WP.16
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (iii): Editing and Imputation in the context of data integration from multiple sources and mixed modes

Micro integration of register-based census data for dwelling and household

Prepared by L.-C. Zhang (lcz@ssb.no) and C. Hendriks (hen@ssb.no), Statistics Norway

Abstract

The coming census 2011 will be register-based in Norway, as in several other European Countries. Relevant data for dwellings and households stem from a number of different administrative sources, among which the Central Population Register (CPR) and the Ground Parcel, Address and Building Register (GAB) are the two most important ones. The existing household statistics are primarily based on the CPR, while the dwelling stock statistics are based on the GAB. Integration of the two sources on the micro level, with explicit linkage between households and dwellings, is important in order to produce detailed census statistics on unoccupied dwellings and housing conditions for the households occupying dwellings. However, various errors exist in the dwelling units in the GAB, the registration of dwellings in the CPR and the household units constructed from the CPR, such that not all the households can be directly linked to the dwellings in the GAB, and multiple households may be linked to the same dwelling in the GAB. Statistical methods of micro integration are needed in order to create a complete ("one-number") census data file of linked dwellings and households. We develop an approach that meets the need for detailed tabulation required by the census, while maintaining the existing 'marginal' statistics on the dwelling stock and the households.

I. Introduction

1. The coming census 2011 will be register-based in Norway, as in several other European Countries such as Denmark, Finland, Slovenia, Austria, *etc.* Relevant data for dwellings and households exist in a number of different administrative sources, regarding residence, family relationship, study and work, buildings, property ownership, *etc.* To simplify the details for the purpose here, one may envisage only *two* data files, to be referred to as the CPR (Central Population Register) and the GAB (Ground Parcel, Address and Building Register). Moreover, consider the CPR to contain all the relevant person and household information, including the person identity number (PIN), the household identity number (HID) and the registered dwelling identity number (DIN) of each person; and consider the GAB to contain all the relevant building and dwelling information, including the DIN, and the PIN of the owner. The two data files can be linked through the DIN. Ideally, i.e. in an error-free universe, the linkage would be unequivocal in the sense that (i) no persons with the same HID in the CPR-file would be linked to different DINs in the GAB-file, (ii) no persons with different HIDs in the CPR-file would be linked to the same DIN in the GAB-file, and (iii) each DIN in the GAB-file that does not link to any PIN (or HID) in the CPR-file would constitute an unoccupied dwelling. In reality, however, all the rules (i) – (iii) are violated, such that statistics that involve *both* households *and* dwellings would not have acceptable accuracy if based on the actual linked data file directly.

2. To clarify the task at hand, we notice that the existing household statistics may be considered to be based on the CPR. The HID is statistically constructed and is subject to the "unit errors" (Zhang, 2011). Since 2006, the CPR has provided register-based detailed household statistics on a yearly basis,

which are fully compatible with the long-standing register-based household statistics in Finland and Denmark. Meanwhile, the dwelling and building stock statistics have been produced based on the GAB on a yearly basis, and the quality is by-and-large regarded as “fit-for-use”. In other words, acceptable ‘marginal’ household or dwelling statistics have been produced based on either the CPR or the GAB in the past. What has been lacking are housing statistics that involve both households and dwellings, such as the average dwelling area per household, *etc*. For the purpose of Census 2011, we set the target of data integration to be the creation of a complete (“one-number”) census data file of linked dwellings and households, which meets the need for detailed tabulation required by the census, while preserving the existing ‘marginal’ statistics on the dwelling stock and the households.

3. The rest of the paper will be organized as follows. In Section II we briefly outline the problems in the data sources and motivate for the task at hand. In Section III we describe a double nearest neighbor imputation method as a general approach of micro integration, in order to establish micro-level linkage between two types of units that are not directly linkable, i.e. between the DINs in the GAB-file that do not directly link to any HIDs in the CPR-file, and the HIDs in the CPR-file that do not directly link to any DINs in the GAB-file. The main results and conclusions will be presented in Section IV.

