WP. 5
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 Mai 2011)

Topic (i): Editing of administrative and Census data

IMPUTING HOUSEHOLD CHARACTERISTICS IN THE REGISTER BASED SURVEY OF THE SWISS POPULATION CENSUS

Invited paper

Prepared by Daniel Kilchmann (Daniel.Kilchmann@bfs.admin.ch), Swiss Federal Statistical Office

I. Introduction

1. The Swiss population Census 2010 is for the first time based on registers and completed with a sample survey for information not available in the administrative data of the municipalities. The administrative data contain demographic characteristics for each resident and an identification number representing the households. The latter is not available in some cases and therefore households must be imputed under coherence constraints with regard to the person characteristics as well as housing and dwelling characteristics. The statistical methods unit of the Swiss Federal Statistical Office evaluated a strategy based on the generation of the household characteristics through combinatorial methods combined with random selection methods. The strategy and its preliminary evaluation on provisional census data will be detailed and the resulting conclusions will be discussed.

II. The register based Swiss Census

2. The new Swiss Census is a combination of different sources and collection modes. The demographic variables are collected through a register based survey (RS). Further information on the household, occupation and education are collected with a sample survey of about 360'000 people called the structural survey of the federal population Census (SSFPC). These two surveys combined with the information from the Federal Housing and Dwelling Register (FHDR) cover the information collected in the past through classical Census'. Yearly sample surveys on specific themes and a smaller yearly sample surveys on actual themes complete the mosaic of the new Swiss Census.

3. The register for the RS is built up since 2010 with data from the municipalities by the means of a secured data channel. The municipalities have to send the respective data every three months starting by 30 September 2010. By 31 December 2010 they had to send all demographic data of their inhabitants with main and secondary residence as well as the corresponding Federal Building Identification Number (FBIN) linked with the FHDR and an identifier for households (HHID).

4. By 31 December 2012 the municipalities have to send for each person the Dwelling Identification Number (DIN) also linked to the FHDR which replaces the household identifier at this moment. The municipalities receive a log file listing the errors in their data. Above fixed thresholds the data are refused and the municipalities have to send a new file, see (POP(2010)).

5. The data of the municipalities are refused with regard to the FBIN and HHID (or DIN) if the missing rate is above 2% for the person data. Several municipalities informed the SFSO that this threshold

for the variable HHID (or DIN) may not be reached without an important loss of data quality. The SFSO suggested therefore to relax the threshold for these municipalities in order to enhance quality instead of quantity. This resulted in missing household information for about 5% of the population.

6. On the contrary, some municipalities and even some cantons sent the DIN for all their inhabitants by the 31 December 2010 already. Furthermore, most municipalities preferred to create the DIN instead of the HHID in their own data base expecting to save ressources as the DIN has to be sent anyway by 31 December 2012. Hence, the DIN is already available for about 80% of the population, whereas the HHID was sent for about 15% of the people.

7. The data quality and the existence of the DIN show a heterogeneous picture and household statistics suffer from this. In the best case, the municipalities sent the DIN for every inhabitant and in the worst case no DIN was sent and the quality of the HHID is poor.

III. **Treatment strategies**

8. The use of the SSFPC does not meet the users needs entirely as it is foreseen to publish classical Census results for the household characteristics by municipality. The missing household information has to be treated by statistical methods as the municipalities already used the maximum of auxiliary information to gather information about the households. The decision how to treat missing household characteristics has to be taken by the end of the year.

9. The aim of imputing household characteristics is to ensure the production of basic household statistics on the municipality level.

10. The demographic variables age, gender, marital status and nationality are available as the most useful auxiliary variables for our purposes on the person side. The number of dwellings, the number of rooms and the surface per dwelling are available on the housing side where the number of rooms variable is of better quality compared to the surface variable. The number of people inhabiting each housing can also be calculated through the FBIN and used for the treatment.

11. Three levels for imputation were considered with their different preconditions and resulting data:

(1) imputing household characteristics on the person level,
(2) imputing the link between households with HHID only and dwellings,
(3) imputing household characteristics to the dwellings.

12. Consistency on the housing/dwelling level and consistency on the person/household level need to be satisfied when imputing household characteristics on the person level. Hence, this strategy was abandoned due to the complexity of this problem.

13. The imputation of the link between households with HHID only and dwellings is not needed for calculating the basic household statistics and is therefore of no priority at the moment.

