Discussion Papers No. 598, December 2009
Statistics Norway, Statistical Methods and Standards

Li-Chun Zhang

A unit-error theory for register-based household statistics

Abstract:
The next round of census 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 the 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 paper we outline a unit-error theory which provides a framework for evaluating the statistical accuracy of these register-based household statistics, and illustrate its use through an application to the Norwegian register household data.

Keywords: Register statistics, statistical accuracy, unit errors, prediction inference

Address: Li-Chun Zhang, Statistics Norway, Statistical Methods and Standards. E-mail: lcz@ssb.no

Discussion Papers	comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no
http://ideas.repec.org/s/ssb/dispap.html

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
NO-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

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 census, systems of inter-linked statistical registers have been developed to produce a wide range of purely register-based statistics (e.g. Wallgren and Wallgren, 2006). 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 as well as potentials 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 statistics. Administrative registers certainly do not provide perfect statistical data. Sampling errors are naturally absent. But there exist a variety of non-sampling 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 the least missing data. We believe that a key issue here, from a statistical methodological point of view, is the *conceptualization* and *measurement* of the *statistical accuracy* in register statistics, which will enable us to apply rigorous statistical concepts such as bias, variance, efficiency and consistency, e.g. as one is able to do when it comes to survey sampling.

In this paper we outline a statistical theory for *unit errors* in register-based household statistics. Unit errors as such are rarely mentioned in survey sampling and census. The main reason may be that while the statistical offices collect their own data in surveys and censuses explicitly for making 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. The central population register (CPR) may contain high-quality information when it comes to judicial and biological relationships within the population. But there will be no record of household relationship of other kinds. One possibility is to link the CPR to the dwelling register (DR). But more or less extensive editing and imputation procedures will be necessary in order to establish a dwelling household register, depending on the quality of the two registers as well as the linkage between them. Errors will occur in the 'constructed' register (dwelling) households whenever people actually do not live together are wrongly grouped into the same register household. We call such errors the unit errors.

As a motivating example, consider the household data in Table 1. Suppose that the dwelling ID is available in the DR, the family ID and person ID are available in the CPR. The statistical unit of interest is household. Suppose that a register household ID has been created, which is marked by * in the table to show that it may be erroneous. In this case we assume that the errors in the register households are due to the poor quality of the DR and its link to the CPR, where the dwelling ID is duplicated for person no. 1 - 5 and missing for person no. 6 - 7.

A few observations are worth noting. (i) The household register (HR) has unit errors for Knut, Lena and Ole: in reality Knut and Lena belong to one household and Ole another, whereas

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

according to the HR Lena and Ole belong to the same household and Knut another. (ii) The unit error might have occurred for all the 7 persons here. 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 statistics, 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) The register household is a unit of central interest in the coming register-based census, a statistical theory that accounts for the uncertainty due to the unit errors in register households is desirable in this respect.

The rest of the paper is organized as follows. In Section 2 we 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 we illustrate the proposed unit-error theory using the Norwegian register household data. Finally, a summary and some discussions will be given in Section 5.

2 A mathematical representation

2.1 Allocation matrix

We assume that the *target (statistical) unit* consists of one or several *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 motivating example above, the target unit is household. The base unit can be person. But it can also be family identified by the family ID, depending on the household definition. 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 units $i = 1, ..., m$. 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. 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.

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

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

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 from the ordering among them. For uniqueness of the allocation matrix, it is necessary to impose a row ordering. Let j_i denote the row number of the ith base unit. For example, in the matrix A above, we have $j_1 = 1$, $j_2 = j_3 = j_4 = 2$, $j_5 = j_6 = 3$ and $j_7 = 4$. We shall assume that, given the ordering of the base units $i = 1, ..., m$, the rows of an allocation matrix are ordered such that $j_{i'} \leq j_i$ provided $i' < i$. Put in another way, this amounts to require that an allocation matrix should be *sequential upper triangular*, where an upper triangular matrix is said to be sequential in addition provided it shall remain upper triangular after deletion of any number of the *first* rows and columns that correspond to the base units therein, as long as it is not all zero afterwards.

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 interest variables can simply be expressed as a linear transformation of X through the allocation matrix. But it can also be a non-linear function of such simple linear transformations. Some examples may help to clarify.

