Journal of Official Statistics, Vol. 27, No. 3, 2011, pp. 415–432

A Unit-Error Theory for Register-Based Household Statistics

Li-Chun Zhang[1]

The next round of censuses will be completely register-based in all the Nordic countries. Household is a key statistical unit in this context, which however does not exist as such in the administrative registers available, and needs to be created by the statistical agency based on various information available in the statistical system. Errors in such register households are thus unavoidable, and will propagate to various induced household statistics. In this article I outline a unit-error theory which provides a framework for evaluating the statistical accuracy of these register-based household statistics, and illustrate its use based on the Norwegian register household data.

Key words: Register statistics; statistical accuracy; unit errors; prediction inference.

1. Introduction

For some decades now administrative registers have been an important data source for official statistics alongside survey sampling and population census. Not only do they provide frames and valuable auxiliary information for sample surveys and censuses, also systems of inter-linked *statistical* registers (i.e. registers for statistical purposes) have been developed on the basis of available administrative sources to produce a wide range of purely register-based statistics (e.g. Wallgren and Wallgren 2007). For instance, the next census will be completely register-based in all the Nordic countries (UNECE 2007). Reduction of response burden, long-term cost efficiency and the potential for detailed spatial-demographic and longitudinal statistics are some of the major advantages associated with the use of administrative registers.

The trend is increasingly being recognized by statistical offices around the world (Holt 2007, Section 3.1.2). However, also being noticed is that there is clearly a lack of *statistical* theories for assessing the quality of register-based statistics. Administrative registers certainly do not provide perfect statistical data. Sampling errors are naturally absent. But there exist a variety of nonsampling errors such as over- and under-coverage, lack of relevance, misclassification, delays and mistakes in the data registration process, inconsistency across the administrative sources, and not least, missing data. I believe that a key issue here, from a statistical methodological point of view, is the *conceptualization* and *measurement* of the *statistical accuracy* in register data, which will enable us to apply rigorous statistical concepts such as bias, variance, efficiency and consistency, as one is able to do when it comes to, for example, survey sampling.

[1] Statistics Norway, Kongensgate 6, PB 8131 Dep, N-0033 Oslo, Norway. Email: lcz@ssb.no

In this article I outline a statistical theory for *unit errors* in register-based household statistics. Unit errors as such are rarely mentioned in the literature of survey sampling and census. The main reason may be that while the statistical offices collect their own data in surveys and censuses explicitly for producing statistics, the administrative data are by default created and maintained by external register owners for administrative purposes. One of the problems this can cause is that the statistical units of interest simply do not exist as such in the administrative registers, and must be established by the statistical agency in order to obtain the relevant statistical data.

Household is a typical example in this respect. In Norway the central population register (CPR) provides the basis of most population statistics. Each resident of Norway is associated with a person identity number (PIN) in the CPR. Persons in the CPR can be grouped into families through the family identity number (FIN) in the CPR. Here by definition a *(CPR) family* consists of persons who are permanently resident in the same dwelling, and who are linked to each other as spouses, cohabitants, registered partners and/or parents and unmarried children (regardless of age). In addition, a family can, at most, consist of two (consecutive) generations and only one married/cohabiting couple. For statistical purposes, however, household is more often of interest. By definition a *(private) dwelling household* consists of persons who are permanently resident in the same dwelling, where the dwelling is not an institution. A household may thus contain more than one family. For instance, two single-person families may constitute a household of cohabiting couples without children, or grandparents (who constitute family on their own) may live in the same household together with a family of younger generations.

Since there does not exist any household identity number in the administrative system, a household register (HR) of private dwelling households needs to be constructed for statistical purposes. A key data source in this respect is the dwelling register (DR), which is a part of the Ground Property, Address and Building Register (SN-GAB). Here by definition a *dwelling* is a residential unit consisting of one or more rooms built or rebuilt as an all-year-round private residence for one or more persons. In Norway there are two types of addresses, street address and cadastral address. Street addresses are used in towns, whereas cadastral addresses are used in the countryside. In both cases, there can be multiple dwelling units at a single address, such that a dwelling unit cannot always be identified by the address alone. This is for instance because traditionally the apartment number was not registered as a part of the street address. The registration of dwelling units in the DR was initiated in connection with the last census in 2001. The dwelling identity number (DIN) was collected in the housing census and stored in the CPR, and it is updated when a person registers for moving at the Municipal Administration. Furthermore, the DIN is updated in the DR as new buildings are constructed and old buildings are rebuilt. However, neither the updating of the two registers nor the communication between the two sources is perfect. As a result the DIN in the CPR can be mistaken or missing. A fair amount of editing and imputation is needed in order to establish the HR. Errors occur in the *register households* (i.e. households according to the HR) whenever people who do not live together are grouped into the same household, and/or when people in the same household are divided into different households. We call such errors the unit errors.

As an illustrative example, consider the household data in Table 1. The ID numbers given are generic, not the real ones in use. The statistical unit of interest is household.

Table 1. Household data at Storgata 99: Reality vs. household register

	Reality						
Dwelling ID	Family ID	Household ID	Person ID	Name	Sex	Age	Income
H101	1	1	1	Astrid	Female	72	y_1
H102	2	2	2	Geir	Male	35	y_2
H102	2	2	3	Jenny	Female	34	y_3
H102	2	2	4	Markus	Male	5	y_4
H201	3	3	5	Knut	Male	29	y_5
H201	4	3	6	Lena	Female	28	y_6
H202	5	4	7	Ole	Male	28	y_7

	Household register						
Dwelling ID	Family ID	Household ID*	Person ID	Name	Sex	Age	Income
H101	1	1	1	Astrid	Female	72	y_1
H101	2	2	2	Geir	Male	35	y_2
H101	2	2	3	Jenny	Female	34	y_3
H101	2	2	4	Markus	Male	5	y_4
H101	3	3	5	Knut	Male	29	y_5
–	4	4	6	Lena	Female	28	y_6
–	5	4	7	Ole	Male	28	y_7

A household ID has been created in the HR, which is marked by * in the table to show that it may be erroneous. The errors here are due to poor quality of the DR since only H101 can be found at Storgata 99, as well as poor registration of the DIN in the CPR. The combined result is that the DIN is duplicated for persons no. 1–5 and missing for persons no. 6–7 in the HR.

