Register-based household and dwelling statistics

Li-Chun Zhang[1,2]

[1] *University of Southampton (L.Zhang@soton.ac.uk)*
[2] *Statistisk sentralbyrå, Norway*
[3] *Universitetet i Oslo*

Population, Household & Dwelling Registers

A brief history in Norway:

- First, census (till 1980)

  [Central Pop. Register: 1964, used in census 1970]

  [**CPR** with family relationship, used in census 1980]

- Then, "virtual" census [admin+survey, 1990]

- The "last" census [Dwelling Register (**DR**), 2001]

- Household Register (**HR**) in 2006

- Register-based census-like statistics in 2011

On definition of dwelling household (**DwHsh**)

Household concept	Source	
	De facto	De jure
Dwelling (place-of-rest)	Sign-of-Life / Survey	CPR
Living (share-a-budget)	Survey	n/a

E1: Two owners of separate dwellings cohabit at one. One DwHsh in fact, two in form. Unrealistic to 'dictate' the registration.

E2: One of a couple weekly commutes. Two DwHsh in fact, one in form. Unreasonable to 'separate' the couple.

E3: A student registered at parents' home. Two DwHsh in fact, one in form. Maybe financially dependent. Which use?

Problems with registered address (**RA**) in CPR

Complete address (CA): each CA identifies a dwelling

- building level: address [e.g. villa, farm house, etc.]
- sub-building: address + DIN [e.g. in apartment building]

CPR-household = persons with the same RA in CPR

1. Not every RA is a CA
2. In case RA = CA, not every RA is correct

Register-based DwHsh targets CPR-household which one could have obtained in the absence of RA-errors.

What if RA-b household registered at RA-a in CPR?

Presence of *other* households registered at (a, b)			
(No a, No b)	(No a, Yes b)	(Yes a, No b)	(Yes a, Yes b)
Unaffected	Unaffected or Under-count	Under-count	Under-count

Overall, address registration errors in CPR *always* lead to net under-count of DwHsh by CPR-households.

NB. Comparison to census/sample survey data

[link at building level; for editing rules; indirect use of survey data]

An illustration at building-level address

Reality

DIN	Family	Household	Person	Name	Sex	Age	Charact.
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

DIN	Family	Household*	Person	Name	Sex	Age	Charact.
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

Illustration of simple decision rules in HR2005

Processing at RA = CA	Total (×1000)
Base Unit (BU)	1931.9
Merging kinship	-23.2
Merging moving date	-33.3
Merging cohabitation tendency	-33.7
Splitting implausible merging	+6.0
Household with registered DIN	1847.7

NB. BU based on CPR Family, RA and Census 2001

Illustration of simple decision rules in HR2005

By 1.1.2005	Household by Size (%)						Total
Source	1	2	3	4	5	6+	(×1000)
CPR Family	47.7	-	-	-	-	-	2215
CPR RA	29.8	29.4	14.8	15.0	7.3	3.7	1681
BU	42.9	25.0	12.4	12.6	5.4	1.6	2095
BU | RA = CA	41.0	25.8	12.8	13.0	5.7	1.6	1932
Processed	35.8	28.6	13.7	13.9	6.2	1.8	1847
Incl. Rest BU	38.1	27.5	13.3	13.4	6.0	1.7	2010
Census 2001	37.7	27.3	13.7	13.6	6.0	1.7	1962

Illustration of simple decision rules in HR2005

Household Size	1960	1970	1980		1990		2001		2005	
1 Person	14,2	21,1	27,9	(29,1)	34,3	(34,0)	37,7	(37,1)	38,1	(38,0)
2 Persons	23,3	25,4	25,8	(31,2)	26,3	(32,8)	27,3	(33,2)	27,5	(33,2)
3 Persons	21,2	18,8	16,3	(15,8)	15,2	(14,9)	13,7	(12,5)	13,3	(11,8)
4 Persons	20,2	17,9	17,9	(16,1)	16,0	(13,2)	13,6	(11,8)	13,4	(11,5)
5 Persons	11,6	10,2	8,3	(5,7)	6,4	(3,7)	6,0	(3,9)	6,0	(4,0)
6+ Persons	9,6	6,7	3,7	(2,1)	1,9	(1,3)	1,7	(1,5)	1,7	(1,4)
Total ('1000)	1 077	1 297	1 524	(2 062)	1 751	(2 265)	1 962	(2 444)	2 010	(2 497)

