Statistics
Netherlands

Discussion Paper

Register-based sampling for household panels

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

2015 | 04

Jan van den Brakel

Summary

In the Netherlands, statistical information about income and wealth is based on two large scale household panels that are completely derived from administrations. A problem with using households as sampling units in the sample design of panels is the instability of these units over time. Changes in the household composition affect the inclusion probabilities required for design-based and model-assisted inference procedures. Such problems are circumvented in the two aforementioned household panels by sampling persons which are followed over time. At each period the household members of these sampled persons are included in the sample. This is equivalent 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 contrasted with the Generalized Weight Share method developed for indirect sampling (Lavallée 1995, 2007). Methods are illustrated with an application to the Dutch Regional Income Survey.

Keywords: probabilities proportional to size, indirect sampling, consistent weighting of persons and households, Regional Income Survey, Generalized Weight Share method.

Index

1. Introduction

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 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, using a large sample based on a small set of the main income components derived in a relatively straightforward manner from tax administrations. Second, the Income Panel Survey (IPS) publishes yearly a precise detailed overview of income and wealth characteristics of the Dutch population on a more aggregated regional level. This survey is based on a large set of variables using all possible income components of households that can be derived from the available administrations in the Netherlands. The derivation of the variables for this survey is more time consuming. Therefore the sample size of this survey is considerably smaller compared to the RIS. Both surveys are designed as a household panel where both person and household based variables about income and wealth are observed.

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). Such panels are used for longitudinal analysis as well as the production of cross-sectional estimates. Using households as sampling units 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 different reasons. As explained by Kalton and Brick (1995), 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, in particular if the panel is used for producing cross-sectional estimates.

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 a later stage are also included in the panel. These individuals are in Lavallée (1995) called cohabitants. As time proceeds, more and more cohabitants are included in the sample and disturb the equal probability design that is used to select the initial sample, Kalton and Brick (1995). 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 households A and B at time t. Not correcting for differences in selection probabilities due to the gradual increasing share of cohabitants in the sample leads to biased inference. Ernst (1989) proposed the Weight Share method to overcome these problems. Lavallée (1995) extended this method to the Generalized Weight Share method as a solution for drawing inference about target populations that are sampled through the use of a frame that refers to a different population.

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 applied. 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. This design is an application of indirect sampling considered in Lavallée (1995, 2007), and Deville and Lavallée (2006).

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. The methodology employed in this paper is particularly useful for register based sampling, since the core persons are included in the sample indefinitely. The sample design is also useful for Web panels, but might require some form of rotation design to avoid problems with panel attrition. The main contribution of this paper to the existing literature is that explicit expressions for the variance of the target parameters are derived using inclusion expectations instead of inclusion probabilities. A measure of the minimum accuracy for an estimated income distribution is proposed and explicit expressions for the minimum sample size are derived. The RIS is used throughout the paper for illustration purposes of the described sampling techniques.

The paper starts in Section 2, with a description of the sample design of the RIS. 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. These inclusion weights are required to construct the π-estimator or Horvitz-Thompson (HT) estimator, initially proposed by Narain (1951) and Horvitz and Thompson (1952). It is also shown that the same weights can be derived as a special case of the Generalized Weight Share method for indirect sampling proposed by Lavallée (1995, 2007). 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 is based on linear weighting using the general regression (GREG) estimator developed by Särndal et al. (1992) and is described in Section 6. An integrated weighting method, initially proposed by Lemaître and Dufour (1987) and further developed by Nieuwenbroek (1993) and Steel and Clark (2007) is applied to obtain equal weights for persons belonging to the same household. In Section 7 variance approximations for the GREG 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 residing in the Netherlands. The sample frame is a register containing all natural persons aged 15 years and over 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. Although an equal probability design is used, stratified sampling is useful to eliminate the variation between strata and to meet minimum precision requirements for the individual strata. Neighbourhoods are the most detailed level of publication for the RIS and therefore used as strata. In Section 4 expressions for minimum sample sizes based on precision requirements are derived. Core persons remain in the panel indefinitely. 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. Information about the household composition of the core persons are obtained from the Municipal Basis Administration (MBA), which is the Dutch government's registry of all residents in the country. Dutch citizens are required by law to report changes in their demographics to their municipalities. The MBA is in combination with the information from tax administrations used to identify household members of the core persons in the sample.

The sample design results in a sample of households where the households are selected with probabilities proportional to the number of persons aged 15 years or older belonging to a household at the current period. 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.