- *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 $\mathbf{1}_{m\times 1}$ yields the target unit sizes, defined as the number of base units

that constitute the target unit. Thus, for A given above, we obtain

$$A\mathbf{1} = (1, 3, 2, 1, 0, 0, 0)^T$$

- *Example 3:* Value matrix $X = \text{Diag}(y)_{m\times m}$ 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

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

The value vector $\mathbf{y} = (y_1, ..., y_7)^T$ 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 non-linear function

$$(A\mathbf{y})//(A\mathbf{1}) = (y_1,\ (y_2 + y_3 + y_4)/3,\ (y_5 + y_6)/2,\ y_7,\ -,\ -,\ -)^T$$

where "//" denotes component-wise division provided non-zero denominator.

- *Example 5:* Value vector of sequels, denoted by $\alpha = (1, 2, ..., m)^T$, 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

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

2.3 Blocking and strata of blocks

In our motivating example above, the street address is used to divide the target population into smaller groups called *blocks*, provided allocation is delimited within each block. That is, 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.

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:

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

which are quite different from the 5 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 household case, information such as whether there exists family nucleus or relatives at a given address (i.e. within a block), age and sex of the residents, the number of children, *etc.*, can all have a strong impact on the frequency of the possible allocation matrices. Thus, for instance, given a block of two persons at a given address, 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.

Finally, the blocks are *absolute* if allocation of base units across the blocks is strictly forbidden. Absolute blocks are suitable provided in reality base units across the blocks never (or virtually never) belong to the same target unit. It is in theory also possible to introduce *soft* blocks that have a high probability of being self-contained, but without being strictly delimiting in every situation. We refer to this as *deep blocking*. For instance, in Table 1, one may consider deep blocking that gives rise to three soft blocks: (i) family no. 1 with Astrid, (ii) family no. 2 with Geir, Jenny and Markus, and (iii) the remaining three single-person families, i.e. Knut, Lena and Ole. Now, in practice the resulting simplifications may tempt one to take greater risks in deep blocking. But valid inference must then also take into account the probability of errors associated with deep blocking. We consider only absolute blocks in this paper.

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 $h = 1, ..., H$. Denote by (hq) the qth block within the hth stratum, where $q = 1, ..., M_h$ and M_h is the number of blocks in the stratum. Denote by A_{hq} the 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

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

Denote by T the corresponding population total, given by

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

Let A^*_{hq} be the allocation matrix that corresponds to the (hq)-th block in the statistical register, which is observed throughout the population. We assume that, within the hth stratum, (A_{hq}, A^*_{hq}) are jointly independently and identically distributed across the blocks, for $q = 1, ..., M_h$. Conditional on the actual statistical register, i.e. $A^*_h = \{A^*_{hq}; q = 1, ..., M_h\}$ and $A^* = \cup_h A^*_h$, the best prediction of the target total T is given by its corresponding conditional expectation

$$E(T|A^*) = \sum_h E(T_h|A^*_h) = \sum_h (\sum_q \mu_{hq}) \qquad \text{where} \quad \mu_{hq} = E(t_{hq}|A^*_{hq}) \tag{1}$$

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

$$V(T|A^*) = \sum_h V(T_h|A^*_h) = \sum_h (\sum_q \tau_{hq}) \qquad \text{where} \quad \tau_{hq} = V(t_{hq}|A^*_{hq}) \tag{2}$$

3.2 Estimation of prediction expectation and variance

In practice, of course, we need to estimate the distribution $f_h(A_{hq}|A^*_{hq})$. 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 $\hat{f}_h(A_{hq}|A^*_{hq})$. An estimate of the prediction expectation $E(T|A^*)$ is then given by

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

i.e. the expectation (1) evaluated at $f_h = \hat{f}_h$. Notice that the audit sample is assumed be of a negligible size compared to the target population. Should this not be the case, the expression (1) should be evaluated conditional on the observed t_{hq}'s, and the prediction expectation is only calculated for the units outside of the audit sample.

For the audit sample, it may be the case that regular surveys that collect household information, such as the Labor Force Survey (LFS), can be linked to the statistical register. This is certainly the situation in the Nordic countries. The households of the survey respondents may then be considered to provide the true allocation matrix. However, from our own experiences, survey households are often subjected to unit errors just like the register households. Several remedies can be considered. Firstly, joint modelling 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}, on 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 of the audit sample is an important question that requires careful considerations. The solution will depend on the quality of the register and survey households available, as well as the additional relevant information in the statistical system, such that it is likely to differ from one country to another.

