Statistical registers by restricted neighbor imputation – an application to the Norwegian Agriculture Survey

Nina Hagesæther[1] *and Li-Chun Zhang. Statistics Norway.*

Abstract

In this paper we implement the method of Zhang and Nordbotten (2008) for constructing a statistical register by the use of neighbor-imputation with restriction (RENI). Empirical results from the Norwegian Agriculture Survey are shown, and further research on the topic is discussed.

1 Introduction

Based on a survey, one may want to estimate the population total or mean for several variables. To use available data more efficiently, we can combine data from administrative sources and statistical surveys. A way to do this is to use a sample survey with target variables and an administrative register with auxiliary variables and construct survey weights for the sample units. This method is common in statistical production at the National Statistical Institutes. An alternative approach is to predict the values of the target variables for every unit outside the sample and construct a statistical register. Statistical tables will then be constructed based on all the population units, each of them having the weight equal to one. If one assumes that the probability of non-response is conditionally independent of the variable of interests given the auxiliary variables, the non-respondent units in the sample can be predicted in the same way as units outside the sample.

Zhang and Nordbotten (2008) proposed three goals they would like to achieve in constructing a statistical register:

1. It should yield efficient estimates of population totals of interest.
2. It should contain correct co-variances among the survey variables, as well as between the survey and auxiliary variables.
3. It should be non-stochastic, such that the statistics can be reproduced on repetition.

The first goal is traditionally associated with the production of statistical tables. The second one is a natural requirement now that a statistical register is a complete micro-data file. The last condition is important for the acceptability and face-value in official statistics: i.e. if a procedure is repeated, the results should be exactly the same. In the paper they present a table where common imputation methods are classified with respect to the triple-goal criterion:

[1] *e-mail: nih@ssb.no*

Table 1: Triple-goal classification of common imputation methods, from Zhang and Nordbotten (2008).

Method	(1)	(2)	(3)
Regression Prediction	Not always	No	Yes
Random regression imputation	No	No, if multivariate	No
Multiple Imputation	Not always	No, if multivariate	No
Predictive mean matching	Not always	No, if multivariate	Yes, in theory
Artificial Neural Network	Usually not	No, if categorical	Yes
Nearest neighbor imputation	Usually not	Yes, non-parametric	Yes, in theory

Nearest neighbor imputation (NNI) emerges as the only practical approach in terms of preserving the co-variances among all the variables. It is therefore a principal method for the construction of a statistical register. Consider the simple situation with a sample $(x_1, \boldsymbol{y}_1)$,..., $(x_r, \boldsymbol{y}_r)$ where $\boldsymbol{y}_i$ is a vector of all target variables known for the respondents r , $1 \leq i \leq r$, and all x-values are observed. The NNI method imputes $\boldsymbol{y}_j$ by $\boldsymbol{y}_i$, where $r+1 \leq j \leq N$ and N is the number of units in the population, and i satisfies

$$| x_i - x_j |= \min_{1 \leq l \leq r} | x_l - x_j |$$

and |.| denotes the chosen distance metric. In the case of multiple nearest neighbors, the donor i is randomly selected as one of them. Chen and Shao (2000) showed that NNI yields asymptotically consistent estimators of the population totals as well as the finite population distributions of interest. However, their empirical results also indicate that the method is often not efficient.

Zhang and Nordbotten suggested a new approach called RENI (restricted neighbor imputation), by imposing restrictions in the imputed totals, which may be obtained separately from the NNI such as through a regression prediction. A simple modification is introduced in order to make the NNI non-stochastic, and additional neighbors are included as potential donors, not only the nearest.

In section 2, the algorithm for the RENI approach is described. In section 3, we show empirical results where RENI is applied to real data from the Norwegian Agriculture Survey. In section 4 we discuss some further work that we have planned.

2 Algorithm for RENI

The algorithm to produce a RENI based statistical register in Zhang and Nordbotten (2008) is described as follows.