The IPS applies a similar sample design with a substantially smaller sample fraction. The RIS as well as the IPS are registered based samples which implies that 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. Core persons as well as their household members are therefore not aware that they are included in these samples. This has the advantage that there are no problems with selective non-response and panel attrition. This also makes it possible to include the core persons indefinitely in the sample. In case of a panel where sampling units must complete a questionnaire, some kind of rotation design would be required in order to avoid selectivity bias due to panel attrition. Also problems with measurement bias associated with data collection where sampling units are asked to complete a questionnaire 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 and the IPS, 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. For the same reason the IPS is traditionally based on a sample of about 80,000 core persons. With the current computational capacity a complete enumeration would still be very demanding but not impossible. The main rationale for conducting this survey as a sample is to maintain the panel for longitudinal analysis.

3. Inclusion weights

3.1 Weighting with inclusion expectations

For design-based inference as well as planning the sample size of the survey, first and second order inclusion probabilities for households and persons are required. Let *M* denote the number of households in the population, *N* the number of persons in the population aged 15 years or over and g_k the number of persons aged 15 years or over that belong to the *k*-th household and N_k the number of persons that belong to the *k*-th household. With the sample design described in Section 2, households *k* can be included more than once but with a maximum g_k. This complicates the derivation of inclusion probabilities since the probability of selecting household *k* is equal to the selection probability of the union of its household members (*k,j*) aged 15 years and over. This probability is defined as:

$$P(k \in s) = P(\bigcup_{j=1}^{g_k}[(k,j) \in s]) = \sum_{j=1}^{g_k} P((k,j) \in s) - \sum_{j=1}^{g_k} \sum_{j'=j+1}^{g_k} P([(k,j) \cap (k,j')] \in s) +$$
$$\sum_{j=1}^{g_k} \sum_{j'=j+1}^{g_k} \sum_{j''=j+j'+1}^{g_k} P([(k,j) \cap (k,j') \cap (k,j'')] \in s) - \ldots$$

This kind of computations can be avoided by using the concept of inclusion expectations instead of inclusion probabilities. Bethlehem (2009), Chapter 2, generalized the HT estimator to the concept of inclusion expectations for sampling with replacement. Let a_k denote the number of times that household *k* is selected in the sample for each element in the population. In the proposed sample design $a_k \in [0,1,\ldots, g_k]$. 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 parameter of interest is the population total, which is defined as

$$t_y = \sum_{k=1}^{M} \sum_{j=1}^{N_k} y_{kj} \equiv \sum_{k=1}^{M} y_k \,. \qquad (1)$$

The HT estimator for the population total in (1) can be defined as

$$\hat{t}_y = \sum_{k=1}^{M} \frac{a_k y_k}{\pi_k} \,. \qquad (2)$$

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

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

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 subsection, 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 aged 15 years or over, n_h the number of core persons selected in the sample from stratum h and g_{kh} the number of persons aged 15 years or over, 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*

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

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

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

A proof of Result 3.1 is given in the appendix.

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}.*

3.2 Generalized Weight Share method

The sample design described in Section 2 can be considered as a special case of indirect sampling, Lavallée (2007). Indirect sampling refers to the situation where the population of interest is sampled through the use of a frame that refers to a different population. The Generalized Weight Share method is developed by Lavallée (1995) to construct weights for these situations and can be used to derive design weights for households and persons in the sample design described in Section 2.

Following the notation of Lavallée (1995) for the case of indirect sampling, there is a population U^A of size N^A from which a sample s^A of size *n* is drawn with selection probabilities π_i^A. In addition, there is the target population U^B of size N^B. This population can be divided in M^B clusters. Each cluster *k* contains N_k^B units, such that $N^B = \sum_{k=1}^{M^B} N_k^B$. The situation for the sample design described in Section 2 is depicted in Figure 1. The clusters are households, U^A is the population of persons aged 15 years and over, and U^B is the population of all persons residing in the Netherlands. Persons in U^A and U^B are depicted as circles, households in U^B are depicted as shaded squares, and the circles within a shaded square visualise persons belonging to the same household. Figure 1 shows respectively, a single person household, a two person household containing for example a divorced parent with a child younger than 15, a two person household containing two adults without children, and a four persons household containing two parents with two children and one of the children is younger than 15 while the other is 15 years or older. The arrows depict the links between the units of U^A and U^B. In the sample design considered in Section 2, each unit in U^A has exactly one unique link with a unit in U^B. Clusters in U^B have at least one link with units in U^A. Links are identified with an indicator variable

$$l_{ij} = \begin{cases} 1 & \text{if there is a link between } i \in U^A \text{ and } j \in U^B \\ 0 & \text{if there is no link between } i \in U^A \text{ and } j \in U^B. \end{cases}$$