When it comes to the prediction uncertainty, under the assumption of negligible audit sample fraction, a naive estimated prediction variance is given by

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

i.e. the prediction variance (2) evaluated at $f_h = \hat{f}_h$. But this is usually an under-estimation 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

$$\tilde{V}(T|A^*) = \sum_h \tilde{V}(T_h|A_h^*) = \sum_h (\lambda_{1h} + \lambda_{2h}) \tag{3}$$

$$\lambda_{1h} = E_{\hat{f}_h}(V_{A_{hq}}(T_h|A_h^*; f_h = \hat{f}_h)) = E_{\hat{f}_h}(\widehat{V}(T_h|A_h^*)) \tag{4}$$

$$\lambda_{2h} = V_{\hat{f}_h}(E_{A_{hq}}(T_h|A_h^*; f_h = \hat{f}_h)) = V_{\hat{f}_h}(\widehat{E}(T_h|A_h^*)) \tag{5}$$

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 $\hat{f}_h$.

3.3 Bootstrap under a simple stratified multinomial model

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

$$\theta_{h,kj} = P[(A_{hq}, A_{hq}^*) = (A_{h,k}, A_{h,j})] \qquad \text{where} \quad \sum_{k,j=1}^{K_h} \theta_{h,kj} = 1 \tag{6}$$

i.e. the probabilities of a multinomial distribution of the pair of allocation matrices. The corresponding estimator, denoted by $\hat{\theta}_{h,kj}$, will depend on the design of the audit sample, denoted by s. In cases where each block in the audit sample has a known inclusion probability, denoted by π_{hq}, a weighted estimate of $\theta_{h,kj}$ can be given as

$$\hat{\theta}_{h,kj} = \sum_{(hq)\in s_h} w_{hq} I_{hq;kj} / \sum_{(hq)\in s_h} w_{hq} \tag{7}$$

where s_h is the sub-sample containing all the blocks that belong to the hth stratum, and $w_{hq} = 1/\pi_{hq}$, and $I_{hq;kj} = 1$ if $(A_{hq}, A_{hq}^*) = (A_{h,k}, A_{h,j})$ and $I_{hq;kj} = 0$ if $(A_{hq}, A_{hq}^*) \neq (A_{h,k}, A_{h,j})$. Notice that $\{\hat{\theta}_{h,kj}; k, j = 1, ..., K_h\}$ is simply the sample empirical mass function of (A_{hq}, A_{hq}^*) provided equal inclusion probability $\pi_{hq} = \pi_h$. In any case, an estimate of the conditional probability of

A_{hq} given A^*_{hq} can then be obtained as

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

Provided the audit sampling fraction is negligible, we may use a stratified bootstrap procedure. The following description is given for the hth stratum, and the procedure is repeated separately in all the strata. Let $A_{s_h} = \{A_{h1}, ..., A_{hn_h}\}$ be the observed allocation matrices, and let $A^*_{s_h} = \{A^*_{h1}, ..., A^*_{hn_h}\}$ be the associated allocation matrices that correspond to the statistical register, where n_h is the number of blocks within s_h. Repeat for $b = 1, ..., B$:

- Draw $(w_{h(i)}, A_{h(i)}, A^*_{h(i)})$, for $i = 1, ..., n_h$, randomly and with replacement from the observed $\{(w_{hq}, A_{hq}, A^*_{hq}); q = 1, ..., n_h\}$.
- Estimate f_h from the bootstrap sample $\{(w_{h(i)}, A_{h(i)}, A^*_{h(i)}); i = 1, ..., n_h\}$, denoted by $\hat{f}^{(b)}_h$.
- Evaluate $\hat{\mu}_{hq}$ and $\hat{\tau}_{hq}$ at $f_h = \hat{f}^{(b)}_h$ to obtain the corresponding $\widehat{E}_{(b)}(T_h|A^*_h)$ and $\widehat{V}_{(b)}(T_h|A^*_h)$.