NB. Denmark in parentheses

Assessing statistical uncertainty

I. Double mixed-effects modelling (Zhang, 2009)

- SPREE (Purcell & Kish, 1980) to GSPREE (Zhang & Chambers, 2004): mixed-effects model relating association structure of CA-household (target) to that of CPR-family (auxiliary)
- differential DIN-missing rates by Municipality and household type: random effects (Municipality by household type) of missing rate

II. A unit-error theory (Zhang, 2011)

- Introducing allocation matrix $\mathbf{A}$
- Uncertainty propagation by $\hat{f}(\mathbf{A}|\mathbf{A}^*)$ given $\mathbf{A}^*$ in HR

Assessing statistical uncertainty

Reality	Astrid	Geir	Jenny	Markus	Knut	Lena	Ole
Hush_1	1	0	0	0	0	0	0
Hush_2	0	1	1	1	0	0	0
Hush_3	0	0	0	0	1	1	0
Hush_4	0	0	0	0	0	0	1

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

Register	Astrid	Geir	Jenny	Markus	Knut	Lena	Ole
Hush_1	1	0	0	0	0	0	0
Hush_2	0	1	1	1	0	0	0
Hush_3	0	0	0	0	1	0	0
Hush_4	0	0	0	0	0	1	1

Assessing statistical uncertainty

Example: To obtain household age composition for 4 age groups (0-18, 18-30, 31-65, 66+), use dummy-index value matrix as follows:

$$\mathbf{X} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad \Rightarrow \quad \mathbf{AX} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 2 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Assessing statistical uncertainty

Kongsvinger	**Household by 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 including estimation uncertainty	**38**	20	10	8	6	5

Kongsvinger	**Household by type**				
	A	B	C	D	E
Proxy Household Register	3050	1791	2124	671	229
Census	3051	1845	2166	699	136
Prediction Expectation	**3100**	1797	2134	713	183
RSEP including estimation uncertainty	**37**	14	12	10	14

(A) Single; (B) Couple without Children; (C) Couple with Children; (D) Single Adult with Children; (E) Others

Integrating Household and Dwelling Registers

CPR			DR
Person	Household	Complete Address	Dwelling
1	1	1	1
2	1	1	1
3	2	2	N/A
4	3	2	N/A
⋮	⋮	⋮	⋮
N	M	D_N	D_P
-	-	-	$D_P + 1$
⋮	⋮	⋮	⋮
-	-	-	D

NB. $M = 2.24 \times 10^6$ and $D = 2.42 \times 10^6$ in 2011

Integrating Household and Dwelling Registers

HD = Perfectly matched household-dwelling set

- HR = HD ∪ HuD [HuD = households without matched dwelling]
- DR = HD ∪ DuH [DuH = dwellings without matched household]

Options and challenges:

- weighting of HD set for statistics, or impute dwelling characteristics for HuD set: lack of coherence/numerical consistency with low-level dwelling statistics
- linkage between HuD and DuH: not directly linkable...

Nearest neighbour linkage (NNL) of units

sets of units $A = \{1, 2, ..., n_A\}$ and $B = \{1, 2, ..., n_B\}$

vector of keys: $\mathbf{x} = (x_1, x_2, ..., x_p)$ available to both A and B

dissimilarity or distance measure between $\mathbf{x}$ and $\mathbf{x}'$: $\|\mathbf{x} - \mathbf{x}'\|$

Method: for each $i \in A$,

1. nearest neighbour (NN): $k = \arg\min_{j \in B} \|\mathbf{x}_i - \mathbf{x}_j\|$
2. nearest neighbour linkage (NNL) $i \leftrightarrow k$

NB. linkage noise if ties (multiple NNs);

NNL from B to A may not result in same $k \leftrightarrow i$

NB. similar to nearest neighbour imputation (Chen & Shao, 2000)

NB. deterministic record linkage requires $\|\mathbf{x}_i - \mathbf{x}_j\| = 0$

Double nearest neighbour linkage (DNNL)

ID (Set A)	ID (Set B)	Keys (Set A)	Keys (Set B)
1		$\mathbf{x}_1$	
...		...	
M_A		$\mathbf{x}_{M_A}$	
M_A+1	M_B+1	$\mathbf{x}_{M_A+1}$	$\mathbf{z}_{M_B+1}$
...	...	...	...
M_A+N	M_B+N	$\mathbf{x}_{M_A+N}$	$\mathbf{z}_{M_B+N}$
	1		$\mathbf{z}_1$
	...		...
	M_B		$\mathbf{z}_{M_B}$