If a unit *i* in U^A is selected in the sample, the entire cluster *k* to which this unit belongs, is included in the sample. The parameter of interest is the population total in U^B and is similar to (1) defined as $t_y = \sum_{k=1}^{M^B} \sum_{j=1}^{N_k^B} y_{kj}$. An estimator for t_y is defined as

$$\hat{t}_y = \sum_{k=1}^{m} \sum_{j=1}^{N_k^B} w_{kj} y_{kj} \, , \tag{3}$$

with *m* the number of unique clusters (households) included in the sample and w_{kj} the weight attached to each unit *j* of cluster *k*. Generally the inverse of the selection probabilities of units (*k,j*) observed in the sample are used as weights in the HT estimator. In this situation not all units in the sample have a known inclusion probability. Firstly not all units in U^B have a link to U^A. Secondly, as time proceeds household compositions change due to marriages divorces, departures of children and cohabitation. As a result, as time proceeds, units with a link to U^A enter the clusters in the sample although they are not initially included in the sample drawn from U^A. For these units inclusion probabilities are not necessarily known. They affect, however, the inclusion expectations of the clusters included in the sample. Reconstruction of the inclusion probabilities requires information of selection probabilities of all units in the population at the moment that the sample is drawn. In many practical situations this information is not available.

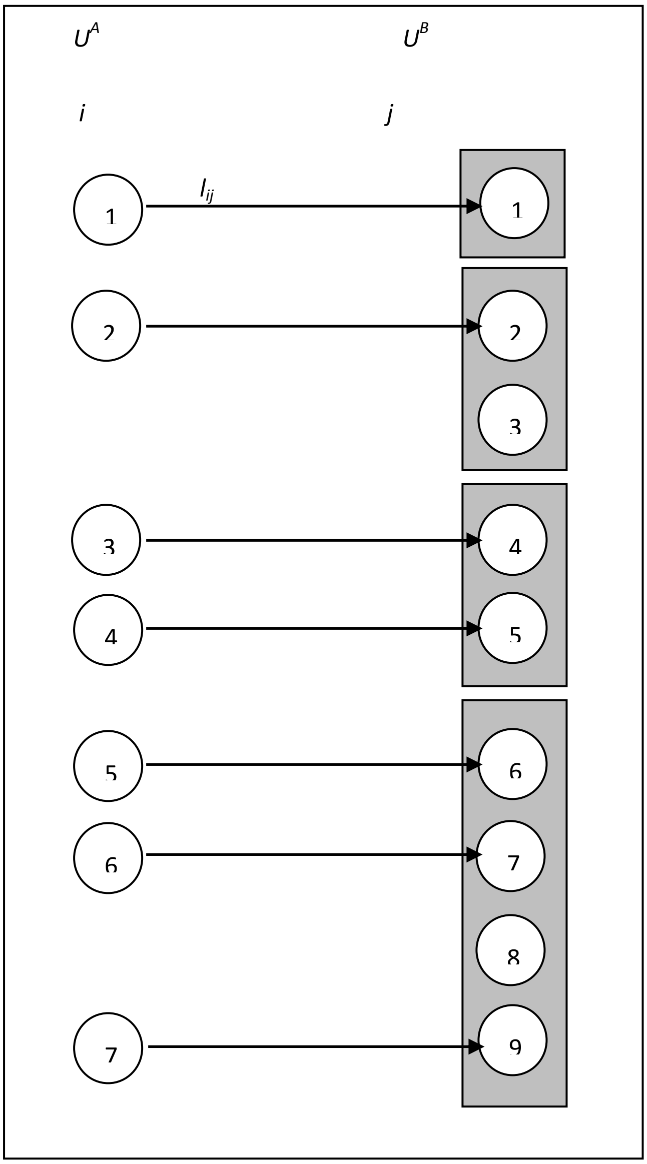

Figure 1: Links between units from the sample frame and units form the target population.

