**Register-based estimation of total dwellings and households**

Li-Chun Zhang, Division for Methods, Statistics Norway, lcz@ssb.no

Johan Fosen, Division for Methods, Statistics Norway, jof@ssb.no

**Abstract**

*Total of dwellings and households are key statistics of a country. Traditionally these figures are produced based on population and housing census, which can be costly. How to produce these statistics based on statistical data originated from the relevant administrative sources is currently a major challenge to the Census Transformation Programmes at a number of European countries. We study the matter based on the Norwegian Address Register and Population Register. A particular difficulty arises due to the fact that dwelling identification at multi-dwelling addresses is problematic, owing to quality issues in the input sources, despite reliable identification of all the addresses. We develop two extensions to the existing capture-recapture methodology in this context. The first one can be characterised as a two-step approach, where one first applies standard capture-recapture methods to obtain an estimate of the resident addresses (i.e. address at which there exists dwellings and possibly households), and then use various missing-data methods to estimate the number of dwellings and households per address. The second approach can be formulated in terms of a log-linear model, under which it becomes possible to estimate at once the sizes of three populations, namely of resident address, dwelling and resident dwelling (i.e. dwelling household). We demonstrate both the approaches using real-life data. These provide potentially options for purely register-based estimation of total dwellings and households, instead of the costly census.*

**Keywords:** Register-based population size estimation, census transformation programme, capture-recapture of clustered elements, log-linear models.

**1. Background**

In Norway the Household Register (HR) was first introduced in 2006. Census-like yearly household statistics can be tabulated from the HR directly. The production approach of HR is micro-level unit harmonisation (Di Zio et al., 2017). A similar production approach is used for dwelling statistics that has been published annually since 2006 (Bloch, 2018). For the HR production process, first various relevant units from two sources are aligned: person, family and residence address form Population Register (PR); address, building and dwelling from the Address and Building Register (ABR). Next, the households (i.e. statistical units) are created by a rule-based deterministic procedure. Several approaches were used to assess the accuracy of the HR; see e.g. Falnes-Dalheim (2014), Zhang (2009) and Zhang (2011).

The aim of this paper is to study statistical estimation methods for dwellings and dwelling households based on multiple registers.

**2. Under-count and over-count**

In a country like Norway, where register-based census-like statistics have become a reality on many topics, the PR (Person register) is a natural starting point for register-based household statistics.

The registered residence address (RRA) is the place where a person lives. In the PR, RRA is based on the place of rest, i.e. where a person normally sleeps during night. An RRA means a *resident dwelling,* and resident dwellings should ideally be in 1-1 correspondence with dwelling households. An *address* is represented by 17 digits and may consist of one or more dwellings. At a single-dwelling address, both the dwelling and the RRA is identified by the address. However, at a multi-dwelling address, the identification in additional contains an additional 9-digit code.

*2.1 Under-count of RRA-households*

Let an RRA-household consist of all the PR-residents who share the same RRA, thus the RRA is distinct from one RRA-household to another. Consider the effects of potential RRA-registration errors. Suppose the RRA of a household is *a* in the PR but should actually be *b*.

**Table 1: Effect on household total if RRA-b household registered at a**.



* (No *a*, No *b*): Clearly, the total number of households is unaffected.
* (No *a*, Yes *b*): By stipulation any household registered at b should have been elsewhere. If there is only one household at b, the household total is unaffected; if there are 2 or more households at b, the total is under-counted.
* (Yes *a*, No *b*): Clearly, the household total is under-counted here.
* (Yes *a*, Yes *b*): There are at least three households, so we have under-count.

It is thus clear that registration errors will always lead to net under-count of dwelling households by direct enumeration of RRA-households. Empirically, the RRA-households have shown to be too few in number and too large in average size, compared to historic census counts and register-based household statistics in Denmark and Finland.

The ABR is a reliable source of buildings (or parts of building) which can be identified by address. However, for various reasons, the ABR has under-coverage of dwellings. This also leads to under-coverage of RRA-households, but the problem needs to be distinguished from under-enumeration by other causes, because it cannot be remedied from juxtaposing the various sources. Under-count of RRA-households can happen in the PR (or any relevant source) at two different levels:

C1. In one source there may be *under-enumeration of addresses* due to registration errors, if there are dwelling households at an address, but there is no RRA at this address, i.e. no one registered living at this address. This can happen whether or not there are multiple dwellings at this address.

C2.1 There may be under-enumeration of dwellings *at an address*. For example, two households may be registered at the same dwelling.