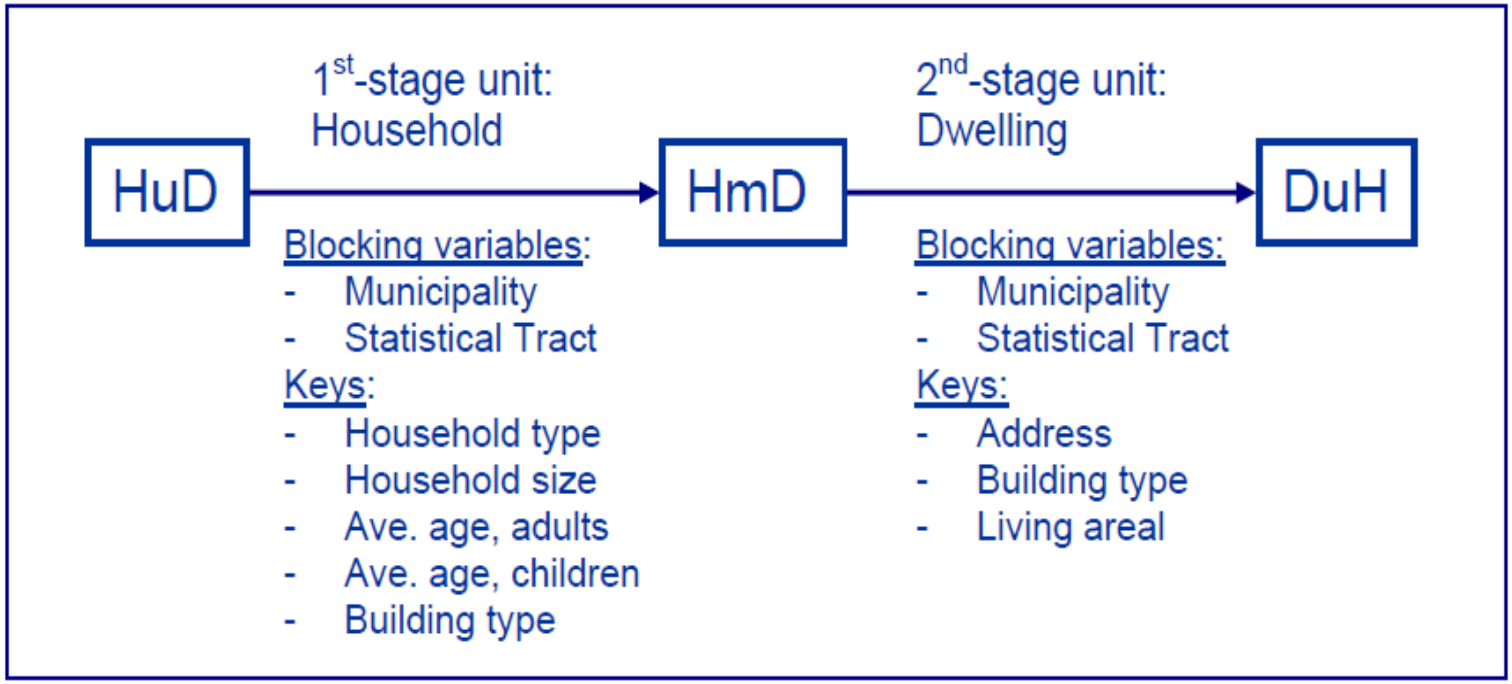
DNNL from A to B: put $R = \{1, ..., N\}$

1. for $i \in \{1, ..., M_A\}$, find NN-match $j \in R$, based on keys $\mathbf{x}_A$

2. for j above, find NN-match $k \in \{1, ..., M_B\}$, based on keys $\mathbf{z}_B$

3. DNNL $i \leftrightarrow k$, where $i \in A$ and $k \in B$ [NB. without common keys]

Implementation in 2011 (Zhang & Hendriks, 2012)



	HuD set: DNNL by 2nd-stage blocking		
HD set	(Street) Address	Census Tract	Municipality
85%	$15\% \times 47.6\%$	$15\% \times 93.5\%$	$15\% \times 97.8\%$

Targets: **Dwelling** *&* **resident dwelling**

Under-counting of resident dwelling in HR/CPR:

1. Registration error at building level
2. Registration error at sub-building level
3. Under-coverage of dwellings in DR

   Source **ABR**: Address Building Register

   Single dwelling in ABR, multiple dwellings in fact

   NB. not a registration error; cannot be registered

Dual system estimator (DSE) with **clustered** elements

	Fixed A	Random B	Matched AB
Cluster (building)	x	n	m
Element (sub-building)	x_U	n_U	m_U