14. No person/household relation needs to be considered when imputing household characteristics to the dwellings. The only constraints for this imputation strategy are based on the number of inhabitants per housing, their demographic characteristics and the housing/dwelling characteristics. This strategy was therefore considered in detail and will be discussed in the following. However, a drawback of this strategy is that the household statistics have to be calculated on the dwelling data. Furthermore, only analysis for which the imputation is designed are possible because the link between people and dwellings will not be available.

IV. Imputing household characteristics to the dwellings

A. Preconditions

15. The following preconditions are needed to impute the household characteristics to the dwellings:

(1) Every person of the RS must have the right FBIN. In oder words, the link between the person and the housing are supposed to be reliable and are supposed to be entirely edited.
(2) The FHDR is complete and its information reflects reality. The FHDR has been consolidated since 2009 with regard to the Census.
(3) The available person-dwelling link, given by the DIN on the person level, are reliable. Only extreme situations detected by macro editing procedures such as when all inhabitants of a municipality live in the same dwelling or have all a dwelling on their own can be detected in the data. These person-dwelling links have to be changed to missing deliberately and the household characteristics have also to be imputed to the concerned dwellings.
(4) There must be enough people with a DIN in each imputation class. What 'enough' means depends on the definition of the imputation classes (regional level) which has to be determined during testing.
(5) The number of rooms is known for each dwelling.

B. Household typology

16. The number of inhabitants by number of rooms and household characteristics based on the demographic variables will be published on the municipality level. These household statistics do not include the relationship of the household members because this information is not available from the municipalities. The relationships between the inhabitants are however collected in the SSFPC.

17. The most basic household characteristic is the number of inhabitants of a dwelling. It was decided to use this household characteristic and a household typology with 22 modalities based on the number of inhabitants combined with their age and gender characteristics as a starting point for the imputation.

C. Methods

18. One housing where its inhabitants are missing the DIN and the HHID will be considered for describing the algorithms for imputing the household characteristics to the dwellings.

19. The imputation was divided into two main steps, see figure 1, in order to lower the complexity of the problem in each of these steps compared to the one step procedure. The latter would impute the household typology directly based on the available auxiliary information. The number of inhabitants per dwelling is imputed during the first main step. The second main step consists of imputing the household typology given the person characteristics linked with the housing and the number of household members per dwelling imputed in the first step.

20. The one step procedure can be seen as seeking the conditional probability function that given n people and m dwellings in one housing with the number of rooms $z_k, k \in \{1, \ldots, m\}$, the household typologies h_k are observed:

$$P(\mathbf{h}|n, m, \mathbf{z}). \tag{1}$$

21. In terms of the most basic household typology (1) is equivalent to (2), with x_k being the number of inhabitants of dwelling k.

$$P(\mathbf{x}|n, m, \mathbf{z}). \tag{2}$$

22. The second main step is equivalent to looking for the conditional probability function that given n people, m dwellings, the number of inhabitants and the number of rooms, z_k, per dwelling, the household typologies h_k, are observed:

$$P(\mathbf{h}|n, m, \mathbf{z}, \mathbf{x}). \tag{3}$$

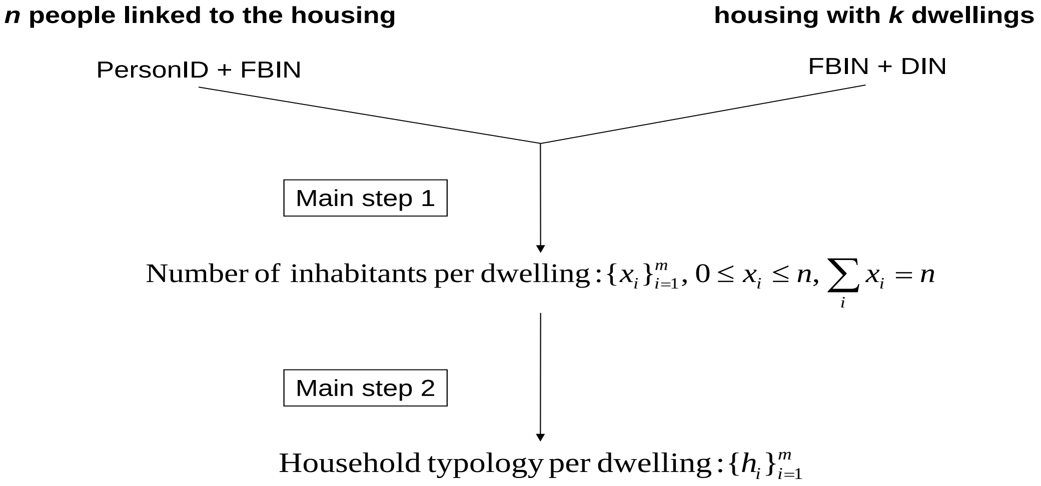
n people linked to the housing
housing with *k* dwellings
PersonID + FBIN
FBIN + DIN
Main step 1
Number of inhabitants per dwelling : $\{x_i\}_{i=1}^{m}$, $0 \leq x_i \leq n$, $\sum_i x_i = n$
Main step 2
Household typology per dwelling : $\{h_i\}_{i=1}^{m}$