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

$$\hat{\lambda}_{1h} = B^{-1} \sum_{b=1}^{B} \widehat{V}_{(b)}(T_h|A^*_h) \tag{8}$$

$$\hat{\lambda}_{2h} = (B-1)^{-1} \sum_{b=1}^{B} \{\widehat{E}_{(b)}(T_h|A^*_h) - B^{-1} \sum_{b=1}^{B} \widehat{E}_{(b)}(T_h|A^*_h)\}^2 \tag{9}$$

4 An application to Norwegian register household data

4.1 Data

The Norwegian Household Register (NHR) is a statistical register created at Statistics Norway. It is based on information from many sources, including the census, the CPR, and the register of Ground Parcels, Addresses, Buildings and Dwellings (SN-GAB). The target statistical unit is household. The units available are persons, CPR-families (including married couples, registered partners and cohabitants who are parents of residing children), and dwellings. A unique identifier is maintained for each type of units, which can be linked to each other at the unit level. The SN-GAB was introduced in connection with the census 2001, and is still subject to severe errors. For instance, the dwelling ID was missing for about 7% of the residents in 2005 when the NHR was first introduced. There exist also a fair amount of registration errors at the street-address level, as well as among the dwellings at a given street address. Due to these errors it is necessary to create the NHR for statistical purpose.

For an illustration of the unit-error theory outlined above, we created the following data. The census 2001 household file provide the target units. A proxy register household file is created for the Municipality Kongsvinger by adapting the procedures for the NHR to only two data sources, namely, the CPR which provides us the family ID at the census time point, and the SN-GAB at

Table 2: Household data for Kongsvinger, Nov. 2001 by census, CPR and a register.

Source: CPR							
	Household size						
Household Type	1	2	3	4	5	6+	Total
Single	4143	0	0	0	0	0	4143
Couple without Children	0	1505	0	0	0	0	1505
Couple with Children	0	0	766	965	279	51	2061
Single Adult with Children	0	557	250	63	13	1	884
Others	0	4	0	0	0	0	4
Total	4143	2066	1016	1028	292	52	8597

Source: Census 2001							
	Household size						
Household Type	1	2	3	4	5	6+	Total
Single	3051	0	0	0	0	0	3051
Couple without Children	0	1845	0	0	0	0	1845
Couple with Children	0	0	826	966	283	61	2166
Single Adult with Children	0	433	197	58	10	1	699
Others	0	41	37	26	17	15	136
Total	3051	2319	1060	1080	310	77	7897

Source: Register							
	Household size						
Household Type	1	2	3	4	5	6+	Total
Single	3050	0	0	0	0	0	3050
Couple without Children	0	1791	0	0	0	0	1791
Couple with Children	0	0	811	977	281	55	2124
Single Adult with Children	0	418	190	52	10	1	671
Others	0	60	60	44	42	23	229
Total	3050	2269	1061	1073	333	79	7865

the street address level. (As mentioned above, the dwelling identity numbers at multiple-dwelling street addresses were not available at the last census time point.)

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 appear as two single-person households, which is why there are many more 1-person households according to the CPR than in the census, i.e. 4143 compared to 3051 in Table 2. The other obvious effect of this is the low number of 2-person households according to the CPR, i.e. 2066 compared to 2319 in the census. The net result is that there are many more households in total according to the CPR, i.e. 8597 compared to 7897 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 register household table is much closer to the census table. With the dwelling register as an extra data source, the actual NHR can be expected to provide even better household data. Yet, while Table 2 gives helpful indications on the quality of the household register, it is not a direct measure of the statistical accuracy.

4.2 Model

We set the base unit (BU) to be the CPR-family. The blocks are individual street address. This is possible because the census employs a 'formal' definition of household, where people are 'placed' at their CPR-addresses. Using the CPR-family as the base unit can be a problem if household is defined for people who actually live at the same addresses. Students are a typical group where the CPR-address (i.e. usually that of the parents) often differs from the actual dwelling address. For Municipality Kongsvinger this gives rise to 8597 base units, distributed over 5638 blocks.

Table 3: An overview of stratum classification.