DSE of clusters and elements, respectively:

$$\widehat{M} = \frac{x}{m}n \qquad\qquad \widehat{N} = \frac{x_U}{m_U}n_U$$

Bias due to under-coverage of dwellings in ABR:

$$\frac{(x_{1U} + x_2)(n_{1U} + n_2)}{m_{1U} + m_2} \underset{<}{\geqq} \frac{(x_{1U} + x_{2U})(n_{1U} + n_{2U})}{m_{1U} + m_{2U}}$$

NB. cardinality(ABR) $x = x_1 + x_2$: under-counting x_{2U}

Two-step estimation approach

Domains of all building-level addresses with dwellings

	In ABR	Out of ABR	Total
In CPR	M_{11} (N_{11})	M_{10} (N_{01})	M_{1+} (N_{1+})
Out of CPR	M_{01} (N_{01})	M_{00} (N_{00})	M_{0+} (N_{0+})
Total	M_{+1} (N_{+1})	M_{+0} (N_{+0})	M (N)

$(M_{11}, M_{10}, M_{01}, M_{00})$: building-level addresses

$(N_{11}, N_{10}, N_{01}, N_{00})$: within-domain elements

$N = X$: no. dwellings

$N = Y$: no. resident dwellings

Two-step estimation approach

Step 1: DSE $\widehat{M}$ and $\widehat{M}_{00}$

$$\widehat{M} = \frac{M_{1+}M_{+1}}{M_{11}} \qquad\qquad \widehat{M}_{00} = \frac{M_{10}M_{01}}{M_{11}}$$

Step 2: Estimation of $\widehat{N}$ given $(\widehat{M}, \widehat{M}_{00})$

(1) missing-completely-at-random (MCAR)

(2) missing-at-random (MAR)

[NB. MACR $\Leftrightarrow$ MAR if DSE $\widehat{M}_{00}$]

(3) 4-way log-linear models

(2) MAR [same estimates as (1) MCAR]

	Dwelling		
	In ABR	Out of ABR	
In CPR	$X_{11} : M_{11}$	$X_{10} : M_{10}$	$\widehat{X}_{10} = \frac{X_{11}}{M_{11}} M_{10}$
Out of CPR	$X_{01} : M_{01}$	$X_{00} : \widehat{M}_{00}$	$\widehat{X}_{00} = \frac{X_{01}}{M_{01}} \widehat{M}_{00}$

Assumption (2.x): $X_k \perp \delta_k^{ABR} | \delta_k^{CPR}$

	Resident Dwelling	
	In ABR	Out of ABR
In CPR	$Y_{11} : M_{11}$	$Y_{10} : M_{10}$
Out of CPR	$Y_{01} : M_{01}$	$Y_{00} : \widehat{M}_{00}$
	$\widehat{Y}_{01} = \frac{Y_{11}}{M_{11}} M_{01}$	$\widehat{Y}_{00} = \frac{Y_{01}}{M_{01}} \widehat{M}_{00}$

Assumption (2.y): $Y_k \perp \delta_k^{CPR} | \delta_k^{ABR}$

(3) Log-linear model-I: $[\delta^{ABR}][\delta^{CPR}][Z^{ABR}Z^{CPR}]$

	In ABR	Out of ABR
In CPR	$Z_k^{CPR} \times Z_k^{ABR} : M_{11}$	$Z_k^{CPR} \times Z_k^{ABR} : M_{10}$
Out of CPR	$Z_k^{CPR} \times Z_k^{ABR} : M_{01}$	$Z_k^{CPR} \times Z_k^{ABR} : \widehat{M}_{00}$

$X_k \mapsto Z_k^{ABR}$: e.g. address with (1, 2, 3+) dwellings

$Y_k \mapsto Z_k^{CPR}$: e.g. address with (1, 2, 3+) resident dwellings

Of M_{10} addresses: observe Z_k^{CPR} (row total) not Z_k^{ABR} (column)

$\widehat{M}_{ij}^{10}/M_{i+}^{10} = M_{ij}^{11}/M_{i+}^{11}$ Assumption (3.x): $Z_k^{ABR} \perp \delta_k^{ABR} | Z_k^{CPR}$

Of M_{01} addresses: observe Z_k^{ABR} (column total) not Z_k^{CPR} (row)

$\widehat{M}_{ij}^{01}/M_{+j}^{01} = M_{ij}^{11}/M_{+j}^{11}$ Assumption (3.y): $Z_k^{CPR} \perp \delta_k^{CPR} | Z_k^{ABR}$

Of $\widehat{M}_{00}$ addresses: observe neither Z_k^{CPR} (row) nor Z_k^{ABR} (column)