FIGURE 1. The two main steps performed to impute household characteristics to the dwellings.

23. The first method considered for each imputation step is called the empirical likelihood method, or shortly likelihood method, where the imputation is based on the empirical likelihood calculated on housings with people all having a DIN.

24. The second method under investigation at the moment is the CUBE method, (Deville, J.-C. and Tillé, Y.(2004)). The following discussion is limited to the likelihood method because the application of the CUBE method to the outlined imputation problems is not yet finished.

V. Imputation of the number of inhabitants

25. The problem of imputing the number of inhabitants to the dwellings can be seen as an allocation problem of n people to m dwellings of one housing. Without taking auxiliary information into account there are $\binom{n+m-1}{n}$ possible allocations of people to the dwellings. In other words, this can be seen as all possible simple random samples with replacement of size n from a population of size m. Therefore, in order to diminish the workload a subsample of all possible allocations may be generated by selecting several simple random samples with replacement with the risk to cover the set of allocations only partially.

A. Empirical likelihood method

26. The empirical likelihood method, also called the simple likelihood method, takes advantage of the observations made above by generating different combinations of the number of inhabitants in the dwellings through repeated simple random sampling with replacement at first.

27. Once possible allocations were determined, a selection probability was calculated for each of them. This selection probability is based on marginal distributions obtained from the dwellings of housings with people all having a DIN and given the number of rooms of the dwellings. These distributions are calculated on a regional level like the municipality or with respect to the municipality typology, see table 7 in the appendix for the municipality typology used.

28. An allocation is then selected randomly based on the empirical likelihood being the product from the marginal distributions.

29. The empirical likelihood method for the imputation of the number of inhabitants per dwelling can be summarized by the five step algorithm given below. This algorithm is applied to each housing with people missing the DIN and the HHID, except of step 3 which is executed only once.

Step 1: Random ordering of the dwellings.

Step 2: Generate d' combinations of the number of inhabitants in the dwellings through repeated simple random sampling with replacement resulting in d different combinations.
Step 3: Based on the housings with people all having a DIN, the relative frequencies of the number of inhabitants given the number of rooms are calculated.
Step 4: Calculate the likelihood for each of the d combinations based on the relative frequencies from step 3.
Step 5: The likelihood from step 4 is used to select one of the combinations randomly and proportionally to the likelihood.

30. Only observed number of inhabitants will be imputed by the means of the use of the empirical relative frequencies. In this way the consistency with observed households is maintained.

31. Example for illustration
Suppose 4 people were linked with a housing having two dwellings, one has 2 and the other one has 3 rooms. Through repeated simple random sampling with replacement the combinations of the number of inhabitants in the two dwellings were generated, see the first two columns of table 2. The relative frequencies calculated on the housings with people all having a DIN are given in table 1.

32. The random number $\tau = 0.4$ was generated for this housing and therefore the combination with τ in the selection interval of L, corresponding to two people for each dwelling for $\tau = 0.4$, is selected with the simple likelihood method as illustrated in table 2.

TABLE 1. Fictitious relative frequencies of the number of people for 2- and 3-room dwellings.

	# people				
# rooms	0	1	2	3	4
2	0.01	0.49	0.46	0.04	0.00
3	0.01	0.59	0.36	0.03	0.01

TABLE 2. The combinations of the number of people per dwelling (# people) in the illustrative example. Dwelling 1 (D1) has 2 rooms, Dwelling 2 (D2) has 3 rooms. '0' stands for a vacant dwelling. The relative frequencies from table 1 were used to calculate the likelihood denoted by $\mathbf{L}$, the root of L, denoted by $\sqrt{\mathbf{L}}$ and the likelihood after calibration, $\mathbf{L_c}$. $\varnothing$ stands for an empty selection interval, i.e. the selection probability is 0.