C2.2 There may be systematic under-enumeration of a dwelling at an address, when it is missing in the ABR altogether. In this case, under-enumeration is not caused by registration errors in the PR (or any source of RRA), because there is simply no RRA that can be registered as dwelling: whenever an address consists of multiple dwellings, the number of RRAs are limited by the number of dwellings in the ABR. Since dwelling identification is only used in connection to ABR, no other source will discover any of the missing dwellings in ABR.

**2.2. Potential over-count of HR**

 The HR aims at comparability with census household statistics from the past, and internationally with other register-based countries including Denmark and Finland. An important objective is to facilitate register-based household income statistics, which ideally is based on a living household definition.[[1]](#footnote-1) To these ends, persons with the same RRA are accepted as members of the same household if they are kins, have the same moving dates or otherwise judged to be ‘admissible’, as long as the resulting household does not appear highly unusual compared to the household types observed in Census 2001. However, persons sharing the same dwelling but not having any ‘plausible’ relationship with each other will be allocated to separate households in the HR, to be referred to as the HR-households. Overall the HR is expected to lead to an over-count of dwelling households.

**3. Data**

In addition to the PR and the ABR, a number of potential sources of dwelling and dwelling households have been examined, including the Post Address Register (PAR), Work and Welfare Register (NAV) and the Broadcast License Register (BLR). PAR suffered from too much dependence with PR and NAV had over-coverage, both properties incompatible with the capture-recapture estimation method used in this paper.

Three datasets for the year 2016 have been prepared for analysis, which are organised around address, person and *building utility unit (BUU)*, respectively. The preparation implied a large amount of micro integration work for pre-processing of each input register, followed by transforming, linking and aligning the multiple input files. The three datasets were:

* Address dataset: Unit is address. For each address: No. PR-residents, No. dwelling RRAs, No. dwellings in BLR; no. of total BLR licenses, No. HR-households; No. dwellings according to the ABR, binary enumeration indicators for each source.
* Person dataset. The unit is PR-resident. Variables: sex, age, RRA in the PR, sign-of-life (SOL) indicator, and other relevant socio-demo characteristics.
* BUU dataset. The unit is BUU. Variables: Planned type of use including dwelling, building characteristics, including type, areal, no. of rooms, etc, binary enumeration indicators for the sources.

**4. Estimation**

Capture-recapture methods are standard methods for population size estimation (e.g. Fienberg, 1972; Nirel & Glickman, 2009; Böhning et al., 2017). In our case the target unit of estimation may be the dwelling or the resident dwelling (i.e. dwelling household), which we refer to as the target *population elements*. Given the contrast between address and dwelling, capture-recapture of the addresses can be envisaged as capture-recapture of population elements in *clusters*. We shall refer to addresses as the *enumeration clusters*. Below we consider four estimators under a two-step estimation approach, in the absence of reliable data at the element level. For simplicity the description below is given for the population total, which can be applied separately to each population stratum in the case of stratification.

***Dual system estimation (DSE)*** is the simplest setting. Denote by  the target population of unknown size. Suppose that the population elements can be enumerated in clusters. Denote the *enumeration clusters* by , which is of unknown size . Let  and  be two enumeration lists. Denote by  and  the numbers of enumeration clusters in  and , respectively, and by  that in both  and . The DSE of the cluster population size  is given by

 

The standard assumptions are given by Wolter (1986), in the context of census and Census Coverage Survey (CCS), where *both* list are considered to be random. For DSE based on administrative lists, these assumptions may be unnecessary strict. For instance, an administrative list may have systematic under-coverage of the target population, due to the nature of the source. Zhang (2017) proposes to treat the administrative list  (i.e. ) as fixed and  (i.e.  and ) as random. Provided constant catch rate () of all the clusters of  in , we have



The other key assumptions required (Zhang, 2017) are perfect matching between  and , and absence of duplicated or erroneous enumeration both in  and .

Having obtained the cluster population size estimate , we estimate the target population size *N* at the second step, where the target element may either be the dwelling or the resident dwelling (i.e. dwelling household). In our case each cluster is a *resident address* with one or multiple dwellings. Let  be the DSE of  based on the ABR and the PR. Denote the four *list domains* of  by , which are of the sizes , as shown in Table 2 where  needs to be estimated and the other three numbers are observed directly.