II. Sources of data and sources of error

4. Unit errors occur in the CPR-file if persons actually belong to different households are assigned the same HID, or if persons actually in the same household are assigned different HIDs. Several approaches have been devised to measure the extent and effect of unit errors (Zhang, 2011; Zhang, 2009). The following main sources of errors in the household data are worth mentioning.

- Relevance or definitional error. The most demanded household concept in socio-economic analysis may be referred to as the *living* household, where the household member share meals and facilities and in other ways function as a real socio-economic unit. Whereas it is the matter of residence that underpins the *dwelling* household concept feasible to the register-based approach. In practical terms the incompatibility is a result of the administrative legislation regarding the residence registration, where for many people it is possible to differentiate between the formal residence (i.e. registered in the CPR) and the actual residence (i.e. much closer to the concept of living household). A typical example of such persons in Norway is the students. Still, this is a problem that is beyond the scope of the task that we have chosen here.

- Coverage error. Traditionally, the typical causes for the coverage errors are the lack of reliable sources for distinguishing between private and institutional residents, and the under-reporting of emigration to a less extent. The political changes in Europe since the last census in 2001, however, have aggravated the problems related to cyclic immigrants, seasonal workers and the likes. The information in the CPR regarding residence or the family relationships is not as reliable for these people as the ‘regular’ or ‘permanent’ residents. Still, the group being rather small in proportion to the whole population, the ‘marginal’ household statistics do not suffer heavily from it. But it creates problems on the micro-level when linking the CPR-file to the GAB-file.

- Unit error. Delay or lack of updating in cases of people moving is initially a selection error at the administrative data source, similar to the sampling error in surveys (Zhang, 2012): it is an event that could have been observed but is not, just like in sampling a unit that could have been included in the sample but is not. In the processed CPR-file at the statistical office, the wrongly registered DINs may cause unit-errors of the involved households. Also the initial measurement errors, i.e. in cases of mistakes in the reported address, could result in wrongly register DINs and unit-errors in the CPR-file. Overall, two types of incidents are most common in the CPR-file: (a) distinct HIDs can share the same DIN, and (b) missing or invalid DINs. The net effect is that the DIN-based households, i.e. households defined by people sharing distinct DINs in the CPR-file, are too large on average, and there are too two many so-called ‘other-type’ of households.

5. When it comes to the GAB-file, the following sources of errors are worth mentioning.

- Coverage error. Demolishing or abandoning of dwellings is under-registered in the GAB. For integration with the household statistics, a particular problem concerns the classification of buildings for dwelling purposes. For example, a house in the countryside that is only used for recreational purposes may not be registered as such in the GAB, which causes over-coverage of the dwelling stock. On the opposite side, under-coverage is the case when extra dwellings created within an existing building are not reported or registered in the GAB. This happens e.g. with many lodgings in the basement or attics of detached houses.

- Measurement error. Both delay (or lack) of certain types of reporting (i.e. initial selection errors) and mistakes in reporting and/or registration (i.e. initial measurement errors) may result in measurement errors (including missing information) regarding the GAB-dwellings. A typical characteristic that suffers from this type of errors is the building year.

- Initial lack of integrability. While the GAB-file has been the basis for building and dwelling stock statistics, it has not been used for other purposes until now. As a result the identification protocol in the GAB was not fully compatible with the DIN. The situation has been improved. But it is likely that the quality of the DINs in the GAB-file is not yet up to the standard one might expect.

6. To summarize marginally, there are 2224730 private households based on the CPR-file for the population census 2011, and 2415859 dwellings at disposal based on the GAB-file for the dwelling census, and accordingly 191129 unoccupied dwellings. Integration of the two sources on the micro level, with explicit linkage between households and dwellings, is important for at least two reasons. Firstly, and obviously, linked households and dwellings are necessary for the housing statistics. Secondly, micro-linkage provides a means to quality assessment.