# people		**rel. freq.**					**selection interval**		
D1	**D2**	**D1**	**D2**	$\mathbf{L}$	$\sqrt{\mathbf{L}}$	$\mathbf{L_c}$	$\mathbf{L}$	$\sqrt{\mathbf{L}}$	$\mathbf{L_c}$
0	4	0.01	0.01	0.00	0.01	0.00	$]0.00; 0.00]$	$]0.00; 0.01]$	$]0.00; 0.00]$
1	3	0.49	0.03	0.01	0.12	0.00	$]0.00; 0.07]$	$]0.01; 0.19]$	$]0.00; 0.00]$
2	2	0.46	0.36	0.17	0.41	1.00	$]0.07; 0.88]$	$]0.19; 0.78]$	$]0.00; 1.00]$
3	1	0.04	0.59	0.02	0.15	0.00	$]0.88; 1.00]$	$]0.78; 1.00]$	$]1.00; 1.00]$
4	0	0.00	0.01	0.00	0.00	0.00	$\varnothing$	$\varnothing$	$\varnothing$

33. The m-th root of the likelihood was used as an alternative in step 5 to balance the selection probabilities of the combinations. This alternative is denoted by $\sqrt[m]{L}$-method in the following. In the example for illustration $m = 2$, see columns $\sqrt{\mathbf{L}}$ in table 2. The same solution is selected as by using the simple likelihood method. The $\sqrt[m]{L}$-method balances the likelihood as expected, resulting in higher probabilities for rare combinations compared to the L-method.

34. Another alternative which is under investigation consists in calibrating the likelihood in step 4, see columns $\mathbf{L_c}$ in table 2. The calibration method, called L_c-method, ensures that

$$\sum_i l_{i,k} = 1, \ \forall k \in \{1, \dots, m\}, \text{ and } \sum_{i,k} n \cdot l_{i,k} = n,$$

where $l_{i,k}$ is the empirical likelihood used for assigning $i - 1$ inhabitants to the dwelling k. It is planned to apply the `calmar2` SAS-macro to perform the calibration, see (INSEE(2004)). The same solution is selected with all three methods in the example. But the selection probability of this solution is close to 1 with the L_c-method. In fact, the probability mass is concentrated to allocations having a high likelihood L with the L_c-method.

VI. Imputation of the household typology

35. To impute the household typology the following preconditions are needed in addition to those listed under item 15:

(6) The number of inhabitants per dwelling is known for all dwellings due to the main step one, see paragraph V.
(7) There are enough people having a DIN to calculate the distribution of the household typology linked to the dwellings given the number of rooms.

36. The number of inhabitants of the dwelling k is denoted by $x_k, k \in \{1, \dots, m\}$, $x_k \geq 0$, such that for a given housing $\sum_k x_k = n$, with n people linked to the housing. There are $\binom{n}{x_k}$ possible person links for each dwelling where the unique link for vacant dwellings is to allocate no people. The number of all possible person links to the dwellings, n_v, is given by the multinomial coefficient:

$$n_v = \binom{n}{x_1 \; x_2 \; \dots \; x_m} = \frac{n!}{x_1! \cdot x_2! \cdot \dots \cdot x_m!}. \tag{4}$$

A. Empirical likelihood method

37. The algorithm of the empirical likelihood method to impute the household typology given the number of inhabitants per dwelling takes advantage of the multinomial distribution of the allocations of people to the dwellings. It can be summarized by the following seven step algorithm applied to each housing except of step 5 which is executed only once.

Step 1: Random ordering of the people and dwellings.
Step 2: Generate d' dwelling links of the n people to the m dwellings through a multinomial distribution.
Step 3: If necessary, calculate the number of people z_k linked to the dwellings and discard the combinations of dwelling links with $z_k \neq x_k$ for at least one dwelling. Therefore, d dwelling link combinations are retained.
Step 4: Calculate the household typology based on the people linked with the dwellings for each of the dwelling links.
Step 5: Based on the housings with people all having a DIN, calculate the relative frequency of the household typology given the number of rooms.
Step 6: Calculate the likelihood for each of the d dwelling link combinations based on the relative frequencies of step 5.
Step 7: The likelihood of step 6 is used to select one of the dwelling link combinations randomly and proportionally to the likelihood.

38. It is not foreseen to use the link between people and dwellings for analysis as the above algorithms do not include the consistency of the household members. The links between the people and the dwellings are generated randomly and can therefore not be considered as true or even likely links. The aim of producing household statistics is however achieved for the imputed household characteristics.