Table 2: Domains of Ω, i.e. all resident addresses

|  |  |  |  |
| --- | --- | --- | --- |
|  | In ABR | Out of ABR | Total |
| In PR | *M*11 | (*N*11) | *M*10 | (*N*01) | *M*1+ (*N*1+) |
| Out of PR | *M*01 | (*N*01) | *M*00 | (*N*00) | *M*0+ (*N*0+) |
| Total | *M*+1 | (*N*+1) | *M*+0 | (*N*+0) | *M* (*N*) |

Let  be the number of dwellings at address , and  the number of resident dwellings at the same address. The total  is the population size with dwellings as the elements, whereas the total  is the population size with resident dwellings as the elements. Having estimated , one may fix the four domains of . The estimation of , which can either be  or , becomes then a problem due to *missing observations* (of either  and ) from the relevant addresses in .

**(i) Missing completely at random (MCAR) assumption.**

Denote by  if address  is in the ABR, and  otherwise. The MCAR assumption for estimation of dwellings is given by



That is, the no. dwellings at an address is independent of whether the address is enumerated in the ABR. Similarly, the MCAR assumption for estimation of resident dwellings is given by

 

where  if address  is in the PR, and  otherwise.

 **(ii) Missing at random (MAR) assumption**  The MCAR assumption can be relaxed to make use of the observed  when estimating , and to make use of the observed  when estimating . This leads to the MAR assumptions

  

It can be shown that provided one estimates  by the DSE , the MAR estimator would coincide with the MCAR estimator.

**(iii)** **Joint estimation under log-linear models**

Let  be a categorical variable derived from, e.g. taking value 1, 2, or 3 for , which represents whether address  has 1, 2, 3 or more dwellings, respectively. Similarly, let  be a categorical variable derived from. Denote by  the cross-classified contingency table. In the present setting,  is missing if  and  is missing if . Joint estimation of  in ,  and  can yield the corresponding total  at the same time. For instance, this can be achieved under the simplest log-linear model

 

which implies the following 3 MAR assumptions at the same time:

 

 

 

**(iv)** **Triple-population size estimation (TPSE)**

The log-linear model  entails three MAR assumptions (iii.1) - (iii.3). It is possible to replace (iii.3) by a more relaxed MAR assumption

 

Together (iii.1), (iii.2) and (iii.3c) correspond to the log-linear model

 

which is less restrictive than the model , where the latter is nested in the former. The model  has been studied by Van der Heijden et al. (2012; 2017) in the context of population (i.e. resident addresses here) size estimation with partially missing covariates. In our case here, the partially missing covariates are the counts of dwellings and resident dwellings at each resident address. The key difference to the model  is the estimates for the domain . Instead of (iii.1), (iii.2) and (iii.3c), the model  can most succinctly be given as

 

where  is the interaction between  and  in , for .

## Uncertainty measurement

The uncertainty of the estimators can be assessed by bootstrapping (Efron & Tibshirani, 1993). The details are omitted here due to space limit.

**5. Some results**

Figure 1 shows the results when performing DSE at cluster level with ABR as list A and PR as list B. We notice that, as expected, the variation is larger for small municipalities. The outliers indicate that for some municipalities the underlying assumptions of dse fails, e.g. the assumption of perfect linkage between clusters identified by both links.

At the second step, we see in Table 3 the results[[2]](#footnote-2) for three estimators: Estimator 1 is the estimator presented in (ii), estimator 2 in (iii) and estimator 3 in (iv) in section 4. The ABR-direct enumeration in Table 3 corresponds to the official number of dwellings published by Statistics Norway, and the conservative alternative of HR-direct enumeration is about 5% lower than the official HR-publication. We see that the estimators only increase the dwelling estimate with about one percent, whereas the estimate for dwelling households are increased by about three percent compared to PR. The three estimators are relatively equal, and the reason is that the proportion of addresses being observed in both lists, i.e. cell (1,1) in Table 2, is very large, thus the ability of the estimators to give different results is limited. Especially close are estimator 2 and 3 since they are only differing by the estimation of cell (2,2) in Table 2, a cell that by step 1 is containing very few addresses, thus giving small differences no matter how differently the number of dwellings and resident dwellings are estimated for this cell.

**Figure 1. Relative difference between dse and ABR for each municipality using ABR as list A and PR as list B. Municipalitites sorted in increasing size**.