The jump-start (JS) phase Denote by R the set of receivers. Denote by D the set of donors. Let x_i be the variable (or variables) based on which the distance metric (and the NNI) is defined.

1. Set the counter $d_i = 0$ for all $i \in D$.

2. For each $j \in R$:

   a) Let m_j be the number of nearest neighbor (NN) donors, where $m_j \geq 1$.

   b) For each NN-donor, increase its counter d_i by $1/m_j$.

3. Let Y_R^0 denote the column vector of marginal restrictions for the receivers. Let y_i be the corresponding vector of variables for $i \in D$. Put

$$d_i' = d_i g_i \qquad \text{and} \qquad g_i = 1 + (Y_R^0 - \tilde{Y}_R)^T \tilde{A}^{-1} y_i$$

where $\tilde{Y}_R = \sum_{i\in D} d_i y_i$ and $\tilde{A} = \sum_{i\in D} d_i y_i y_i^T$. It is easily verified that $\sum_{i\in D} d_i' y_i = Y_R^0$.

4. Let $d_i' = a_i + u_i$, where a_i is the largest integer that satisfies $a_i \leq d_i$. Sort the receivers in the increasing order of m_j. For $j = 1,...,|R|$:

   a) Find the first NN donor i with $a_j \geq 1$. Impute $y_j^* = y_i$, and decrease a_i by 1.

   b) Do nothing if there is no NN-donor with positive a_i.

The fine-tune (FT) phase Denote by R' the remaining set of receivers that have not yet been imputed. Extend x_i to x_i' so that for each $j \in R'$ there is now a unique ordering among the potential donors by x_i and, some additional information z_i. For instance, z_i can be the post zip code, the identification number of unit, and so on. Notice that z_i is not considered informative. The unique ordering ensures that the procedure is non-stochastic.

1. Set k =1. For each $j \in R'$, find the NN-donor i and set $y_j^* = y_i$. Let Δ_1 be the distance between $Y_{R'}^0 = Y_R^0 - \sum_{j\in R\setminus R'} y_j^*$ and $Y_{R';k=1}^* = \sum_{j\in R'} y_j^*$ according to some chosen metric.

2. Set k =2. For $j \in R'$, let $D_{j;k=2}$ contain its two closest NN-donors.

   a) For each $j \in R'$, find the NN-donor i and set $I_j^* = i$.

   b) For $j = 1,...,|R'|$, set $I_j^* = i$ for $i \in D_{j;k=2}$ that yields a closer imputed total to $Y_{R'}^0$.

   c) Repeat step 2b until no changes can be made. Calculate Δ_2 between $Y_{R'}^0$ and $Y_{R';k=2}^*$.

3. Stop if Δ_2= 0, and use the imputations from Step 2. Otherwise, stop if $\Delta_2 \geq \Delta_1$, and use the imputation from Step 1. Otherwise, set k = 3 and let $D_{j;k=3}$ contain the three closest NN-donors for $j \in R'$.

   a) For each $j \in R'$, find the NN-donor i and set $I_j^* = i$.

   b) For $j = 1,...,|R'|$, set $I_j^* = i$ for $i \in D_{j;k=3}$ that yields a closer imputed total to $Y_{R'}^0$.

   c) Repeat step 3b until no changes can be made. Calculate Δ_3 between $Y_{R'}^0$ and $Y_{R';k=3}^*$.