The Generalized Weight Share method can be used to derive non-zero weights for all units in the sample. This method starts with deriving initial weights, which are defined as

$$w^*_{kj} = \begin{cases} \dfrac{\delta_i^A}{\pi_i^A} & \text{if } (k, j) \text{ has a link with } i \in U^A \\ 0 & \text{otherwise} \end{cases},$$

with δ_i^A an indicator variable that is equal to one if *i* is included in the sample s^A and zero otherwise. This expression follows directly from Lavallée (1994), equation (2) in combination with the fact that in this application each unit in U^A has exactly one unique link with a unit in U^B, see Figure 1. In a second step a so-called basic weight for each cluster *k* is derived as the mean of all initial weights within each cluster:

$$w_k = \frac{\sum_{j=1}^{N_k^B} w_{kj}^*}{\sum_{j=1}^{N_k^B} l_{kj}},$$

which follows from Lavallée (1995), equation (3). Finally all persons *j* that belong to the same household *k* receive the same weight assigned to their household, i.e. $w_{kj} = w_k$ for all $j \in k$. A proof that the use of the basic weights in (3) is an unbiased estimator for the population total is also given by Lavallée (1995).

Let $\sum_{j=1}^{N_k^B} l_{kj} = g_k$ denote the number of persons in household *k* aged 15 years and older and a_k the number of core persons in household *k*, i.e. the number of persons in household *k* that are included in sample s^A. Since s^A is drawn by means of stratified simple random sampling, it follows that $\pi_i^A = n_h^A / N_h^A$ with N_h^A the number of persons aged 15 years and older in the population of stratum *h*, and n_h^A the number of core persons selected in the sample from stratum *h*. Then it follows that

$$w_k = \frac{a_k}{g_k} \frac{N_h^A}{n_h^A}. \qquad (4)$$

Inserting the first order inclusion expectation from Result 3.1 into (2) gives the same HT estimator as derived with the Generalized Weight Share method, i.e. inserting (4) into (3).

The derivation of the inclusion expectations in Subsection 3.1 applies to stratified sampling of households with inclusion expectations proportional to household size and is a special case of the Generalized Weight Share method. An argument to apply a design as outlined in Section 2 is that sampling households proportional to household size is efficient for target variables that are positively correlated with household size. It is useful to have explicit expressions for the first and second order inclusion expectations for sample size determination.

Lavallée (1995) also provides variance expressions for (3) based on the Generalized Weight Share method. This expression is based on the first and second order inclusion probabilities of the sample units drawn from U^A and a transformation of the target variable. As a result the property that clusters are drawn proportional to their size is not made explicit as well as the fact they are drawn partially with replacement. In section 7 it is pointed out that the variance expressions in Lavallée (1995) for this application are equal to the variance expressions based on the inclusion expectations derived in Result 3.1.

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 or area *r* are defined as

$$P_{lr} = \frac{M_{lr}}{M_{+r}}, \; l=1,\ldots,L, \qquad (5)$$

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 area *r*. This income distribution is estimated as

$$\hat{P}_{lr} = \frac{\hat{M}_{lr}}{M_{+r}}, \; l=1,\ldots,L, \qquad (6)$$

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

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

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 similarly to (5), (6), with M_{lr} the number of persons from area *r*, belonging to the *l*-th income category. The HT estimator for M_{lr} is now defined as

$$\hat{M}_{lr} = \sum_{h \in r} \sum_{k=1}^{m_h} \frac{1}{\pi_{kh}} \sum_{j=1}^{N_k} y_{kjhl},$$

Where $y_{kjhl} = 1$ if person *j* from household *k* and stratum *h* is classified to the *l*-th income class and $y_{kjhl} = 0$ otherwise.

For sample size determination, precision specifications for the estimated income distributions are required. For stratified sampling designs, Neyman allocations are often considered to determine minimum sample sizes and optimal allocations to meet precision requirements at aggregated levels, Cochran (1977). Power allocations are useful to find the right balance between precision requirements for aggregates and strata, Bankier (1988). In this application the minimum sample size is based on precision requirements for the individual strata, i.e. neighbourhoods.

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. As an alternative the square root of the mean over the variances of the estimated income classes of an income distribution is proposed as a precision measure for the estimated income distributions. With this measure the influence of the most imprecise income class on the minimum sample size will be reduced. The square root of the mean over the variances of the estimated income classes of an income distribution is called the average standard error measure and is defined as:

$$s = \sqrt{\frac{1}{L}\sum_{l=1}^{L} V(\hat{P}_{lr})} \ . \tag{7}$$

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 areas for which income distributions are published, precision requirements for sample size determination are specified at this 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 neighbourhood.