A few things are worth noting. (i) The HR contains unit errors for Knut, Lena and Ole: in reality Knut and Lena belong to one household and Ole to another, whereas according to the HR Lena and Ole belong to the same household and Knut to another. (ii) The unit error might have occurred for all the seven persons at Storgata 99. The register households are actually correct for Astrid, and for Geir, Jenny and Markus, but one would not be able to know that for sure, given possible mistakes in the dwelling IDs. A statistical theory is therefore needed in order to evaluate the uncertainty in register household data, no matter how good the quality of the underlying registers may be, as long as they are not error-free in reality. (iii) Unit errors in households will carry over to all household statistics such as household income or population demographic statistics, which may or may not have severe consequences. A unit-error theory should enable us to propagate the uncertainty to such induced household statistics. (iv) Household is a unit of central interest in the coming register-based census, so a statistical theory that accounts for the uncertainty due to the unit errors in register households is desirable in this respect.

I have named the household errors in Table 1 unit errors. The mapping from the persons (or families) to the households has some resemblances to the matching of two sets of units in record linkage (e.g. CENEX-Project 2006–2008). There is nevertheless an important difference. In record linkage, the numbers of units in the two data sets are fixed. In my case, however, the number of households at a given address is generally unknown and needs to be determined at the same time as the households are being ‘constructed.’ A unit-error problem therefore involves more than just mismatching between two sets of fixed units. There are two reasons why the number of households is unknown. Firstly, the DR is not perfect. There are both missing and wrongly registered DINs in the DR, as well as delays in updating, such that the number of dwelling units at a given address cannot be known for sure. Secondly, even when the errors in the DR are disregarded, it is not true that the number of dwelling households will always be the same as the number of dwelling units. For instance, in the Netherlands, the address is considered complete enough to function as the DIN. Yet people at the same address may be classified into different households, as was done in the last Virtual Census (Harmsen and Isarels 2003). Moreover, unit errors will almost certainly arise in a longitudinal perspective, because the updating of the DIN in the population register is not perfect (e.g. Van der Laan et al. 2009). The problem of unit errors is thus also relevant in the Dutch case.

Finally, we notice that unit errors are not limited to the household data. As mentioned earlier, the issue may be relevant whenever the statistical unit of interest cannot be found in the existing sources, and needs to be created by the statistical agency. For instance, in many countries the business unit enterprise needs to be ‘constructed’ from smaller legal units. While it may be possible to determine all the legal units involved at the enterprise-group (EG) level, it is not given how many enterprises an EG should be divided into.

The rest of the article is organized as follows. In Section 2 I introduce a mathematical representation of the unit errors in register households, as well as the various household statistics derived from the register households. A prediction inference framework is outlined in Section 3. In Section 4, I illustrate the proposed unit-error theory using the Norwegian register household data. Finally, a summary and some discussion will be set forth in Section 5.

2. A Mathematical Representation

2.1. Allocation Matrix

We assume that the *target (statistical) unit* consists of one or more *base units*. The base units are atomic components that are never to be broken up when the target units are being created. Unit errors arise then from *allocating* base units into wrong target units. In our illustrative example above, the target unit is household. The base unit can be person. But it can also be family identified by the family ID, provided the adopted definitions allow for that. If applicable, the latter choice is more convenient because it reduces the combinatorial complexity of *allocation*.

We may express the mapping from base units to target units by means of an *allocation matrix* A, where $a_{ji} = 1$ if the base unit i is allocated to the target unit j, and $a_{ji} = 0$ otherwise, for base unit $i = 1, \ldots, m$. The allocation matrix has dimension $m \times m$ and can be up to rank m, in which case it is a *permutation matrix*, i.e. a matrix obtainable from the identity matrix through a row permutation, and every base unit constitutes a target unit by itself. But there will be redundant rows of zeros if there are fewer target units than base units. Notice that we do not assume that the number of target units is known.

Given persons as the base units, listed as in Table 1, the correct allocation matrix, denoted by A, and the matrix that corresponds to the household register, denoted by A^*, are given by

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad A^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

In this way, errors in the target units now correspond to errors in the allocation matrix A^*.

Now, the target units obviously remain the same under any row permutation of the allocation matrix, except for the ordering among them. For uniqueness of the allocation matrix, it is necessary to impose a row ordering. Given the ordering of the base units, indexed as $i = 1, \ldots, m$, let k_j be the column number of the *first* nonzero element in the jth row of the allocation matrix, provided it exists. In other words, k_j is the index of the first base unit in the jth target unit. For example, in the matrix A above, we have $k_1 = 1$, $k_2 = 2$, $k_3 = 5$ and $k_4 = 7$. Whereas in A^*, we have

$k_1^* = 1 = k_1$, $k_2^* = 2 = k_2$, $k_3^* = 5 = k_3$ and $\mathrm{k}_4^* = 6 \neq k_4$. We assume that the rows of an allocation matrix are ordered such that $k_{j'} \leq k_j$ provided $j' < j$. By such an ordering an allocation matrix becomes *sequential upper triangular*, where an upper triangular matrix is said to be sequential in addition provided it will remain upper triangular after deletion of any number of the *first* rows and columns, as long as it is not all zero afterwards.