39. Example for illustration
There are $\binom{4}{2\;2} = \frac{4!}{2! \cdot 2!} = 6$ possible links between the people and the dwellings in the example for illustration under item 31, if two inhabitants were imputed to both the two- and the three-room dwelling. The possible links with two inhabitants in each dwelling are generated with the multinomial distribution and the person identifiers of table 3, see the first two columns of table 5. The algorithm for calculating the household typology uses the person characteristics of table 3 to generate the columns 3 and 4 of the table 5.

TABLE 3. Demographic characteristics of the people of the example for illustration. 'F' stands for female, 'M' for male.

Id	age	gender
1	35	M
2	29	F
3	22	F
4	10	M

40. The relative frequencies of the distribution of the household typology given in table 4 are used to construct columns 5 and 6, 'rel. freq.', of table 5. An allocation of the household typologies is selected randomly and proportional to the likelihood. Suppose, that the housing was assigned the random number $\tau = 0.4$, then the allocation of the household typologies with the selection interval containing τ is selected. Therefore, the household typology h_4 is imputed to the two-room dwelling and the household typology h_3 is imputed to the three-room dwelling with this likelihood method.

TABLE 4. Fictitious empirical relative frequencies of the household typology in 2- and 3-room dwellings. 'adult' stands for an adult person, 'vacant' for a vacant dwelling and 'rooms' for the number of rooms.

				Household typologie						
				F -	M -	2 adults -	F -	2 adults -		3 adults
rooms	vacant	1 adults	2 adults	child	child	child	2 children	2 children	3 adults	child
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}
2	0.01	0.49	0.26	0.16	0.04	0.02	0.02	0.00	0.00	0.00
3	0.01	0.59	0.30	0.05	0.01	0.03	0.00	0.01	0.00	0.00

TABLE 5. Possible links between people and dwellings given that both dwellings have two inhabitants. The first two columns refer to the identifiers in table 3. Dwelling 1 has two rooms and dwelling 2 has three rooms. The notation is the same as in table 2.

HH composition		HH-typologie		rel. freq.					selection interval		
D1	D2	D1	D2	D1	D2	$\mathbf{L}$	$\sqrt{\mathbf{L}}$	$\mathbf{L_c}$	$\mathbf{L}$	$\sqrt{\mathbf{L}}$	$\mathbf{L_c}$
1,2	3,4	h_3	h_4	0.26	0.05	0.01	0.11	0.00	]0.00;0.10]	]0.00;0.14]	]0.00;0.00]
1,3	2,4	h_3	h_4	0.26	0.05	0.01	0.11	0.00	]0.10;0.19]	]0.14;0.28]	]0.00;0.00]
1,4	2,3	h_5	h_3	0.04	0.30	0.01	0.11	0.00	]0.19;0.28]	]0.28;0.41]	]0.00;0.00]
2,3	1,4	h_3	h_5	0.26	0.01	0.00	0.05	0.00	]0.28;0.30]	]0.41;0.47]	]0.00;0.00]
2,4	1,3	h_4	h_3	0.16	0.30	0.05	0.22	0.50	]0.30;0.65]	]0.47;0.74]	]0.00;0.50]
3,4	1,2	h_4	h_3	0.16	0.30	0.05	0.22	0.50	]0.65;1.00]	]0.74;1.00]	]0.50;1.00]

41. Again the m-th root of the likelihood, $\sqrt[m]{L}$-method, was used as an alternative in step 7 to balance the selection probabilities of the combinations, see columns $\sqrt{\mathbf{L}}$ in table 5. The solution h_5 for the two-room dwelling and h_3 for the three-room dwelling is selected with $\tau = 0.4$. The selection of a different solution compared to the solution of the L-method is due to the balancing of the distribution by this alternative method.

42. A calibration method, called L_c-method, is also under investigation for the imputation of the household typology given the number of inhabitants per dwelling. The L_c-method consists in calibrating the likelihood in step 6, see columns $\mathbf{L_c}$ in table 5. The calibration ensures that

$$\sum_i l_{i,k} = 1, \, \forall k,$$

where $l_{i,k}$ is the empirical likelihood used for assigning the household typology of the i-th combination to the dwelling k. It is planned to apply the `calmar2` SAS-macro to perform the calibration. The same solution is selected in the example for illustration with the L- and the L_c-method. Similar to the application of the L_c-method to the problem of imputing the number of inhabitants, the probability mass is concentrated to combinations where L his high.