Table 1. Addresses by number of dwelling units and number of households

	Number of Households						
Number of Dwellings	0	1	2	3	4	5+	Total
0	0	11419	1878	690	332	481	14800
1	185778	1150995	59451	5656	1145	726	1403751
2	6210	57501	87891	17055	2702	933	172292
3	792	2500	4288	6469	3258	1603	18910
4	534	797	1301	2869	6315	2459	14275
5+	985	664	760	1065	1891	41452	46817
Total	194299	1223876	155569	33804	15643	47654	1670845

7. Table 1 provides an illustration. The unit of tabulation is "address", where one may distinguish between a *single-dwelling address (SDA)* which uniquely identifies a dwelling and a *multiple-dwelling address (MDA)* at which there are at least two dwellings. One part of the DIN identifies the address, and the other the part identifies dwelling or dwellings beyond address. The second part was introduced in connection with the last census in 2001. In any case, on linking the CPR- and GAB-file at the address level, it is possible to count, at each distinct address, how many HIDs can be found in the CPR-file, and how many DINs can be found in the GAB-file.

8. On the one hand, out of 1670845 addresses in Table 1, 194299 are 'unoccupied', which is already larger than the 191129 unoccupied dwellings derived from the marginal stocks. To this one may add all the 'unoccupied' dwellings at the addresses that correspond to the other cells below the main diagonal of Table 1. On the other hand, there are 14800 'invalid' addresses registered in the CPR-file, which can not be found in the GAB-file at all. In addition, at all the addresses corresponding to the other cells above the main diagonal of Table 1, there are more HIDs in the CPR-file than DINs in the GAB-file. It is possible to 'collapse' the households at such an address according to the limit imposed by the number of DINs. But the resulting household statistics would have been unacceptable.

9. There are thus clear evidences of the problems in the input data files. Had all the DINs in the GAB-file been valid, the problems would have to be caused by the various errors in the CPR-file. In reality, the validity of the DINs in the GAB-file can not be taken for granted, and the problems are not restricted to the CPR-file. To achieve better integration and improve the quality at a more fundamental level, one needs to revise both the CPR- and the GAB-files, and consequently the marginal household and building and dwelling stock statistics. Unfortunately, given the time limit, it is not possible for us to deal with the problems in this way for Census 2011.

10. We have therefore restricted ourselves to the task of creating a complete census data file of linked dwellings and households, which meets the need for detailed tabulation required by the census, while preserving the existing 'marginal' statistics on the households and the dwelling stock. The approach we have developed consists of two stages.

- At the first stage, two types of micro-linkages between the HIDs in the CPR-file and the DINs in the GAB-file are obtained *deductively*.
    - The overwhelming majority type corresponds to the case where the DIN associated with an HID in the CPR-file matches uniquely with a DIN in the GAB-file.
    - The other type corresponds to the case where an HID in the CPR-file identifies exactly the same people of the household as a DIN from the GAB-file, although the DIN associated with this HID in the CPR-file does not match with the corresponding DIN from the GAB-file. The DIN from the GAB-file is accepted as the true one provided some of the household members can be identified as the owner.

  Afterwards, the CPR-file is divided in two: the first set of HIDs are the ones that have been linked to the GAB-file, to be referred to as the HD set; whereas the second set are the HIDs that still have not been linked to the GAB-file, to be referred to as the HuD set. Similarly, the GAB-file is partitioned into a set that is in one-to-one correspondence with HD, and another set of DINs that have not been linked to the CPR-file, to be referred to as the DuH set.

- At the second stage, micro-linkages between HuD and DuH are obtained *statistically*, using the method described in Section III. Notice that there are now no keys by which HuD and DuH can be linked to each other unequivocally. Indeed, there are many elements of the HuD set that can not be linked to any elements of the DuH set at all, and *vice versa.*

11. An alternative to establishing explicit micro-linkages could be to produce the housing statistics based on the data associated with the HD set, which contains of about 85% of all the private households. For instance, a weighting scheme can be devised to achieve suitable adjustments. But this would have violated the policy of census production based on a complete data matrix, with all its practical complications as consequences. An alternative is to impute only dwelling characteristics for the HuD set. A complete census data file can be obtained in this way. But one would have to deal with the potential inconsistency towards the dwelling characteristics that can be derived directly from the GAB-file, which can be messy whether the restrictions are imposed initially during imputation or afterwards during tabulation. For instance, in some local areas the number of DINs in the GAB-file is actually smaller than the number of HIDs in the CPR-file, which obviously would impose constraints on the imputation (or tabulation) of dwellings characteristics for the HuD there.