We notice that the allocation matrix is a generalization of the permutation matrix mentioned earlier. A permutation matrix contains a single element of ones in each row and column. The allocation matrix, however, may contain multiple elements of ones in the same row as well as possible rows of all zeros, but it still has a single element of one in each column because a base unit can only belong to one and only one target unit.

A permutation matrix is useful in a special record linkage situation, where the two sets of units are identical (Chambers 2008). Given the ordering of the two sets of units and without losing generality, let the correct but unknown linkage be given by the vector $\mathbf{y} = (y_1, \ldots, y_d)$, where y_i is the order of the unit in the *second* set to which the ith unit in the *first* set should be linked and d is the number of units in both sets. Notice that $\mathbf{y}$ must be a permutation of the vector $(1,2,\ldots,d)$ since each unit in the first set must be matched to a distinct unit in the second set. Let $\mathbf{y}^*$ be the actual linkage, which is also a permutation of the vector $(1,2,\ldots,d)$. In particular, $\mathbf{y}^*$ can be obtained from $\mathbf{y}$ through a linear transformation using a permutation matrix $B_{d\times d}$, i.e. $\mathbf{y}^* = B\mathbf{y}$, where correct linkage of the ith unit in the first set corresponds to $b_{ii} = 1$, and an incorrect linkage entails that $b_{ij} = 1$ for some $j \neq i$. A simple parametric model of B can be given by $P(b_{ii} = 1) = \lambda$ and $P(b_{ij} = 1) = \gamma$ for $j \neq i$, where $\lambda + (d-1)\gamma = 1$. That is, one assumes that all possible incorrect linkages are equally likely. Depending on the available information about the linkage process, more sophisticated models for linkage errors can be formulated. See Chambers (2008) for a more detailed discussion.

In the application later, use is made of a simple multinomial model, where each distinct allocation matrix is assigned its own probability, after conditioning on certain characteristics of the base units involved and the corresponding allocation matrix in the register. The multinomial model of the true allocation matrix and the associated inference will be explained in Section 3.

2.2. Value Matrix and Statistical Variables of Interest

To facilitate statistics of the units of interest, we define a *value matrix*, or *vector*, X for the involved base units, such that the statistical variables of interest can be obtained as a function of the allocation matrix and X. Often the variables of interest can simply be expressed as a linear transformation of X through the allocation matrix. But it can also be a nonlinear function of such simple linear transformations. Some examples may help to clarify this.

Example 1 Value matrix $X = I_{m\times m}$, i.e. the identity matrix, yields target unit inclusion, indicating which base units are included in which target unit by definition of the allocation matrix.

Example 2 Value vector $1_{m\times 1}$ yields the target unit sizes, defined as the number of base units that constitute the target unit. Thus, for A given above, we obtain

$$A1 = (1, 3, 2, 1, 0, 0, 0)^T$$

Example 3 Value matrix $X = \mathrm{Diag}(y)_{m \times m}$ can be used to group the y-values of the base units according to which target unit they belong to. Thus, for A above, we obtain

$$A\,\mathrm{Diag}(y) = \begin{pmatrix} y_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & y_2 & y_3 & y_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & y_5 & y_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & y_7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The value vector $\mathbf{y} = (y_1, \ldots, y_7)^T$ yields the target unit y-total, such as household income, by

$$A\mathbf{y} = (y_1, y_2 + y_3 + y_4, y_5 + y_6, y_7, 0, 0, 0)^T$$

Example 4 The target unit mean y-value, such as mean household income above, can thus be given as a nonlinear function

$$(A\mathbf{y})//(A1) = (y_1, (y_2 + y_3 + y_4)/3, (y_5 + y_6)/2, y_7, -, -, -)^T$$

where "//" denotes component-wise division provided a nonzero denominator.

Example 5 Value vector of sequels, denoted by $\alpha = (1,2, \ldots ,m)^T$, yields target unit identifier when multiplied on the left by the transpose of the allocation matrix. For A above, we obtain

$$A^T\alpha = (1, 2, 2, 2, 3, 3, 4)^T$$

Example 6 Suppose in the example above we would like to obtain household age composition for 4 age groups: <18, 18–30, 31–65 and >66. We may use the dummy-index value matrix

$$X = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad \text{giving} \quad AX = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

2.3. Blocking and Strata of Blocks

According to the adopted household definition, all the household members must be resident in the same dwelling unit. Since no dwelling unit can be divided between different addresses, the addresses (such as "Storgata 99" in our illustrative example) in the DR can be used to divide the target population into smaller groups called *blocks* in our application later, where allocation (of households) is delimited within each block. In other words, no base units from different blocks can be allocated to the same target unit. Blocking is important in practice because it reduces the dimension of the data. Notice that blocks here do not refer to building blocks or street blocks as such. Rather the blocks form a conceptual division of the target population. Using the Dutch household data as an example, a block may be a dwelling unit, which may consist of more than one household. Or, to take another example in business statistics, each enterprise-group may form a block consisting of all the legal units that can possibly join each other in one or another enterprise.

Next, *Strata* of blocks can be formed that have strong stratum-specific distributional characteristics of the allocation matrix. The number of base units inside a block is naturally a stratum variable. For instance, there are only two possible allocation matrices for blocks of two base units:

$$A_1 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \quad A_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

which are quite different from the five possible allocation matrices for blocks of three base units:

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

But also other auxiliary information can be used as stratum variables. In the case of household, whether there are parents with children within a block, the age and sex combinations of the residents, etc. can all have an impact on the frequency of the possible allocation matrices. For instance, given a block of two persons, the chance that the true allocation matrix is A_1 above, i.e. they belong to the same household, will be close to unity provided they are married to each other according to the CPR. Whereas the probability can be much lower otherwise. An overview of the definition of the strata in our application later can be found in Table 3.

