.0851	(0.0129)	0.0818	(0.0106)	0.0800	(0.0103)
2	0.1343	(0.0153)	0.1162	(0.0116)	0.1165	(0.0116)
3	0.1014	(0.0120)	0.0951	(0.0107)	0.0977	(0.0108)
4	0.0879	(0.0107)	0.0866	(0.0100)	0.0883	(0.0101)
5	0.0966	(0.0102)	0.0989	(0.0098)	0.1020	(0.0101)
6	0.1058	(0.0104)	0.1090	(0.0100)	0.1118	(0.0102)
7	0.1191	(0.0103)	0.1257	(0.0100)	0.1254	(0.0100)
8	0.1110	(0.0098)	0.1172	(0.0095)	0.1147	(0.0093)
9	0.0768	(0.0078)	0.0821	(0.0078)	0.0803	(0.0078)
10	0.0820	(0.0083)	0.0873	(0.0080)	0.0836	(0.0078)
HHinc	28207	(618)	28901	(520)	29026	(490)
Pinc	24056	(456)	24219	(396)	24459	(393)

Sevenum 2008

Variable	HT		GREG		GREG consistent (L&D)	
Inc. distr. hh. 1	0.0920	(0.0133)	0.0843	(0.0110)	0.0798	(0.0107)
2	0.1331	(0.0154)	0.1187	(0.0119)	0.1199	(0.0119)
3	0.1071	(0.0124)	0.1001	(0.0107)	0.1038	(0.0109)
4	0.0733	(0.0097)	0.0711	(0.0089)	0.0752	(0.0087)
5	0.0865	(0.0098)	0.0866	(0.0091)	0.0898	(0.0091)
6	0.1098	(0.0104)	0.1176	(0.0103)	0.1206	(0.0104)
7	0.1347	(0.0114)	0.1421	(0.0112)	0.1411	(0.0112)
8	0.0946	(0.0090)	0.1011	(0.0089)	0.0996	(0.0089)
9	0.0786	(0.0081)	0.0838	(0.0081)	0.0813	(0.0081)
10	0.0904	(0.0088)	0.0948	(0.0085)	0.0889	(0.0082)
HHinc	31466	(795)	32372	(715)	32536	(694)
Pinc	24980	(468)	25482	(426)	25644	(455)

Table 4: Estimation results RIS for Sevenum (small Dutch municipality), standard errors between brackets.

9. Discussion

Households are inappropriate as sampling units in panels conducted to collect information at the level of households or persons. Since the internal composition of households changes over time the reconstruction of the correct inclusion probabilities, required for design-based and model-assisted inference, can become very complicated or even impossible. To avoid these complications, a sample design is prosed in this paper where persons are drawn through a self-weighted sample design. At each point in time, the household members of these so-called core persons are included in the sample. This results in a sample where households can be drawn more than once but with a maximum that is equal to the household size.

It is shown that first and second order inclusion expectations for this sample design can be derived in a relatively straightforward manner from the household composition of the core persons at each point in time. No additional information about the history of changes in the household composition in the past is required. These inclusion expectations can be used in a similar way in design-based and model-assisted inference as the more common inclusion probabilities. Expressions for minimum sample sizes to meet a pre-specified precision for estimated distributions as well as the expected number of unique households in a sample are derived.

In this context weighting procedures that enforce equal regression weights for persons within the same household are relevant in order to enforce consistency between person based and household based estimates. In this paper an approach that is slightly more general compared to the procedure proposed by Lemaître and Dufour (1987) is described. It also provides an interpretation of the method of Lemaître and Dufour since it shows that the underlying regression model assumes a residual variance that is proportional to the household size. An application to the RIS illustrates that this assumption is reasonable for household based estimates since it decreases the standard error of the GREG estimates. For person based characteristics Lemaître and Dufour increases the standard errors, which is the price paid for enforcing consistency between person and household based parameters.

10. References

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, pp. 251-260.

Bethlehem, J.G. (2009). *Applied survey methods*, John Wiley & Sons, New Yersey.

Cochran, W.G., (1977). *Sampling techniques*, John Wiley & Sons, New York.

Horvitz, D.G., and D.J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Ass*ociation, 47, pp. 663-685.

Lemaître, G. and Dufour, J. (1987). An integrated method for weighting persons and families. Survey Methodology, 13, pp. 199-207.

Lynn, P. (2009). Methods for longitudinal surveys, in *Methodology of longitudinal surveys,* Ed. P. Lynn. Chichester, Wiley.

Mood, A.M., F.A. Graybill and D.C. Boes, (1974). *Introduction to the theory of statistics*. McGraw-Hill, Singapore.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, pp. 169-174.

Nieuwenbroek, N. J. (1993). An integrated method for weighting characteristics of persons and households using the linear regression estimator. Research paper, BPA nr.: 8555-93-M1-1, Statistics Netherlands, Heerlen.

Särndal, C.-E., and S. Lundström (2005). *Estimation in Surveys with Nonresponse*. New-York: Wiley.

Särndal, C.E., and B. Swensson (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *International Statistical Review*, 55, pp. 279-294.

Särndal, C.-E., B. Swensson and J. Wretman, (1992). *Model assisted survey sampling*. Springer-Verlag, New-York.

Wallgren and Wallgren (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*. New York: Wiley.