**Table 3. The estimated number of dwellings and dwelling households at the national level.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimator | Relative w.r.t. direct enum. | Relative wr.t. ABR-direct enum. | Relative wr.t. dwelling estimates | Absolute |
| Dwellings |  |  |  |  |  |
|  | ABR-direct enumeration | 1.0000 | 1.0000 | 1.0000 | 2 248 422 |
|  | Estimator 1 | 1.0119 | 1.0119 | 1.0000 | 2 275 214 |
|  | Estimator 2 | 1.0120 | 1.0120 | 1.0000 | 2 275 433 |
|  | Estimator 3 | 1.0119 | 1.0119 | 1.0000 | 2 275 289 |
| Dwelling households |  |  |  |  |  |
|  | PR-direct enumeration | 1.0000 | 0.8892 | 0.8892 | 1 999 356 |
|  | Estimator 1 | 1.0381 | 0.9231 | 0.9122 | 2 075 520 |
|  | Estimator 2 | 1.0355 | 0.9208 | 0.9099 | 2 070 330 |
|  | Estimator 3 | 1.0354 | 0.9207 | 0.9098 | 2 070 129 |
|  | Conservative HR-direct enumeration[[3]](#footnote-3) | 1.1188 | 0.9949 |  | 2 236 916 |

**6. Conclusions and discussions**

A two-step approach has been developed for the estimation of dwellings and resident dwellings (dwelling households) using capture-recapture methodology, due to the lack of reliable recapture data at the element level of interest. There are only small differences between the alternative estimates when applied to the Norwegian data, due to the large overlaps between ABR and PR at the address level in most of the municipalities.Further analysis (not detailed here) show that the “formal” dwelling household estimates may suffer from under-coverage error caused by the under-registration of dwellings in the ABR. Moreover, the conceptual difference between dwelling and living households is another important reason for the discrepancy between these estimates and the HR-based statistics.

**7. Acknowledgements**

This works has been funded by the Eurostat grant no. 07112.2016.004-2016.594 for the action entitled - Improvement of the use of administrative sources - ESS.VIP ADMIN WP6 pilot studies and applications.

Thanks to Johan Åmberg, Leif Rusten and Grete Smerud at Statistics Norway for preparing data sets. Thanks also to colleagues at Statistics Norway for valuable discussions: Vilni Verner Holst Bloch, Anders Falnes-Dalheim, Anders Haglund, Anders Holmberg, Roger Jensen, Paul Inge Severeide, Grete Smerud, Harald Utne and Johan Åmberg.

**8. References**

Bloch, V.V.H. (2018), Registerbasert boligstatistikk – Dokumentasjon og videre arbeid. Documents 2018/02 (in Norwegian), Statistics Norway.

Böhning, D., Van der Heijden, P.G.M. and Bunge, J. (2017) Capture-Recapture Methods for Social and Medical Sciences. CRC Press.

Di Zio, M., Zhang, L.-C. and De Waal, T. (2017). Statistical methods for combination of multiple sources of administrative and survey data. The Journal of International Association of Official Statisticians, 76:17-26.

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

Falnes-Dalheim, A. (2014). Dokumentasjon av registerbasert familie- og husholdningsstatistikk. Internal Technical Report (in Norwegian), Statistics Norway.

Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. Biometrika, 59, 409-439.

Nirel, R. and Glickman, H. (2009). Sample surveys and censuses. In Sample Surveys: Design, Methods and Applications, Vol 29A (eds. D. Pfeffermann and C.R. Rao), Chapter 21, pp. 539-565.

Wolter, K. (1986). Some coverage error models for census data. Journal of the American Statistical Association, 81, 338-346.

Van der Heijden, P. G. M., Whittaker, J., Cruyff, M. J. L. F., Bakker, B., and van der Vliet, R. (2012). People born in the Middle East but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates. Annals of Applied Statistics, 6, 831-852.

Zhang, L.-C. (2017). A note on dual system population size estimator. *To appear*.

Zhang, L.-C. and Dunne, J. (2017). Trimmed Dual System Estimation. In Capturerecapture Methods for the Social and Medical Sciences, eds. D. Böhning, P.G.M. Van der Heijden and J. Bunge, pp. 239-259. CRC Chapman and Hall.

Zhang, L.-C. (2015). On modelling register coverage errors. Journal of Official Statistics, 31, 381-396.

Zhang, L.-C. (2011). A Unit-Error Theory for Register-Based Household Statistics. Journal of Journal of Official Statistics, Vol. 27, No. 3, 2011, pp. 415-432.

Zhang, L.-C. (2009) Estimates for small area compositions subjected to informative missing data. Survey Methodology, vol. 35, pp. 191-201.

1. the subset of the dwelling household sharing the household budget. [↑](#footnote-ref-1)
2. For technical programming reasons, 31 small municipalities of the 428 municipalities have not been calculated in this Section. These municipalities add up to only about 0.4% of the households in Norway.. [↑](#footnote-ref-2)
3. A slightly modified conservative alternative to the official HR-enumeration that gave 2 348 797 households in 2016. [↑](#footnote-ref-3)