3. Inference

3.1. Prediction Expectation and Variance of a Target Population Total

Suppose that the population is divided into strata of blocks, denoted by $h = 1, \ldots, H$. Denote by (hq) the qth block within the hth stratum, where $q = 1, \ldots, M_h$ and M_h is the number of blocks in the stratum. Denote by A_{hq} the true allocation matrix for block (hq), and denote by X_{hq} the corresponding value matrix (or vector), such that the values of interest associated with the corresponding target units can be given as a function of $A_{hq}X_{hq}$, denoted by

$$t_{hq} = g_h(A_{hq}X_{hq})$$

Consider as the target of interest a population total T, given by

$$T = \sum_{h=1}^{H} T_h = \sum_h \left(\sum_{q=1}^{M_h} t_{hq} \right) = \sum_h \sum_q g_h(A_{hq}X_{hq}) \tag{1}$$

Of course, other expressions of T are also possible, such as a weighted average over (hq) or a nonlinear function of several population totals. The corresponding expectation and variance expressions will then be different from (2) and (3) below, but can be derived similarly.

Notice that the functional g_h is allowed to vary in different strata. As an example in the application later, let T be the vector of population totals of households by size k, for $k = 1, 2, \ldots, K$. Let each address form a separate block. Let $X_{hq} = (x_1, x_2, \ldots, x_{m_{hq}})^T$ represent the sizes of the families (i.e. base units), where m_{hq} is the number of families at address (hq). Suppose first that the strata are formed such that $m_{hq} = m_h$ is a constant in each stratum h. Let $\mathrm{n}_{hq} = A_{hq}X_{hq} = (n_1, \ldots, n_{mh})^T$, where n_i is the size of household i at address (hq) for $1 \leq i \leq m_h$. Notice that $n_i = 0$ if the ith row in A_{hq} contains only zeros. Then $t_{hq} = g_h(A_{hq}X_{hq})$ can be given by a vector of length K, where the kth component is given by $\sum_{i=1}^{m_h} I_{n_i=k}$, and $I_{n_i=k} = 1$ if $n_i = k$ and $I_{n_i=k} = 0$ otherwise. Suppose next that m_{hq} may vary in each stratum h, i.e. the number of families at each address may vary for the addresses in a stratum. Let $m_0 = \max_{h,q} m_{hq}$. Let now X_{hq} be a vector of constant length m_0, where $x_i = 0$ for $m_{hq} < i \leq m_0$. Consequently, the vector n_{hq} will have the length m_0, and $g_h(A_{hq}X_{hq})$ can take the same form $g(A_{hq}X_{hq})_{m_0 \times 1}$, where the kth component is given by $\sum_{i=1}^{m_0} I_{n_i=k}$, and $I_{n_i=k} = 1$ if $n_i = k$ and 0 otherwise.

Let A^*_{hq} be the allocation matrix at block (hq) in a statistical register such as the HR, where A^*_{hq} is known throughout the population. We assume that, within the hth stratum, (A_{hq}, A^*_{hq}) are jointly independently and identically distributed across the blocks, for $q = 1, \ldots, M_h$. Then, conditional on $A^*_h = \left\{ A^*_{hq}; q = 1, \ldots, M_h \right\}$ and $A^* = U^H_{h=1} A^*_h$, the best prediction of the target total T is given by its conditional expectation

$$E(T|A^*) = \sum_h E\left(T_h|A^*_h\right) = \sum_h \left(\sum_q \mu_{hq} \right) \quad \text{where} \quad \mu_{hq} = E\left(t_{hq}|A^*_{hq}\right) \tag{2}$$

taken with respect to the conditional distribution of A_{hq} given A^*_{hq}, denoted by $f_h\left(A_{hq}|A^*_{hq}\right)$. Moreover, let $V(T|A^*)$ denote the prediction variance with respect to the same conditional distribution. We have

$$V(T|A^*) = \sum_h V\left(T_h|A^*_h\right) = \sum_h \left(\sum_q \tau_{hq} \right) \quad \text{where} \quad \tau_{hq} = V\left(t_{hq}|A^*_{hq}\right) \tag{3}$$

In the application later, independence across the blocks means that the way households are formed by the families at one address does not depend on those at another address. This seems reasonable. The assumption of identical distribution depends on how well the addresses can be divided into homogeneous strata. In the application, the strata are formed on the basis of an analysis of the relationship between census households and CPR families, where one seeks to identify groups of addresses with clearly distinct distributions of the pairwise allocation matrices. As mentioned earlier, the number of base units inside a block, i.e. families at a given address, is a natural stratum variable. The other most important factor turns out to be whether or not there are couples at a given address according to the CPR, to be referred to as the CPR couples. Of course, as with all statistical modeling, some degree of misspecification will necessarily be present. More discussion will be offered in Section 4.3.

3.2. *Estimation of Prediction Expectation and Variance*

In practice, of course, we need to estimate the distribution $f_h\left(A_{hq}|A^*_{hq}\right)$. Suppose we have available an *audit sample*, where A_{hq} can be identified. It is then possible to obtain an estimate of f_h, denoted by $\hat{f}_h\left(A_{hq}|A^*_{hq}\right)$. An estimate of the prediction expectation $E(T|A*)$ is then given by

$$\hat{E}(T|A*) = \sum_h \hat{E}(T_h|A^*_h) = \sum_h \left(\sum_q \hat{\mu}_{hq}\right) \quad \text{where} \quad \hat{\mu}_{hq} = E\left(t_{hq}|A^*_{hq}; f_h = \hat{f}_h\right)$$

i.e. the expectation (2) evaluated at $f_h = \hat{f}_h$. Notice that, given the audit sample, the best prediction is no longer given by (2). Denote by s the audit sample, and denote by s_h the subsample in stratum h. The best prediction of T conditional on both s and A* is given by

$$\sum_h \sum_{q \in s_h} t_{hq} + \sum_h \sum_{q \notin s_h} \mathrm{E}\left(t_{hq}|A^*_h\right)$$