VII. Simulation study

A. Simulation setup

43. A two stage sample of 504 municipalities at the first stage and 11’687 housings with all inhabitants linked to the dwellings selected at the second stage was used for the simulation study. The sample was drawn in a population of 2’551 municipalities and 87’932 housings where one-dwelling housings and housings with at least one household with more than 20 members were excluded. The sampled housings are occupied by 91’538 people linked to 43’614 dwellings (households) through the DIN from a total of 50’464 dwellings belonging to the sample of housings.

44. It is assumed that the missing rate in the RS data will not be higher than 10%. Therefore, housings with all inhabitants missing the DIN were generated randomly among the sample of housings with a rate of 10% and a MCAR and a MAR non-response mechanism where the latter depended on the municipality typology, see table 7 in the appendix for the municipality typology.

45. The municipality typology was also used to build the imputation classes as a starting point. The imputation of the number of inhabitants was then applied with the generation of 100 combinations of this number and with the L- and the $\sqrt[m]{L}$-method.

46. The generation of the missing DIN and the imputation was repeated 50 times to evaluate the accuracy of the imputation. The SAS programs run for about three hours to perform the simulation study. First preliminary analysis were realized by means of the imputation error, equation (5), where $x^*_{i,k}$ is the true number of inhabitants of dwelling k of the housing i and $\hat{x}_{i,k}$ is its imputed counterpart.

$$IE = x^*_{i,k} - \hat{x}_{i,k} \tag{5}$$

Note that the relative average imputation error, equation (6), is always 0 even on the housing level. This is due to the characteristics of the imputation method where the number of inhabitants of the housing is always distributed to the dwellings. Therefore, differences are canceled out on the housing level.

$$RAIE = \frac{\sum_{i,k} x^*_{i,k} - \hat{x}_{i,k}}{\sum_{i,k} x^*_{i,k}} \tag{6}$$

47. Furthermore, the comparison between original frequencies and the frequencies observed after imputation of the number of people by the number of rooms they inhabit was carried out to asses the imputation accuracy. At the moment, there are however no quality criteria available for this assessment.

48. The household typology could not yet be imputed because its implementation is not finished at the moment.

B. Preliminary results

49. The first preliminary results show that the statistics listed in table 6 are comparable for both methods and both non-response mechanisms. However, the minimum of the L-method is higher compared to the L_c-method and it has also a lower standard deviation under both non-response mechanisms. These results are also illustrated in figure 2 in the appendix. The imputation may result in quite different number of inhabitants, see the minimum and maximum of the imputation error in table 6. This happens if at least one big household occupies the building, which occurs for about 1% of the imputed data.

50. The following discussion is limited to the L-method for its lower standard deviation compared to the L_c-method. The mean of the relative frequencies of the number of inhabitants over all municipalities based on the 50 imputation loops compared to the original relative frequencies, figure 3, shows all values in a small band width around the diagonal. The mean of the difference between original and imputed data of the relative frequencies of the number of people by number of rooms, figure 4, is for the big bulge of the data in the interval [-0.5%; 0.5%] and spreads out to -2% to 2% in extreme situations. These extreme situations have to be investigated in detail. The standard deviation of the imputation is

highest for the 2- to 5-room dwellings, figure 4, which which make about 90% of the test data and are therefore most concerned by the imputation in this simulation study.

51. The preliminary simulation results are rather encouraging but need to undergo deeper analysis. A sensitivity analysis of the parameters of the imputation method should also enable to decide on their choice with respect to data quality and the feasibility of the imputation procedure (running time).

TABLE 6. Some statistics of the imputation error, IE, under the MCAR and MAR non-response mechanism with the L- and $\sqrt[m]{L}$-method.

statistic	NR-mechanism			
	MCAR		MAR	
of IE	$\mathbf{L}$	$\sqrt[m]{\mathbf{L}}$	$\mathbf{L}$	$\sqrt[m]{\mathbf{L}}$
$\max$	20	20	11	12
P_{95}	2	3	2	3
Q_3	1	1	1	1
med	0	0	0	0
Q_1	-1	-1	-1	-1
P_5	-2	-3	-2	-3
$\min$	-7	-17	-9	-11
σ	1.51	1.82	1.48	1.74

VIII. Conclusions

52. There are three strategies for the imputation of household characteristics in the register based survey of the Swiss population Census. Imputation on the person level would result in data similar to the data where no imputation is needed but showed to be too complex. The second strategy, imputing the link between households with HHID only and the dwellings is of minor interest at the moment, as the basic household statistics can be calculated without that link. The third strategy, imputing household characteristics to the dwellings, having a lower complexity, was discussed in detail.