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 standard error measure* s_h *in (7) 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 HT estimator, Särndal et al. (1992), Section 2.8:

$$V(\hat{P}_{lh}) = \frac{1}{M_h^2}\sum_{k=1}^{M_h}\sum_{k'=1}^{M_h}(\pi_{kk'h} - \pi_{kh}\pi_{k'h})\frac{y_{khl}}{\pi_{kh}}\frac{y_{k'hl}}{\pi_{k'h}} \ . \tag{8}$$

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 (8) can be simplified to

$$V(\hat{P}_{lh}) = \frac{N_h - n_h}{n_h}\frac{1}{N_h - n_h}\left(\frac{N_h}{M_h^2}\sum_{k=1}^{M_h}\frac{y_{khl}}{g_{kh}} - \left(\frac{M_{lh}}{M_h}\right)^2\right). \tag{9}$$

Result 4.1 is obtained by inserting (9) into (7). ■

Remark: If $g_{kh} = 1$ for all households in the population of stratum *h*, then it follows that $M_h = N_h$ and that formula (9) simplifies to

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

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

Minimum sample size requirements based on Result 4.1 require information about the income distribution and its variance from preceding periods. Since this information is generally not available at the design phase of a panel, it is useful to have an upper bound for the average standard error measure for the income distribution in Result 4.1. This is comparable to taking the variance for a parameter defined as a proportion equal to its maximum value at 0.5 for calculating the minimum sample size for a survey.

Result 4.2: *An upper bound for the average standard error measure* s_h *for an income distribution, specified in Result 4.1 is given by*

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

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

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

$$\sum_{l=1}^{L} \sum_{k=1}^{M_h} \frac{y_{khl}}{g_{kh}} = \sum_{l=1}^{L} \sum_{t=1}^{T} \sum_{k=1}^{M_{th}} \frac{y_{khl}}{t} = \sum_{t=1}^{T} \frac{M_{th}}{t}. \qquad (10)$$

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

$$\sum_{l=1}^{L} \left(\frac{M_{lh}}{M_h} \right)^2 = \sum_{l=1}^{L} P_{lh}^2 \geq \frac{1}{L}. \qquad (11)$$

Result 4.2 is obtained by inserting (10) and (11) in the expression for *s* in Result 4.1. ■

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

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

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 standard error 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 standard error 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 (11).

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

$$n_h \geq \frac{\left(\frac{N_h}{M_h}\right)^2 \sum_{t=1}^{T} \frac{M_{th}}{t} - \frac{N_h}{L}}{(N_h - 1)L\Delta_h^2 + \frac{N_h}{M_h^2} \sum_{t=1}^{T} \frac{M_{th}}{t} - \frac{1}{L}} . \tag{12}$$

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 (12) 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

$$D_h = \mathrm{E}(\sum_{t=1}^{T}\sum_{k=1}^{M_{th}} I_{tkh}) = \sum_{t=1}^{T} M_{th}\tilde{\pi}_{th}$$

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

■

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*

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

The expected number of unique households and persons are random variables. It would therefore be useful to have an uncertainty measure for these expected values. Variance expressions for Results 4.3 and 4.4 are however not straightforward and therefore left for further research.

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 0.16. With this sample size, the maximum value for the average standard error measure s_h at the level of neighbourhoods amounts to about 0.01 for the estimated household income distributions. With a total population of about 12 million persons, this resulted in a sample size of about 2.1 million core persons and an expected sample size of about 4.6 million 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. A first requirement for correct cross-sectional inference with this panel is to have correct first and second order inclusion expectations for the sampling units, which are derived in Section 3. A second requirement for correct cross-sectional inference is to keep the panel representative for the target population. To this end, it is determined on a yearly basis 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 0.16 is selected. These core persons are added to the panel of the RIS, with the purpose to maintain a representative sample.

6. Linear weighting

For household surveys like the RIS, 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:

$$\hat{t}_y = \sum_{h=1}^{H} \sum_{k \in 1}^{m_h} \sum_{j \in k} w_{kjh} y_{kjh} \, , \tag{13}$$

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

$$\hat{t}_y = \sum_{h=1}^{H} \sum_{k=1}^{m_h} w_{kh} y_{kh} \, , \tag{14}$$

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

Weights are obtained by means of the GREG estimator to use auxiliary variables which are observed in the sample and for which the population totals are known from other sources, Särndal et al. (1992). 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. To ensure that relationships between household variables and person variables are reflected in estimated totals, it is relevant to apply a weighting method which yields one unique household weight for all its household members. 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). This can be achieved with the so-called integrated weighting methods.