An estimate is obtained on replacing $E\left(t_{hq}|\mathrm{A}^*_h\right)$ by $\hat{\mu}_{hq}$ for $q \notin s_h$. However, the difference from $\hat{E}(T|A*)$ above is small provided the audit sample is of a negligible size compared to the population. For ease of exposition, we concentrate on $\hat{E}(T|A*)$ under this assumption. When it comes to the prediction uncertainty, a naive estimated prediction variance is given by

$$\hat{V}(T|A*) = \sum_h \hat{V}\left(T_h|A^*_h\right) = \sum_h \left(\sum_q \hat{\tau}_{hq}\right) \quad \text{where} \quad \hat{\tau}_{hq} = V\left(t_{hq}|A^*_{hq}; f_h = \hat{f}_h\right)$$

i.e. the prediction variance (3) evaluated at $f_h = \hat{f}_h$. But this is usually an underestimation of the true prediction uncertainty because it ignores the uncertainty in the estimation of f_h. An estimate of the prediction variance that takes this into account is given by

$$\tilde{V}(T|A*) = \sum_h \tilde{V}\left(T_h|A^*_h\right) = \sum_h (\lambda_{1h} + \lambda_{2h}) \tag{4}$$

$$\lambda_{1h} = E_{\hat{f}_h}\left(V_{A_{hq}}\left(T_h|A^*_h; f_h = \hat{f}_h\right)\right) = E_{\hat{f}_h}\left(\hat{V}\left(T_h|A^*_h\right)\right) \tag{5}$$

$$\lambda_{2h} = V_{\hat{f}_h}\left(E_{A_{hq}}\left(T_h|A^*_h; f_h = \hat{f}_h\right)\right) = V_{\hat{f}_h}\left(\hat{E}\left(T_h|A^*_h\right)\right) \tag{6}$$

where $E_{A_{hq}}$ and $V_{A_{hq}}$ are expectation and variance with respect to A_{hq} that are evaluated at $f_h = \hat{f}_h$, and $E_{\hat{f}_h}$ and $V_{\hat{f}_h}$, are with respect to the distribution of the estimated $\hat{f}_h$.

For the audit sample, it may be the case that regular surveys that collect the target information can be linked to the statistical register, such as the Norwegian Labor Force Survey (LFS) that collects household data once a year. Otherwise, the audit sample requires its own data collection. There is then the issue regarding the design of the audit sample. Disproportionate allocation of the stratum sample sizes should be considered, in order to handle the varying within-stratum variations efficiently, taking into account both the prediction variance (3) and the estimation variance (6). Next, given the audit sample, there may be an issue of potential measurement errors in the observed allocation matrix. For instance, from my own experience, survey households collected in the LFS are often subjected to unit errors just like the register households. Several remedies can be considered. Firstly, joint modeling of the latent true allocation matrix A_{hq} and the observed allocation matrices, say A^*_{hq} from the register and A'_{hq} from the survey, can be explored. Secondly, experts can review the collected survey households A'_{hq}, against the background of the register households A^*_{hq} and other relevant information available, in order to arrive at the revised households. Such expert-revised households often have a higher quality than the directly collected survey households, such that they can plausibly be treated as the true households. Thirdly, it is still possible to verify the most tricky cases by extra field work, which however will raise the issue of cost. In short, the design and measurement of the audit sample is an important question that requires careful consideration. The solution will depend on the quality of the register and survey data available, as well as the additional relevant information in the statistical system, such that it is likely to differ from one country to another, as well as from one subject to another.

3.3. Bootstrap Under a Simple Stratified Multinomial Model

We assume a simple stratified multinomial model for the stratum distribution $f_h\left(A_{hq}, A^*_{hq}\right)$. More explicitly, suppose that there are K_h possible allocation matrices for the hth stratum of blocks, denoted by $A_{h,k}$ for $k = 1,2,\ldots,K_h$. For $1 \leq k, j \leq K_h$, let

$$\theta_{h,kj} = P\left[\left(A_{hq}, A^*_{hq}\right) = (A_{h,k}, A_{h,j})\right] \quad \text{where} \quad \sum_{k,j=1}^{K_h} \theta_{h,kj} = 1 \tag{7}$$

i.e. the probabilities of a multinomial distribution of the pair of allocation matrices.

Under our model-based prediction approach, had we observed $\left(A_{hq}, A^*_{hq}\right)$ throughout the population, we would have used the maximum likelihood estimate (MLE) of $\theta_{h,kj}$ given by

$$\tilde{\theta}_{h,kj} = \sum_{(hq)\in U_h} I_{hq;kj}/N_h$$

where $I_{hq;kj} = 1$ if $\left(A_{hq}, A^*_{hq}\right) = (A_{h,k}, A_{h,j})$ and $I_{hq;kj} = 0$ if $\left(A_{hq}, A^*_{hq}\right) \neq (A_{h,k}, A_{h,j})$, and U_h is the population stratum h and N_h the number of blocks in U_h. In practice, however, we observe A_{hq} only in the audit sample s, and we may need to take its design into consideration. The key question is whether the distribution $f_h\left(A_{hq}, A^*_{hq}\right)$ in the subpopulation U_h holds also in the corresponding subsample s_h. Provided this is the case, we may derive the MLE from the sample empirical distribution of $\left(A_{hq}, A^*_{hq}\right)$ directly. Otherwise, provided each block (hq) has a known inclusion probability in the audit sample, denoted by π_{hq}, we may use an estimate given by

$$\hat{\theta}_{h,kj} = \sum_{(hq)\in s_h} w_{hq} I_{hq;kj} \Big/ \sum_{(hq)\in s_h} w_{hq} \tag{8}$$