4. Stop if Δ_3= 0, and use the imputations from Step 3. Otherwise, stop if $\Delta_3 \geq \Delta_2$, and use the imputation from Step 2. Otherwise, set k = 4 and let $D_{j;k=4}$ contain the four closest NN-donors for $j \in R'$ …

The following observations are worth noting:

- The JS is designed to speed up the process. Y_R^0 may consist of optional target variables.
- At each iteration of the FT phase the donor is the one among the k nearest neighbors that best satisfies the restrictions. The consistency of the NNI remains, as long as the difference between the conditional expectations of a unit and its k-th nearest neighbor is bounded by the 'distance' between them through a finite constant.
- Deviation from the marginal restrictions of the imputed totals can be reduced in two ways. Firstly, one may increase k to allow for greater combinatorial flexibility. Secondly, one may reduce the amount of imputations achieved by the JS, or even skipping completely over it.
- In constructing Δ, one can put a larger weight on the target variables that are considered more important. In return these will be closer to the restriction, on the expense of the other variables considered less important. The weight is called variable-weight and denoted by W.
- If one wants to introduce restrictions on a more detailed level than the imputation class, this can be incorporated in the algorithm. Y_R^0 can be extended to include sub-populations of the imputation class, and the Δ can be calculated also with respect to the sub-populations. A weight between 0 and 1, called marginal weight and denoted by W', may be included, to balance between the imputation class total and the sub-population total. If W' = 0, no sub-population totals are included in Δ, and if W' = 1, only sub-population totals are included.

3 Empirical results

We have tested RENI on the Norwegian Agriculture Survey 2006.

Agriculture Survey 2006. There are 10206 responding units which are the donors. There are 40993 non-responding and out-of-sample units which are the receivers. The Agriculture Survey 2006 has 84 variables of interest: 42 of them are size variables and the other 42 are indicator variables depending on whether the size is positive. Leasing, investment and maintenance were the most important topics including the following variables, all given in Norwegian kroner (VAT excluded):

- Leasing: leasing (paid leasing rent for machines), leasing1 (value of new leasing contract, contract 1) and leasing2 (value of new leasing contract, contract 2).
- Investment: i_fixed (sum invested in fixed technical equipment in outbuildings), i_build (sum invested in outbuildings), mach_new (purchase of other new machines and tools) and mach_used (purchase of other used machines and tools). We have also included sale_mach (sales of other machines and tools).
- Maintenance: m_build (maintenance of buildings and fixed equipment), m_tractor (maintenance of tractors and combine harvesters), m_car (maintenance of cars and trucks) and m_mach (maintenance of machines and tools).

Estimated numbers are published at the following classification levels:

- Farming system/activity (FA): 1=grain and oilseed crops, 2=other agricultural crops, 3=garden crops, 4=cattle, milk production, 5=cattle, meat production, 6=cattle, milk and meat production, 7=sheep, 8=other roughage animals, 9=swine and poultry, 10=mixed crop production, 11=mixed farm animal production, 12=crop and farm animal production
- Class of farmland in decares: 1=0-4, 2=5-49, 3=50-99, 4=100-199, 5=200-499 and 6=500+
- County

The imputation class is FA. This means that receivers can only get a donor with the same farming activity. Due to limited space we show only results four different imputation classes:

- FA-2: 2660 receivers and 727 donors. The sample fraction is about the same as for the whole population, and the proportion of receivers is about average.
- FA-4: 9984 receivers and 3296 donors. This is the largest imputation class in this study.
- FA-10: 384 receivers and 243 donors. This is the smallest imputation class in this study, with a high sampling fraction above 60 percent.
- FA-11: 586 receivers and 340 donors. This is another one of the smallest imputation classes.

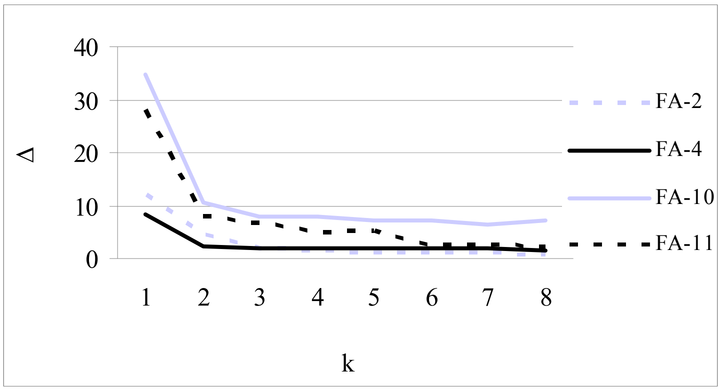
The variable used to find the nearest neighbor within each imputation class is farmland in decares and an identity number of the farms, where the latter of which is non-informative. The restrictions are the published totals of each imputation class, obtained by stratified ratio estimation.