Lemaître and Dufour (1987) applied an integrated weighting method at the person's level and replaced the original auxiliary variables defined at the person level by the corresponding household mean. In this way, members of the same household have the same inclusion expectation and share the same auxiliary information, and therefore the resulting regression weights are forced to be the same. Nieuwenbroek (1993) proposed a slightly more general approach by applying the linear weighting method at the household level, where the auxiliary information of person based characteristics is aggregated at the household level. Nieuwenbroek (1993) mentions that the linear weighting method at the household level is equal to the linear weighting method of Lemaître and Dufour at the person level, if the residual variance of the regression model at the household level is chosen proportional to the number of persons within the household. Integrated weighting of person and household surveys is further generalized by Steel and Clark (2007) and Estevao and Särndal (2006). Steel and Clark (2007) addressed the issue whether the cosmetic benefits of integrated weighting result in an increased design variance of the GREG estimates. They showed that large sample design variances obtained by linear weighting at the household level is less than or equal to the design variance obtained with linear weighting at the person level. For small samples there can be a small increase in the design variance due to integrated weighting. As a result there is little or no loss in efficiency by applying an integrated weighting method.

In this paper the integrated weighting approach at the household level is applied. 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 GREG estimator is derived from a linear regression model that specifies the relation between the target variable and the available auxiliary variables for which population totals are known, and is defined as:

$$y_{kh} = \mathbf{x}_{kh}^{t}\boldsymbol{\beta} + e_{kh}\text{ , with }\mathrm{E}_{m}(e_{kh}) = 0\text{ , }\mathrm{V}_{m}(e_{kh}) = \sigma_{kh}^{2}\text{ .} \qquad (15)$$

In (15) $\boldsymbol{\beta}$ 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. In this application, the variance structure is taken proportional to the household size, i.e. $\sigma_{hk}^{2} = g_{kh}\sigma^{2}$. In this case, the weighting applied at the household level is equal to Lemaître and Dufours method as shown by Nieuwenbroek (1993).

Regression weights for the households are finally obtained by

$$w_{kh} = \frac{1}{\pi_{kh}}\left(1 + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}\boldsymbol{\pi}})^{t}\left(\sum_{k=1}^{m}\frac{\mathbf{x}_{kh}\mathbf{x}_{kh}^{t}}{\pi_{kh}g_{kh}}\right)^{-1}\frac{\mathbf{x}_{kh}}{g_{kh}}\right),$$

with $\mathbf{t}_{\mathbf{x}}$ a *q* vector containing the known population totals of the auxiliary variables $\mathbf{x}$, $\hat{\mathbf{t}}_{\mathbf{x}\boldsymbol{\pi}}$ the HT estimator for $\mathbf{t}_{\mathbf{x}}$. The weights calculated at the household level can be used for weighting person based characteristics of the corresponding household members, using formula (13) since $w_{kjh} = w_{kh}$ for all persons belonging to the same household *k*.

7. Variance estimation

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

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_z}, \tag{16}$$

where $\hat{t}_y$ and $\hat{t}_z$ are GREG estimators defined by (13) or (14) in the case of person-based or household-based target variables, respectively.

Result 7.1: *The variance of (16) 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 GREG estimators is given by (Särndal et al. 1992, Section 7.13):

$$V(\hat{R}) = \frac{1}{t_z^2}\sum_{h=1}^{H}\sum_{k=1}^{N_h}\sum_{h'=1}^{H}\sum_{k'=1}^{N_{h'}}(\pi_{kk'hh'} - \pi_{kh}\pi_{k'h'})\frac{e_{kh}}{\pi_{kh}}\frac{e_{k'h'}}{\pi_{k'h'}}. \tag{17}$$

After inserting first and second order inclusion expectations specified in Result 3.1, it follows that (17) 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 HT type estimators for* $\mathbf{b}_y$ *and* $\mathbf{b}_z$*, defined by (6.5).*