where $w_{hq} = 1/\pi_{hq}$. This is known as the pseudo MLE (Skinner 1989), since the estimator (8) is targeted at the population based MLE $\tilde{\theta}_{h,kj}$. Notice that, as long as the design is noninformative in the sense that $f_h\left(A_{hq}, A^*_{hq}\right)$ is the same in U_h and s_h, one could set $w_{hq} \equiv 1$ and ignore the possible varying inclusion probabilities. This is certainly the case provided there is constant within-stratum inclusion probability, i.e. $\pi_{hq} = \pi_h$. In any case, it follows from $\hat{\theta}_{h,kj}$ by (8) that an estimate of the conditional probability of A_{hq} given A^*_{hq} can be obtained as

$$\hat{f}_h\left(A_{hq} = A_{h,k} | A^*_{hq} = A_{h,j}\right) = \hat{\theta}_{h,kj} \Big/ \sum_g \hat{\theta}_{h,gj} \tag{9}$$

To estimate the variances (5) and (6), we use a stratified bootstrap procedure. The following description concerns the hth stratum, and the procedure is repeated separately in all the strata. Let $A_{s_h} = \{A_{h1}, \ldots, A_{hn_h}\}$ be the observed allocation matrices, and let $A^*_{s_h} = \left\{A^*_{h1}, \ldots, A^*_{hn_h}\right\}$ be the associated allocation matrices in the statistical register, where n_h is the number of blocks in s_h. Repeat for $b = 1, \ldots, B$:

- Draw a bootstrap sample $\left(w_{h(i)}, A_{h(i)}, A^*_{h(i)}\right)$, for $i = 1, \ldots, n_h$, randomly and with replacement from the observed $\left\{\left(w_{hq}, A_{hq}, A^*_{hq}\right); q = 1, \ldots, n_h\right\}$.
- Estimate f_h from $\left\{\left(w_{h(i)}, A_{h(i)}, A^*_{h(i)}\right); i = 1, \ldots, n_h\right\}$ by (8) and (9), denoted by $\hat{f}_h^{(b)}$
- Evaluate $\hat{\mu}_{hq}$ and $\hat{\tau}_{hq}$ at $f_h = \hat{f}_h^{(b)}$ to obtain the corresponding $\hat{E}_{(b)}\left(T_h | A^*_h\right)$ and $\hat{V}_{(b)}\left(T_h | A^*_h\right)$ by substitution for μ_{hq} in (2) and τ_{hq} in (3), respectively.

Given all the B sets of independent bootstrap replicates, we obtain

$$\hat{\lambda}_{1h} = B^{-1} \sum_{b=1}^{B} \hat{V}_{(b)}\left(T_h | A^*_h\right) \tag{10}$$

$$\hat{\lambda}_{2h} = (B-1)^{-1} \sum_{b=1}^{B} \left\{ \hat{E}_{(b)}\left(T_h | A^*_h\right) - B^{-1} \sum_{b=1}^{B} \hat{E}_{(b)}\left(T_h | A^*_h\right) \right\}^2 \tag{11}$$

Provided $w_h = 1$, this is the standard procedure under a model-based approach (Efron and Tibshirani 1993). On the other hand, in cases where design based adjustment is necessary as in (8), the bootstrap is approximately consistent with regard to the design provided (i) the audit sampling fraction n_h/N_h is negligible in each stratum h, and (ii) n_h is not too small, say $n_h \geq 25$, under some single-stage sampling design. Otherwise, adjustment to the bootstrap may be necessary. Indeed, alternative resampling methods such as jackknife may also be considered. See Shao (1996) for an overview of resampling methods for sample surveys, including justifications for the approximate consistency of the bootstrap procedure under conditions (i) and (ii).

4. An Illustration Using Norwegian Register Household Data

4.1. Data

The HR is a statistical register created at Statistics Norway on the basis of a number of sources including the last census in 2001, the CPR and the DR. For an illustration of the unit-error theory outlined above, however, I have created the following data. The census 2001 household file provides the target units. A proxy HR is created for the Municipality of Kongsvinger by adapting the procedures for the HR to only two data sources, namely the CPR, which provides the family ID at the census time point, and the SN-GAB which provides the addresses. There are no DINs at multiple-dwelling addresses for the last census time point, because the registration of the DINs in the DR was not sufficient

Table 2. Household data for Kongsvinger, Nov. 2001: By Census, CPR and proxy HR

	Source: CPR						
	Household size						
Household type	1	2	3	4	5	6+	Total
Single	4,143	0	0	0	0	0	4,143
Couple without children	0	1,505	0	0	0	0	1,505
Couple with children	0	0	766	965	279	51	2,061
Single adult with children	0	557	250	63	13	1	884
Others	0	1	0	0	0	0	4
Total	4,143	2,066	1,016	1,028	292	52	8,597
	Source: Census 2001						
	Household size						
Household type	1	2	3	4	5	6+	Total
Single	3,051	0	0	0	0	0	3,051
Couple without children	0	1,845	0	0	0	0	1,845
Couple with children	0	0	826	966	283	61	2,166
Single adult with children	0	433	197	58	10	1	699
Others	0	41	37	26	17	15	136
Total	3,051	2,319	1,060	1,080	310	77	7,897
	Source: Proxy household register						
	Household size						
Household type	1	2	3	4	5	6+	Total
Single	3,050	0	0	0	0	0	3,050
Couple without children	0	1,791	0	0	0	0	1,791
Couple with children	0	0	811	977	281	55	2,124
Single adult with children	0	418	190	52	10	1	671
Others	0	60	60	44	42	23	229
Total	3,050	2,269	1,061	1,073	333	79	7,865

until a long time after the census, which is part of the reason that the HR was first established in 2005.