$\widehat{M}_{ij}^{00}/\widehat{M}_{00} = M_{ij}^{11}/M_{11}$ Assumption (3.xy): $(\delta_k^{CPR}, \delta_k^{ABR}) \perp (Z_k^{CPR}, Z_k^{ABR})$

Model-II: $[\delta^{ABR}Z^{CPR}][\delta^{CPR}Z^{ABR}][Z^{ABR}Z^{CPR}]$

TPSE: Triple population size estimator

Relax Assumption (3.xy): $(\delta_k^{CPR}, \delta_k^{ABR}) \perp (Z_k^{CPR}, Z_k^{ABR})$ by

Assumption (3.xy'): $\delta_k^{CPR} \perp \delta_k^{ABR} | (Z_k^{CPR}, Z_k^{ABR})$

- Same $\widehat{M}_{ij}^{10}$ among M_{10} addresses by Assumption (3.x)
- Same $\widehat{M}_{ij}^{01}$ among M_{01} addresses by Assumption (3.y)
- Of M_{00} addresses, 'DSE' by Assumption (3.xy'):

$$\widehat{M}_{ij}^{00} M_{ij}^{11} = \widehat{M}_{ij}^{10} \widehat{M}_{ij}^{01}$$

NB. Equivalent to *constant interaction* between Z_k^{CPR} and Z_k^{ABR}

NB. Simultaneous estimation of address, dwelling and Resident Dwelling

see also Van der Heijden et al. (2018), Van der Heijden et al. (2018a)

Which estimator?

MCAR & MAR: can be rejected if heterogeneity observed

NB. MACR $\Leftrightarrow$ MAR if DSE $\widehat{M}_{00}$

Between the two log-linear models:

- both fit observed data $\{M_{ij}^{11}, M_{i+}^{10}, M_{+j}^{01}\}$ perfectly
- TPSE-model more relaxed assumption *a priori*
- difference limited by $\frac{\widehat{M}_{00}}{\widehat{M}}$, e.g. of 394 municipalities:

25%	50%	75%	95%	96%	97%	98%	99%	99.5%
.00015	.00037	.00084	.00383	.00521	.00586	.01373	.03751	.08795

Which estimator?

Municipality	$\widehat{M}_{00} : \widehat{M}$	Dwellings		
		MAR:MCAR	Model-I:MCAR	TPSE:MCAR
Oppdal	0.0351	1.0000	0.9799	0.9713
Kongsvinger	0.0001	1.0000	1.0159	1.0165
Trondheim	0.00005	1.0000	0.9932	0.9932
Oslo	0.0001	1.0000	0.9950	0.9949
Municipality	$\widehat{M}_{00} : \widehat{M}$	Resident Dwellings		
		MAR:MCAR	Model-I:MCAR	TPSE:MCAR
Oppdal	0.0351	1.0000	0.9963	0.9896
Kongsvinger	0.0001	1.0000	1.0064	1.0066
Trondheim	0.00005	1.0000	0.9986	0.9985
Oslo	0.0001	1.0000	0.9906	0.9906

Discrepancy to HR: error in registers and household definition

References

[1] Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, vol. **16**, pp. 113-131.

[2] Van der Heijden. P.G.M., Smith, P.A., Cruyff, M. and Bakker, B.F.M. (2018). An Overview of Population Size Estimation where Linking Registers Results in Incomplete Covariates, with an Application to Mode of Transport of Serious Road Casualties. *Journal of Official Statistics*, vol. **34**, pp. 239-263.

[3] Van der Heijden. P.G.M., Smith, P.A., Whittaker, J., Cruyff, M. and Bakker, B.F.M. (2018a). Dual and multiple system estimation. In *Analysis of Integrated Data, eds. L.-C. Zhang and R-L Chambers*. Chapman & Hall/CRC. *To appear.*

[4] Purcell, N.J. and Kish, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*, Vol. **48**, pp. 3 - 18.

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

[6] Zhang, L.-C. (2011). A unit-error theory for register-based household statistics. *Journal of Official Statistics*, vol. **27**, pp. 415-432.

[7] Zhang, L.-C. and Chambers, R.L. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society, Series B*, vol. **66**, pp. 479-496.

[8] Zhang, L.-C. and Hendriks, C. (2012). Micro integration of register-based census data for dwelling and household. *UNECE: Work Session on Statistical Data Editing, 2012.*

[9] Zhang, L.-C. and Fosen, J. (2018). *Register-based estimation of dwellings and households*. Methodological report for grant agreement 07112.2016.044-2016.594 Improvement of the use of administrative sources (ESS.VIP.ADMIN WP6)