53. Splitting up the imputation of household characteristics to the dwellings into two main steps lowers the complexity of each step compared to the one step procedure, where the household typology would be imputed directly based on the person, housing and dwelling characteristics.

54. The two empirical likelihood methods investigated, called the simple likelihood, L-method, and the m-th root of the likelihood, $\sqrt[m]{L}$-method, show slightly different results in terms of the standard deviation of the imputation error. Compared to the L-method the $\sqrt[m]{L}$-method balances the selection probabilities of the possible solutions. A further likelihood method currently under investigation, which consists in calibrating the likelihood, showed that the probability mass is concentrated at the solutions with high L in the example for illustration. Based on preliminary results, the L-method seems to be the most accurate but the accuracy of the imputation methods has to be assessed in more detail.

55. The discussed methods and their applications have to be further developed and evaluated by the end of the year. A decision about their usability can only be taken at that moment. The following steps are planned at the moment to be carried out.

(1) Finish the implementation of the second main step, that is the imputation of the household typology given the number of inhabitants per dwelling.
(2) Further investigate the application of the CUBE method and the calibration of the likelihood.
(3) Analyse the sensitivity of the imputation procedure with respect to different parameters.
(4) The accuracy of different imputation classes should be evaluated in detail.
(5) Take into account special configurations where the housings show a mixture of people with DIN, HHID and without any identifier.
(6) Include partially formed households, identified by a special HHID, in the treatment.
(7) Subject matter specialists together with the methodological unit have to define quality criteria for the assessment of the imputation procedure.

IX. Appendix

TABLE 7. Swiss municipality typology.

Typology #	label
1	centre (big town)
2	suburban municipalities
3	municipalities with high income population
4	peri-urban municipalites
5	touristic centres
6	industrial and third sector municipalities
7	rural municipalities with commuting population
8	municipalities with a mix of agrarian and other type
9	exclusively agrarian municipalities

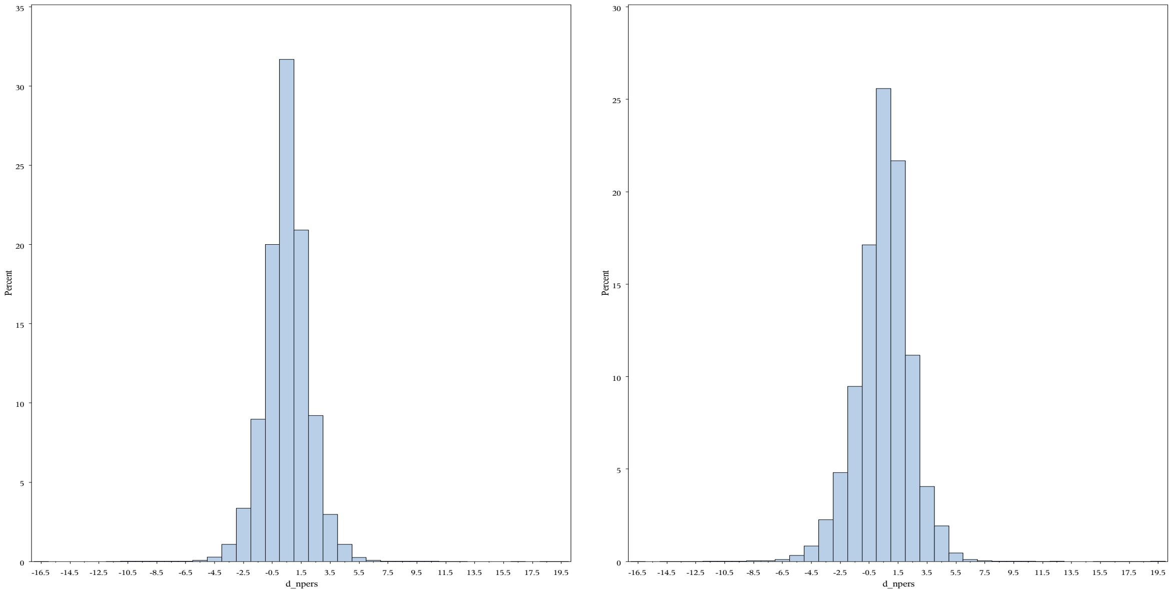

FIGURE 2. Distribution of the imputation error 'd_npers', equation (5), over all 50 simulation loops under the MCAR mechanism for the L-method, left panel, and the L_c-method, right panel. Attention: the scales of the vertical axes are different.