Proof: An estimator for the variance approximation (17) is given by (Särndal et al. 1992, Section 7.13):

$$V(\hat{R}) = \frac{1}{\hat{t}_z^2}\sum_{h=1}^{H}\sum_{k=1}^{N_h}\sum_{h'=1}^{H}\sum_{k'=1}^{N_{h'}}\frac{(\pi_{kk'hh'} - \pi_{kh}\pi_{k'h'})}{\pi_{kk'hh'}}\frac{c_{kh}\hat{e}_{kh}}{\pi_{kh}}\frac{c_{k'h'}\hat{e}_{k'h'}}{\pi_{k'h'}}, \tag{18}$$

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 (18) 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. ■

The same expressions for the variance can be derived from the variance expressions proposed for the Generalized Weight Share method in the case of indirect sampling. In Lavallée (1995), variance expressions for the HT estimator are based on the sampling design used to select the sample s^A of n units from population U^A with transformed target variables, say z_i. In this application each unit in U^A has exactly one link with a unit in U^B. As a result z_i in Lavallée (1995) is in this case defined as the sum over the target variables of all elements in cluster *k*, divided by the number of units in cluster *k* with a link to population U^A, i.e. $z_i = y_k / g_k$ for al $i \in U^A$ that have a link with cluster $k \in U^B$. Inserting the first and second order inclusion probabilities for stratified simple random sampling without replacement and the transformed variables z_i (where the target variable y_k is preplaced by the residual of the regression on the cluster totals e_k) in the variance formula for a ratio gives Result 7.1. Result 7.2 follows in a similar way.

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 quantile points of the national distribution (abbreviated as Inc. distr. hh.)
- Mean standardized income of households (abbreviated as HHinc)
- Mean standardized income of persons (abbreviated as Pinc)

For all three variables standardized income is used, which is defined as the disposable income corrected for differences in household size and composition. Standardized income is a generally applied measure for welfare and income. According to the definition, all members of the same household have the same standardized income because all members of the same household receive or share the same amount of welfare. As a result the mean household income and the mean personal income are very close. In the latter, the standardized income of larger households have a larger share.

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 marginals of different publication tables. Inclusion expectations are based on the formulas derived in Subsection 3.1. 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 numbers 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 slightly increases the standard error compared to the regular GREG estimator. This suggests that the assumed variance structure for the residuals in the underlying regression model in the case of integrated weighting better fits to the household-based variables than the person-based variables.

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 due to their instability over time inappropriate as sampling units in panels conducted to collect information at the level of households or persons. In this paper, a sample design is proposed 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. Households are included with expectations proportional to the household size. First and second order inclusion expectations for households are derived under stratified simple random sampling of core persons. These inclusion expectations can be used in a similar way in design-based and model-assisted inference as the more common inclusion probabilities.

The sample design in this paper is a special case of indirect sampling, Lavallée (1995, 2007). In the case of a self-weighted sample design 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. In the case of more complex sample designs, the Generalized Weight Share method, developed by Lavallée (1995, 2007), is required to construct inclusion weights at each point in time.

Since core persons remain in the panel indefinitely, this sample design is particularly appropriate for register-based household panels where all required information is derived from administrations. For interview-based household panels some kind of rotation design is required to cope with problems like panel attrition. 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 for individual strata, which are the most detailed areas for which figures are published. A topic for further research is to combine this mean standard error measure with a Neyman allocation or power allocations to have expressions for the minimum sample size based on precision requirements for estimated distributions at aggregates of strata.

In the context of household surveys and panels, 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 integrated weighting approach based on the procedure proposed by Lemaître and Dufour (1987) is applied to the RIS. In this application standard errors obtained with Lemaitre and Dufour are smaller compared to a non-integrated weighting procedure for household based estimates. For person based estimates, standard errors can be slightly larger. These results are in line with Steel and Clark (2007), who showed that the large sample design variance of integrated weighting at the household level are smaller than or equal to the design variance obtained with non-integrated weighting at the person level. In their simulation they also report small increases of the design-variances due to integrated weighting in the case of small sample sizes. The additional advantage of integrated weighting is that totals for household and person based income, which can be derived directly from their means, are consistent.

References

Bankier, M.D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42, pp. 174-177.

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

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

Deville, J.C. and P. Lavallée (2006). Indirect Sampling: The Foundations of the Generalized Weight Share Method, *Survey Methodology*, 32, pp. 165-176.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. In Panel Surveys. (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). John Wiley & Sons, New York, pp. 135-159.

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.

Kalton, G. and J.M. Brick (1995). Weighting Schemes for Household Panel Surveys. *Survey Methodology*, 21, pp. 33-44.

Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households using the Weight Share Method. *Survey Methodology*, 21, pp. 25-32.

Lavallée, P. (2007). *Indirect Sampling*, Springer Verlag, New York.

Lemaître, G. and J. Dufour (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., 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.

Technical appendix: Proof of Result 3.1

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

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

with a_{kh} the number of times that household *k* from stratum *h* is selected. The enumerator of the ratio in (A.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 (A.1) can be expressed as

$$\pi_{kh} = \sum_{i=1}^{g_{kh}} g_{kh} \frac{\binom{g_{kh} - 1}{i - 1}\binom{N_h - g_{kh}}{n_h - i}}{\binom{N_h}{n_h}}. \qquad \text{(A.2)}$$

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

$$\sum_{j=0}^{n} \binom{a}{j}\binom{b}{n - j} = \binom{a + b}{n}. \qquad \text{(A.3)}$$

By changing to $j = i - 1$ and applying formula (A.3), it follows that (A.2) can be simplified to

$$\pi_{kh} = g_{hk} \sum_{j=0}^{g_{hk} - 1} \frac{\binom{g_{hk} - 1}{j}\binom{N_h - g_{hk}}{n - j - 1}}{\binom{N_h}{n_h}} = g_{kh} \frac{\binom{N_h - 1}{n_h - 1}}{\binom{N_h}{n_h}} = g_{kh} \frac{n_h}{N_h}. \qquad \text{(A.4)}$$

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

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

(A.5)

Using similar arguments as specified following equation (A.1), the ratio in (A.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 (A.5) can be simplified to

$$\pi_{kk'h} = \sum_{i=1}^{g_{kh}} \sum_{i'=1}^{g_{k'h}} g_{kh} g_{k'h} \frac{\binom{g_{kh}-1}{i-1}\binom{g_{k'h}-1}{i'-1}\binom{N_h - g_{kh} - g_{k'h}}{n_h - i - i'}}{\binom{N_h}{n_h}}. \qquad \text{(A.6)}$$

By changing to $j = i - 1$ and $j' = i' - 1$ and applying formula (A.3) twice, it follows that (A.6) simplifies to

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

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

(A.7)

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

$$\pi_{kkh} = \mathrm{E}(a_{kh} a_{kh}) = \mathrm{E}(a_{kh}(a_{kh} - 1)) + \mathrm{E}(a_{kh}). \qquad \text{(A.8)}$$

An expression for the first order inclusion expectation $\mathrm{E}(a_{kh})$ is already given by (A.4). The first term on the right hand side of (A.8) can be elaborated as follows:

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

$$= \sum_{i=2}^{g_{kh}} g_{kh}(g_{kh} - 1) \frac{\binom{g_{kh} - 2}{i - 2}\binom{N_h - g_{kh}}{n_h - i}}{\binom{N_h}{n_h}}. \qquad \text{(A.9)}$$

Changing to $j = i - 2$ and applying formula (A.3) gives

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

$$= g_{kh}(g_{kh}-1)\frac{n_h(n_h-1)}{N_h(N_h-1)}. \tag{ A.10}$$

Inserting the expressions (A.4) and (A.10) into (A.8) gives

$$\pi_{kkh} = g_{kh}(g_{kh}-1)\frac{n_h(n_h-1)}{N_h(N_h-1)} + g_{kh}\frac{n_h}{N_h}. \tag{A.11}$$

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

$$\pi_{kk'hh'} = \mathrm{E}(a_{kh}a_{k'h'}) = \mathrm{E}(a_{kh})\mathrm{E}(a_{k'h'}) = g_{kh}g_{k'h'}\frac{n_h n_{h'}}{N_h N_{h'}}. \tag{A.12}$$

This result is straightforward, since samples in different strata are drawn independently from each other. Collecting the results obtained in formula's (A.4), (A.8), (A.11), and (A.12), proves Result 3.1. ∎

Explanation of symbols

.	Data not available
*	Provisional figure
**	Revised provisional figure (but not definite)
x	Publication prohibited (confidential figure)
–	Nil
–	(Between two figures) inclusive
0 (0.0)	Less than half of unit concerned
empty cell	Not applicable
2014–2015	2014 to 2015 inclusive
2014/2015	Average for 2014 to 2015 inclusive
2014/’15	Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015
2012/’13–2014/’15	Crop year, financial year, etc., 2012/’13 to 2014/’15 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

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

Prepress
Studio BCO, Den Haag

Design
Edenspiekermann

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