Group	Block Size	Further Classification	Blocks	Base Units
(I)	1	-	4351	4351
(II)	2	Without any CPR-couple	526	1052
		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	1676

We assume the stratified multinomial model (6). The strata are formed based on an analysis of the relationship between census households and CPR-families. The aim is similar to strata formation in sampling design, i.e. to minimize the within-stratum variation while maximizing the between-strata variation. That is, we search to identify groups of blocks with clearly distinct distributions of the pairwise allocation matrices. As mentioned earlier, the number of base units inside a block (i.e. block size) is a natural stratum variable. The other most important factor turns out to be whether or not there exist couples within a block according to the CPR, to be referred to as the CPR-couple. Notice a CPR-couple may be 'constructed' based on information of children in cases where the couple are neither married to each other nor registered as partners.

Table 3 provides an overview of the stratum classification and their distribution. The stratum of blocks with only 1 base unit (or CPR-family) contains just below 80% of all the blocks, and about 50% of all the base units. Unit errors are confined to the rest blocks and base units. The next big group comprises of blocks of 2 base units, further divided into 3 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 blocks of 3, 4, ... base units, such that the effective sample sizes for the estimation of the corresponding stratum-specific multinomial distributions are rather small. It is thus quite clear that self-representing audit sample can hardly be efficient for the purpose here. In practice, disproportional allocation of the stratum audit sample sizes should be considered.

4.3 Result

We now apply the outlined inference approach to the Kongsvinger data. Of course, in reality one would only calculate the prediction expectation and variance for the units outside of the audit sample. Our purpose here is illustrative. 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?

Table 4: Household counts by size for Municipality Kongsvinger.

	Household size					
	1	2	3	4	5	6+
Proxy Household Register	3050	2269	1061	1073	333	79
Census	3051	2319	1060	1080	310	77
Prediction Expectation	3100	2314	1053	1063	317	81
RSEP (I) without estimation uncertainty	30	17	10	8	6	5
RSEP (II) including estimation uncertainty	38	20	10	8	6	5

The results are given in Table 4. The first row gives the counts of households by size according to the proxy household register. The second row gives the same counts according to the census 2001. The third row gives the estimated prediction expectations, given by (1), of the corresponding household population. 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 based on the data from Kongsvinger. The fourth row gives the naive root squared errors of prediction (RSEP) given by (2), evaluated at the estimated stratum-specific distributions of the allocation matrices as if these were known. Finally, the last row gives the RSEPs given by (3) - (5) using the bootstrap procedure, which take into account the estimation uncertainty. The increase of the RESP is the largest for the number of 1-person households, and second for the number of 2-person households. The increments are not noteworthy for the other household counts. In either case, the census counts are well within the respective 95% prediction intervals based on the Normal approximation, except from the number of 4-person households which is just on the border of the interval. The proxy household register counts are for the most part well within the same prediction intervals, except from the numbers of 2- and 5-person households. These two are also the ones that actually differ most from the census counts.

5 Summary and discussions

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. The issue is certainly relevant for 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 bring sound statistical methodological foundations to register statistics.

Several inter-related topics deserve future investigations. First of all there is the design of audit sample. As noticed earlier, disproportionate allocation of stratum sample sizes should be considered, in order to handle the varying within-stratum variations efficiently. The identification of the target units in the audit sample may require a different approach than traditional sample

survey. Expert review may prove to be a more cost-efficient 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 advanced 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 the 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 such assessments, and the analysis can provide valuable information for the producer of statistics. Nevertheless, whether or not to actually adjust the register statistics as a consequence of this evaluation will be a question that requires much more consideration, where the statistical accuracy as well as its underlying assumptions need to be placed under a bigger context.

The allocation matrix is a generalization of the permutation matrix that is suitable for representation of probability-linkage errors. Instead of a single value of one in each row and column, the allocation matrix allows for multiple values of one per row as well as possible rows of all zeros, but still a single value of one in each column because a base unit must belong to one target unit and that only. Parametric and, hence, parsimonious models of the permutation matrices have been developed (Chambers, 2008), where concentration of values of one along the diagonal (i.e. correct linkage) may be expected. Parametric modelling of the allocation matrices seems a bit more complicated, but it is certainly welcome due to the potential improvement of estimation efficiency. These can possibly replace the simple multinomial model that has been used here.

References

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

Holt, D. (2007). The official statistics Olympic challenge: Wider, deeper, Quicker, Better, Cheaper. (With discussions). *The American Statistician*, **61**, 1–15.

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. (2006). *Register-based Statistics - Administrative Data for Statistical Purposes*. John Wiley & Sons, Ltd.