III. Double nearest neighbor imputation

12. It is convenient to present figuratively as in Table 2 the general setting for our problem outlined above. Between two sets of units, denoted by A and B, there are *N* identified *matches*. The problem is to find additional matches (or linkages) between the remaining M_A units in set A and M_B units in set B. However, the information (or *keys*) associated with the M_A units in set A does not exist for the M_B units in set B, neither do the keys associated with the M_B units in set B exist for the M_A units in set A. Direct comparison (i.e. matching) between these remaining units are thus not possible.

In what we call *double nearest-neighbor imputation (DNNI),* the identified matches between the two sets function as the *1st-stage* donors. For instance, starting from set A, one would find for each unit $i = 1, 2, ..., M_A$ a nearest-neighbor (NN) among the matched units $j = M_A + 1, M_A + 2, ..., M_A + N$, denoted by $j(i)$ or $i \rightarrow j$, based on the keys $\mathbf{x}$ that are associated with set A. At the 2nd-stage, for each $j(i)$ where $i = 1, 2, ..., M_A$, i.e. now the 2nd-stage receptor, one finds its NN among the rest units of set B, i.e. $k = 1, 2, ..., M_B$, based on the keys $\mathbf{z}$ associated with set B, denoted by $k(j(i))$ or $i \rightarrow j \rightarrow k$. Notice that each linked $k(j(i))$ needs to be removed from the 2nd-stage donor set, so that it would not be linked to another remaining unit in set A, and the 2nd-stage NN-matching is terminated if either the set of receptors or the set of donors is 'emptied'. In this way, micro-linkages can be established for all the units of the smaller one of the two remaining sets, despite direct comparisons between the two are impossible. It does not matter if M_A and M_B are equal or not, and it is equally possible to carry out the DNNI the other way around, i.e. $k \rightarrow j \rightarrow i$.

Table 2. General setting for double nearest-neighbor imputation. (Empty cells marked over)

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	▨	$\mathbf{z}_{M_B}$

13. A few general remarks are worth noting. First of all, it should be pointed out that in practice the population is always partitioned into *blocks,* and the DNNI applied separately within the blocks, so that micro linkages will only be established among the two sets of units that belong to the same block. In our application of DNNI for census household and dwelling micro-linkage, the blocks can be the Municipalities or, at a lower aggregation level, the Statistical Tracts.

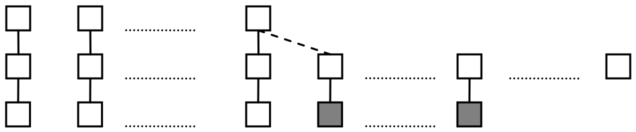

Figure 1: Illustration of DNNI with unique matches for either remaining set.

14. Next, there is a question regarding the 'direction' of DNNI, i.e. $i \rightarrow j \rightarrow k$ or $k \rightarrow j \rightarrow i$? It is instructive to consider the extreme case (Figure 1), where the 1st-stage donor set is so rich and the NN-criteria at both stages are so detailed that unique NN-matches can be found in all cases for either 1st-stage linkage. This is in fact the asymptotic setting implied by the regularity conditions required for the consistency of the NNI method (Chen and Shao, 2000), of course, used twice here. Now, suppose that one of the remaining sets is smaller than the other, represented by the top and bottom rows in Figure 1, respectively. Each square corresponds to a unit even though, strictly speaking, the two sets have different units, and the 'extra' units in the larger set are in shadow. The identified matches are represented by the middle row. The solid lines connecting the units indicate the NN-matches when each remaining set is linked to the units in the middle. First, starting from the smaller set, one would obtain the same NN-matches at the second stage as those indicated by the solid lines in Figure 1, now that the NN-matches are

all unique. Next, starting from the larger set, one would obtain the same matches, *provided the first units to be linked are all not in shadow*. Otherwise, if one of the units in shadow is attempted at the 1st-stage, it would pick out a unit in the middle that is not among the NN-matches for the smaller set. The 2nd-stage can still be accomplished, and an NN-match will be found among the smaller set, e.g. as illustrated by the dashed line in Figure 1, but it would not be among the 'best' NN-matches represented by the solid lines. So the order of the DNNI does matter, and it seems likely that starting from the smaller set would be preferable in most cases.