The household data for Kongsvinger are shown in Table 2. We notice the following. First, the CPR has a serious deficiency when it comes to cohabitants without children. Such a couple appears as two single-person households, which is why there are many more single-person households according to the CPR than in the census, i.e. 4,143 compared to 3,051 in Table 2. The other obvious effect of this is the low number of 2-person households according to the CPR, i.e. 2,066 compared to 2,319 in the census. The net result is that there are many more households in total according to the CPR, i.e. 8,597 compared to 7,897 in the census. Next, the procedures underlying the creation of the household register seem to be able to capitalize on the relevant information in the statistical system. The two-way proxy HR table is much closer to the census table. With the dwelling register as an extra data source, the actual HR can be expected to provide even better household data. Yet, while Table 2 gives helpful indications as to the quality of the household register, it is not a direct measure of the statistical accuracy.

4.2. Model

We set the base unit to be the CPR family. The blocks are set to be each individual address. For the Municipality of Kongsvinger this gives rise to 8,597 base units, distributed over 5,638 blocks.

We assume the stratified multinomial model (7). As mentioned earlier, the strata are formed on the basis of an analysis of the relationship between the census households and the CPR families. The number of base units at a given address and whether or not there exist couples according to the CPR are used to define the strata. Table 3 provides an overview of the stratum classification and distribution, where the strata are listed in three groups. The strata are completely listed for Group (I) and Group (II), whereas Group (III) is in fact further divided into a number of strata according the *block size*, i.e. the number of CPR families at a given address — to save space these are not listed here individually. The stratum of blocks with only one base unit contains just below 80% of all the blocks, and about 50% of all the base units. Unit errors are confined to the rest of the blocks and base units. The next big group comprises blocks of two base units, further divided into three strata. Together they make up about 15% of the blocks and 20% of the base units. The last group of around 7% blocks is further divided into strata of blocks with 3, 4,. . . base units, such that the stratum sample sizes for the estimation of the corresponding stratum-specific multinomial distributions are rather small. This illustrates the point

Table 3. An overview of stratum classification

Group	Block size	Further classification	Blocks	Base units
(I)	1	–	4,351	4,351
(II)	2	Without any CPR-couple	526	1,052
		With CPR-couple and 1 register household	117	234
		With CPR-couple and 2 register households	235	470
(III)	3+	Without any CPR-couple	155	814
		With CPR-couple	254	1,676

mentioned earlier that the self-weighting audit sample can hardly be efficient in the present context. Disproportional allocation of the stratum audit sample sizes must be considered in practice. No attempt has been made to improve the audit sample design in this study, due to limited resources available for data preparation.

4.3. Results

We now apply the outlined inference approach to the Kongsvinger data. The questions that we are trying to answer are: (i) what is the expected household population given that the associated household register looks like the one for Kongsvinger, based on the relationships between the two sources that are observed in the audit sample, and (ii) what is the associated uncertainty?

The results are given in Table 4, for household totals by size and type. The first row (in each part) gives the counts according to the proxy HR. The second row gives the same counts according to the 2001 census. The third row gives the corresponding estimated prediction expectations, given by (2). Notice that the set of actual census counts can be regarded as one particular realization among all possible household populations associated with the given household register. The calculation is carried out under the stratified multinomial model that is fitted on the basis of the data from Kongsvinger. The fourth row gives the naive root squared errors of prediction (RSEP) given by (3), evaluated at the estimated stratum-specific distributions of the allocation matrices as if these were known.

Table 4. Household counts by size and type for the municipality of Kongsvinger

	Household size					
	1	2	3	4	5	6+
Proxy household register	3,050	2,269	1,061	1,073	333	79
Census	3,051	2,319	1,060	1,080	310	77
Prediction expectation	3,100	2,314	1,053	1,063	317	81
RSEP (I) without estimation uncertainty	30	17	10	8	6	5
RSEP (II) including estimation uncertainty	38	20	10	8	6	5
	Household type					
	I	II	III	IV	V	
Proxy household register	3,050	1,791	2,124	671	229	
Census	3,051	1,845	2,166	699	136	
Prediction expectation	3,100	1,797	2,134	713	183	
RSEP (I) without estimation uncertainty	25	11	9	8	11	
RSEP (II) including estimation uncertainty	37	14	12	10	14	

Denotation of household type: (I) Single; (II) Couple without children; (III) Couple with children; (IV) Single adult with children; (V) Others.

Finally, the last row gives the RSEPs given by (4)–(6) using the bootstrap procedure, which take into account the estimation uncertainty.

Comparisons between the corresponding household counts by type show that there are not as many couples (i.e. Type II and Type III) in the proxy HR as in the census, whereas, in terms of households by size, there are too many large households (i.e. 5+ persons) and too few two-person households in the proxy HR compared to the census. A typical large proxy register household involves the older generation. For example, a type-V household of five persons may contain two CPR families: (a) a single parent with two children, and (b) two grandparents. In reality (i.e. in the census) the two CPR families may constitute two households, i.e. (a) being a type-IV household of size 3 and (b) a type-II household of size 2. But one is unable to 'separate' them in the proxy HR due to the known kinship. Notice that the problem is less severe in the actual HR due to the additional information on dwelling units, which could have placed the two families in two separate blocks to start with. As a typical example that contributes to the under-count of couples in proxy HR, consider the following three CPR families at a given address: (A) single female, (B) single male, (C) single father with children. In reality (A, B) may be a type-II household of size 2, and (C) is a type-IV household. Suppose additional information shows that (A) and (C) moved to the address on the same date. The proxy HR is unable to capitalize on this information because a simple check on the same date of moving will turn out false as long as (C) is also included in the judgement, so that the three CPR families constitute three households there. Again, had the dwelling identities been available, one might have been able to place (A, B) and (C) in separate blocks to start with.