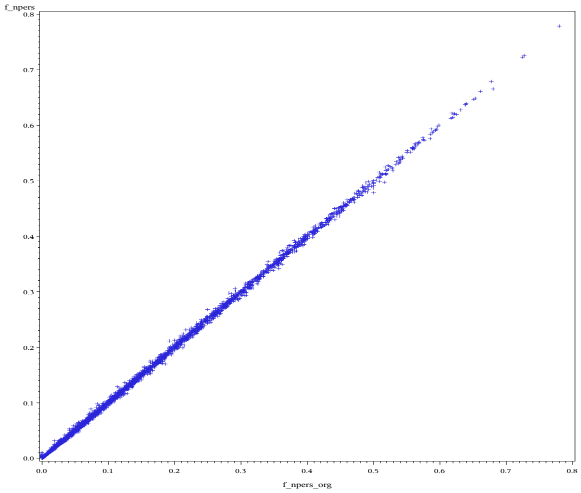

FIGURE 3. Comparison between the original mean relative frequencies of the number of inhabitants of the dwelling per municipality, ’f_npers_org’, and the same statistic after imputation with the L-method under the MCAR mechanism, ’f_npers’.

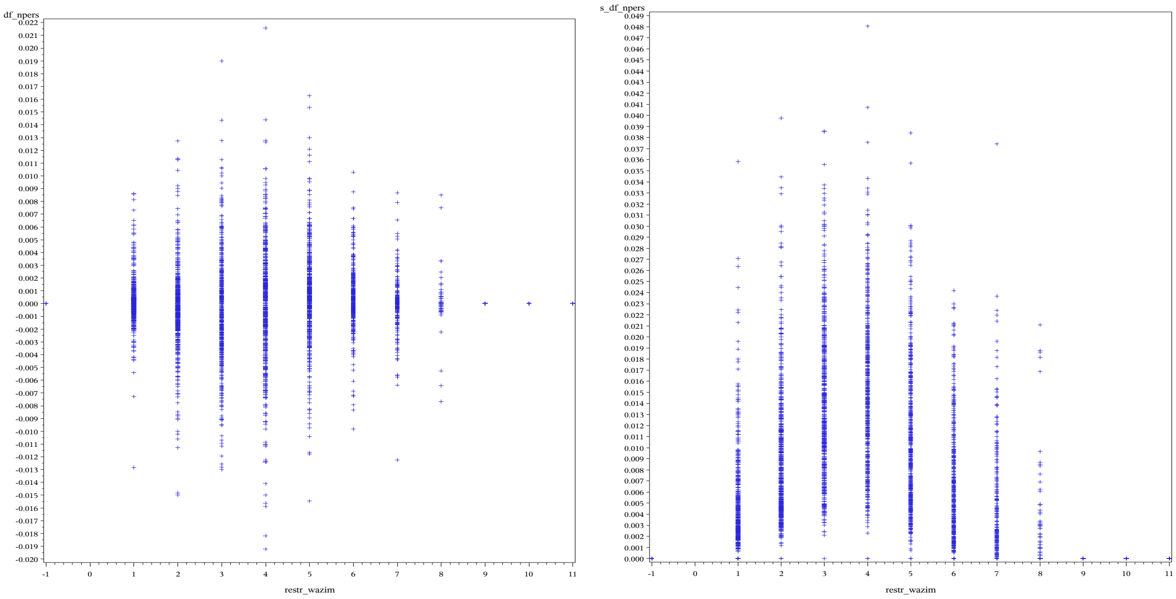

FIGURE 4. Mean difference between the original and imputed relative frequencies of the number of inhabitants of the dwelling by the number of inhabitants (’restr_wazim’), left panel, and its standard deviation, right panel. The L-method was applied under the MCAR mechanism. Note: there are no 0-room dwellings by definition.

References

Deville, J.-C. and Tillé, Y. Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912, 2004.

INSEE. *La macro CALMAR V2.0*. Institut National de la Statistique et des Études Économiques, PARIS, 2004. URL http://www.insee.fr/fr/methodes/default.asp?page=outils/calmar/accueil_calmar.htm.

POP. Validierung und Lieferung an die Statistik: Erklärung der Fehlermeldungen V09, 2010. URL http://www.bfs.admin.ch/bfs/portal/de/index/news/00/00/07.html.