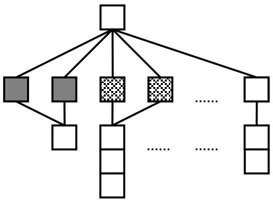

Figure 2: A general illustration of DNNI for a given unit.

15. The finiteness of any real data set means that *ties* exist among NN-matches. Consequently there is *linkage noise* that requires consideration. Figure 2 illustrates the general situation of DNNI for a given unit. All the middle row units are ties of NN-matches for the given unit at the 1st-stage. The 2nd-stage ties are stacked downwards. The 1st-stage ties do not introduce linkage noise if the 2nd-stage NN-matches is unique nevertheless (i.e. the shadowed cases), because exactly the same linkage would be established regardless of the 'route' it takes. i.e. $i \to j \to k$ is equivalent to $i \to l \to k$. The 1st-stage ties may or may not introduce linkage noise provided the characteristics of interest are identical for all the potential 2nd-stage NN-matches, even though these are different units, i.e. $i \to j \to k$ may be statistically equivalent to $i \to l \to g$ provided he characteristics of interest are the same for unit $k(j(i))$ and $g(l(i))$. The cross-marked units in Figure 2 illustrate this possibility. Otherwise, there is definitely linkage noise. Figure 2 shows that the linkage noise for a given unit can be measured exactly by going through all its potential DNNI-matches. Since it seems likely that, on average, more 1st-stage ties would entail greater linkage noise for the DNNI procedure, in practice one might choose to start with the set that is associated with the stronger set of keys on this account.

16. A more theoretically motivated investigation can be outlined as follows. Let $f(\mathbf{x},\mathbf{z})$ denote the joint distribution of the keys that holds for the remaining sets of units, and suppose $M = M_A = M_B$ to focus on the linkage noise. To start the DNNI from set A is to condition on $f(\mathbf{x}_1,...,\mathbf{x}_M)$, so that a model of the linkage noise may be derived from, say, $f(\mathbf{z}_1,...,\mathbf{z}_M \mid \mathbf{x}_1,...,\mathbf{x}_M) = \prod_i f(\mathbf{z}_i \mid \mathbf{x}_i)$. Whereas, to start from set B is to condition on $f(\mathbf{z}_1,...,\mathbf{z}_M)$, so that a model of the linkage noise may be derived, similarly, from $f(\mathbf{x}_1,...,\mathbf{x}_M \mid \mathbf{z}_1,...,\mathbf{z}_M) = \prod_i f(\mathbf{x}_i \mid \mathbf{z}_i)$. Since the joint distribution $f(\mathbf{x},\mathbf{z})$ is all the same, the choice with less linkage noise can be based on comparisons between suitable noise-measures derived from $f(\mathbf{x} \mid \mathbf{z})$ and $f(\mathbf{z} \mid \mathbf{x})$.

17. Finally, having chosen the DNNI in a given application, how should one evaluate the variance of an estimator based on the resulting linked data set? Consider a target total for the units of set A. It depends on the micro-linkages only if the variable of interest involves characteristics that are initially associated with the units of set B. In our application for census household and dwelling data, only the housing statistics are affected by the DNNI results, but not the 'marginal' household statistics, if set A consists of the households. Likewise, a target total for the units of set B depends on the micro-linkages only if it involves characteristics of the units of set A. Moreover, a sub-total for the identified matched units is also constant of the DNNI results. Therefore, in the variance due to DNNI, we propose to condition on all the initial keys and other relevant variables that are initially given with set A and B. A target total for the units of set A and its variance due to micro-linkage may then be given as

$$Y = \sum_{i=M_A+1}^{M_A+N} y_i + \sum_{i=1}^{M_A} y_i^* \qquad \text{and} \qquad V(Y) = V(\sum_{i=1}^{M_A} y_i^*)$$

where $y_i^* = y_i$ is a constant if it is given with set A, but it may be variable if it is obtained from $k(j(i))$. Notice that the variance becomes zero provided unique NN-matches as in Figure 1, conforming to the fact that there is no linkage noise in this case.