When it comes to the prediction expectations estimated under the stratified multinomial model, it can be seen that many type-V proxy register households are 'broken up' into smaller households, but not enough couples are established among them at the same time. The net result is that too many proxy households of type V are turned into single-person (type I) and single-parent (type IV) households. To understand how this may happen, consider again the three CPR families (A), (B) and (C) above. Suppose the conditional expected household composition at this address is an average over the following five possibilities: (i) three separate households (A), (B) and (C), (ii) two households (A, B) and (C), where household (A, B) is a type-II household, (iii) two households (A, C) and (B), where (A, C) form a type-III household, (iv) two households (A) and (B, C), where (B, C) will be classified as type V, and (v) one type-V household (A, B, C). Under the stratified multinomial model, these possibilities are weighted according to the overall frequencies of the corresponding allocation matrices in the audit sample. On the other hand, conditional on e.g., the additional information on the dates of moving, the chance that case (ii) may be the census classification will become much greater than its unconditional frequency in the audit sample. Or, conditional on the additional information of sex-age combinations among (A), (B) and (C), the chance of case (iv) may be greatly reduced compared to that of (iii). The stratified multinomial model is always somewhat misspecified unless such relevant information is incorporated into the model, and part of the differences between the census counts and the corresponding expectations in Table 4 can be attributed to such misspecifications. Again, the effects will be more limited with smaller blocks at the level of dwelling units. Nevertheless, deeper stratification where useful additional information is incorporated is a task for model development prior to the next census.

Comparisons between the two sets of RSEPs in Table 4 show that the naive RSEP (I) clearly underestimates the uncertainty. There are two observations to be made. First, the underestimation of RSEP (I) is not noteworthy for the counts of larger households, i.e. households of 3+ persons. The reason is that different household compositions may be compensating for each other when it comes to household counts by size, in such a way that the totals of households with 3+ persons are fairly robust towards alternative estimated distributions of the allocation matrices, and λ_{2h} is small compared to the corresponding λ_{1h} for these totals. Next, a single-person household is just the same as a household of size one, but the two counts and the associated expectations and RSEPs are being estimated as part of two different target functions, i.e. one for household counts by size and another for household counts by type. The prediction expectations are identical (i.e. 3,100) in both cases as expected. Also, the RSEPs (II) are very close to each other, i.e. 38 for household by size and 37 for household by type, and the difference is due to the Monte Carlo errors associated with the bootstrap. Meanwhile, the naive RSEPs (I) are seen to be quite different, i.e. 30 for household by size and 25 for household by type, because the decomposition of the prediction variance (4)–(6) does depend on the functionals by which a target statistic is calculated.

5. Summary and Discussion

In the above we have outlined a unit-error theory that provides a framework for evaluating the statistical accuracy of register-based household statistics, and illustrated its use through an application to the Norwegian register household data. This is certainly relevant to the coming register-based census which will be the case in a number of countries, including all the Nordic ones. It is also one step in the broad effort to give register statistics a sound statistical methodological foundation.

Several interrelated topics deserve further investigation. First of all there is the design of the audit sample. If feasible, disproportionate allocation of stratum sample sizes should be considered, in order to handle the wide range of within-stratum variations efficiently. The identification of the target units in the audit sample may require a different approach than that of the traditional sample survey. Expert review may prove to be a more costefficient alternative in many situations. In-field data collection may be necessary only for the most difficult cases.

A related matter is statistics on detailed levels. It is convenient to assume that the relationship between reality and statistical register is the same everywhere in the population. But this can potentially be misleading. One may need to develop more sophisticated models that are able to account for the between-area or -domain variations in the distribution of the allocation matrices. Alternative design of the audit sample may be explored in this regard.

No matter how good a statistical register may be, there is always a possibility that some statistics may not be as accurate as the others, as the results in Table 4 have illustrated. A statistical inferential framework can help to make the requisite assessments, and the analysis can provide valuable information for the producer of statistics. Whether or not to actually adjust the register statistics as a consequence of this evaluation will be a question that requires careful considerations, where the statistical accuracy needs to be set against the other quality dimensions, and the potential misspecifications of the underlying statistical assumptions need to be taken into account.

6. References

CENEX-Project (2006–2008). http://cenex-isad.istat.it. Project “European Centres of Networks and Excellence (CENEX)”, Area “Integration of survey and administrative data”.

Chambers, R.L. (2008). Regression Analysis of Probability-Linked Data. Technical report, Statistics New Zealand, Official Statistical Research Series, Vol. 4, 2008.

Efron, B. and Tibshirani, R. (1993). An Introduction to the Bootstrap. London: Chapman and Hall.

Harmsen, C. and Isarels, A. (2003). Register-Based Household Statistics. In European Population Conference 2003, Warsaw, Poland.

Holt, D. (2007). The Official Statistics Olympic Challenge: Wider, Deeper, Quicker, Better, Cheaper. (With Discussions). The American Statistician, 61, 1–15.

Shao, J. (1996). Resampling Methods in Sample Surveys (With Discussion). Statistics, 27, 203–254.

Skinner, C.J. (1989). Domain Means, Regression and Multivariate Analysis. In Analysis of Complex Surveys, C. Skinner, D. Holt, and T. Smith (eds). New York: John Wiley & Sons, 59–87.

UNECE (2007). Register-Based Statistics in the Nordic Countries: Review of Best Practices with Focus on Population and Social Statistics. United Nations Publication, ISBN 978-92-1-116963-8.

Wallgren, A. and Wallgren, B. (2007). Register-based Statistics – Administrative Data for Statistical Purposes. Chichester: John Wiley & Sons, Ltd.

Van der Laan, J., Harmsen, C., and Kvijvenhoven, L. (2009). Deriving Longitudinal Consistent Household Statistics from Register Information. In The 57th Session of the International Statistical Institute, Durban, South Africa.

Received November 2009
Revised July 2010