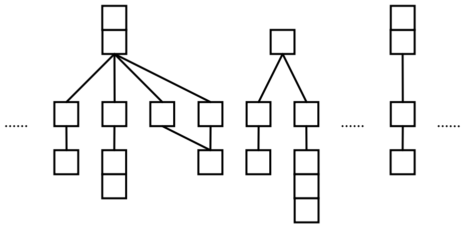

Figure 3: An illustration of inter-unit dependence under DNNI.

18. A general issue for this variance lies with the inter-units dependence of the DNNI procedure, as illustrated in Figure 3, where the initial units that share the same 1st-stage NN-matches are stacked upwards. When more than one units share the same 1st-stage NN-matches, micro-linkage for one of them obviously depends on the linkages for the others since, unlike the usual NNI-applications, each distinct 2nd-stage donor can only be selected once for the purpose of micro-linkage. The real data sets being finite, it is possible to go through all the possible micro-linkages for the units sharing the same 1st-stage NN-matches, in order to obtain the exact joint distribution of the y_i^*'s involved. But this can be time-consuming. A convenient simplification is to ignore the inter-units dependence altogether, following which an *approximate* variance may be given by

$$AV(Y) = \sum_{i=1}^{M_A} V(y_i^*)$$

The approximate variance is appropriate asymptotically under the setting that yields the situation illustrated in Figure 1. The exploration required for the exact ‘marginal’ variance of each y_i^* for $i = 1, 2, ..., M_A$, as in Figure 2, may be manageable. Otherwise, it remains to develop an appropriate model-based approach to $V(y_i^*)$, but we shall not go into details here given the limit of this paper.

IV. Results and conclusions

19. For the census application, there should be fewer households than dwellings in the respective the input data files. Also the keys are stronger for households than dwellings. On both accounts the DNNI should be carried out from the households to the dwellings. The HD set (i.e. identified matches and 1st-stage donors) contains 1899977 households (i.e. about 85 % of all private households). The HuD set consists of 324753 households, because there are 49500 households with unknown “household type”, for which micro-linkage is not required. The DuH set contains 515882 dwellings.

20. Figure 4 provides an overview of the actual scheme for DNNI. Two blocking variables are considered: Municipality and Statistical Tract. At the 1st-stage, up to four key variables are used for household NN-matching: household type, household size, average age of adults, and average age of children. In addition, the households are arranged by the dates on which they have moved to the address registered in the CPR, so that the most recently moved-in households will be the first ones to be linked to the ‘available’ dwellings at the same address. At the 2nd-stage, up to three key variables are used for dwelling NN-matching: address, building type, and living areal.

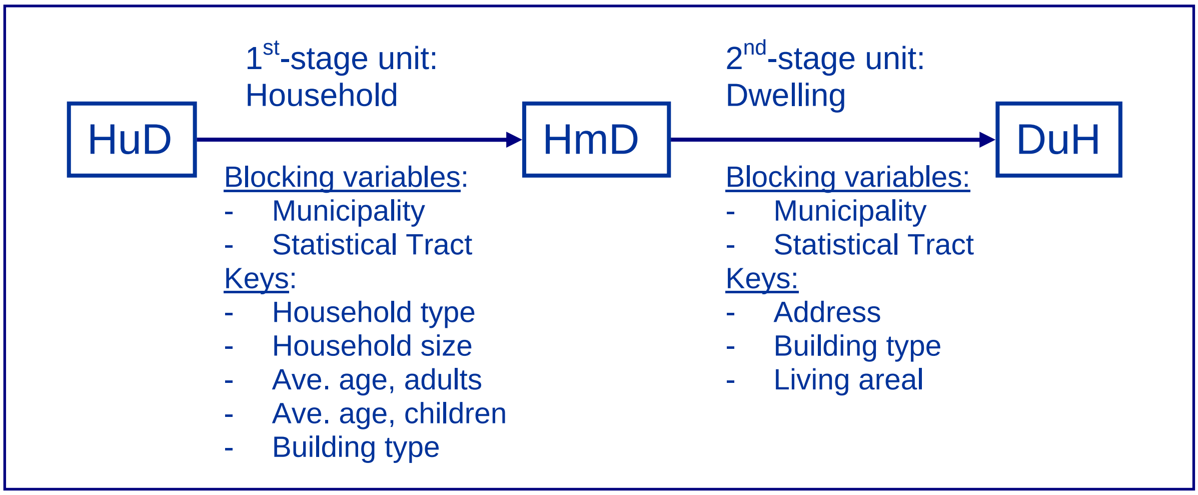

Figure 4. Application of DNNI for micro-linkageof households and dwellings in Census 2011.

Table 3. Main results of DNNI micro-linkage for census households and dwellings.

1st-Stage	***Characteristics of Nearest Neighbor-Match***	***Number***
Group 1	Exact Match on All Key Variables	303826
Group 2	Same Statistical Tract and Imperfect Match on All Key Variables	9022
Group 3	Same Municipality and Imperfect Match on All Key Variables	5574
Group 4	Outside Municipality and Imperfect Match on All Key Variables	6305
Group 5	Non-Match (below Minimum-Match Criterion for Key Variables)	26
2nd-Stage	*Characterization of Nearest-Neighbor-Match*	*Number*
Group 1	Same Statistical Tract and Same Address	159767
Group 2	Same Statistical Tract Different Address, Match on Building Type and Areal	65532
Group 3	Same Statistical Tract Different Address, Match on Building Type or Areal	65149
Group 4	Same Statistical Tract Only	15246
Group 5	Same Municipality Different Tract, Match on Building Type and Areal	2738
Group 6	Same Municipality Different Tract, Match on Building Type or Areal	6558
Group 7	Same Municipality ONLY	2298
Group 8	Non-Match	7465

21. Some results are given in Table 3. A couple of comments regarding the blocking are worthwhile. At the 1st-stage, blocking is not necessary in concept but preferable for statistical considerations. For instance, if two NN-households exist but one belongs to the same Statistical Tract and the other not, then the one from the same Statistical Tract is preferred, because the housing condition must to some extent vary across the country. Thus, Group 1 consists actually of sub-groups of NN-matches from the same Statistical Tract, the same Municipality but outside of the same Statistical Tract, and so on. When exact matches based on the household keys do not exist, there is a choice whether to lift the blocking or to drop the keys. The scheme presented here opts for the latter, albeit with a gradually decreasing degree of key-matching within each block. Alternative more refined schemes may achieve an even better balance between locality and household characteristics. However, we do not expect big impacts on the final results, since only about 6% of the 1st-stage NN-matches are not exact on all the key variables.

22. Blocking is necessary at the 2nd-stage to ensure consistency of the resulting housing statistics. Group 8 are the non-matches at the Municipality level. That is, for 7465 households, no 'available' dwellings can be found within the same Municipality at all, regardless of the dwelling characteristics. This reflects the problem mentioned earlier, i.e. in some local areas the number of dwelling stock according to the GAB-file is smaller that the number of households according to the CPR-file, and the problem concerns a handful Municipalities out of 429. An apparent consequence is that in these Municipalities there will be no "unoccupied" dwellings at all, which is not plausible. A 'superficial' remedy is to also declare Group 5-7 as "non-matches" at the 2nd-stage, and to classify all the non-matched

households to have “unknown” dwelling information. This would ensure consistency of the resulting housing statistics at the statistical track level. The existence of such unknowns would convey the message of potential errors to the users. A more ‘discrete’ remedy would be to revise the input files in those Municipalities in order to get rid of such non-matches afterwards. But it is still an *ad hoc* solution.

23. In conclusion, the method of DNNI described here provides a general approach to the problem of micro-linkage between different types of units for which direct matching is impossible to start with. To ensure the results of data integration on a more fundamental level, however, one must also look for means to improve the quality of the source data.

V. References

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

Zhang, L.-C. (2012). Topics of statistical theories for register-based statistics and data integration. *Statistica Neerlandica, vol.* **66**, pp. 41 – 63.

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

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