3.1 Performance in satisfying the restrictions

First we examine how well the algorithm is able to satisfy the imposed restrictions. This is important in order to realize the efficiency goal.

Figure 1 below shows the effect of k on Δ, where Δ is calculated as the un-weighted sum of relative differences between the restriction and the imputed class total for all variables. Recall that k=1 means that only the nearest neighbor may be the donor, k=2 means that the donor may be selected among the two nearest neighbors, and so on. It is seen that Δ decreases rapidly as k increases from 1 to 3, and there is basically no change in Δ beyond k=6.

Figure 1: How Δ varies with k, for farming activities (FA) 2, 4 10 and 11.



We examine next how the JS affects the computation time. Five different scenarios of the JS phase where used, varying from no JS at all to JS with respect to all variables of interest. The results are summarized in Table 2 - 5, where Δ is the same as above. We notice that the number of units that is left for the FT phase depends mostly on whether the JS is employed or not, but less on the exact restrictions used in the JS phase. The computation time depends on number of iterations and number of receivers at the FT phase. Twice as many receivers, or twice as many iterations, takes about twice as long time --- the relationship is linear as shown in Figure 2. Since the number of iterations does not seem to depend on the JS at all, the JS can significantly reduce the computation time for large imputation classes such as FA-4. Meanwhile, the restrictions in the JS do affect the final Δ. Generally, it seems that the size variables more readily put the JS in the right direction, and using all the restrictions at the JS achieves the best fitting of the final imputed totals. In conclusion, the JS performed as intended, and using all the restrictions at the JS yielded final imputation totals that were basically as good as without the JS at all. The latter option requires of course much more computation time in large imputation classes.

Table 2: Shows how Δ and computation time depend on restrictions and number of units in the fine-tune phase. JS = jump-start, FT = fine-tune. Farming activity 2.

	Seconds	Iterations	Δ, k=6	Δ, k=1	Restrictions	# in JS	# in FT
No JS	83	4	1,16	22,95	-	-	2660
With JS	17	6	5,57	15,48	5 size variables	2345	315
With JS	14	5	1,48	12,07	All size variables	2335	325
With JS	9	4	5,23	12,84	All indicator variables	2411	249
With JS	11	4	1,09	11,95	All variables	2345	315

Table 3: Shows how Δ and computation time depend on restrictions and number of units in the fine-tune phase. JS = jump-start, FT = fine-tune. Farming activity 4.

	Seconds	Iterations	Δ, k=6	Δ, k=1	Restrictions	# in JS	# in FT
No JS	1546	3	4,17	33,6	-	-	9984
With JS	299	4	4,2	11,34	5 size variables	8420	1564
With JS	300	4	2,29	8,08	All size variables	8457	1527
With JS	317	4	3,93	10,17	All indicator variables	8580	1404
With JS	296	4	1,97	8,43	All variables	8556	1428

Table 4: Shows how Δ and computation time depend on restrictions and number of units in the fine-tune phase. JS = jump-start, FT = fine-tune. Farming activity 10.

	Seconds	Iterations	Δ, k=6	Δ, k=1	Restrictions	# in JS	# in FT
No JS	3,5	4	6,11	37,14	-	-	384
With JS	1,5	5	7,45	32,18	5 size variables	260	124
With JS	1,5	4	6,13	33,06	All size variables	253	131
With JS	1	4	8,95	30,08	All indicator variables	281	103
With JS	1,5	4	7,29	34,87	All variables	232	152

Table 5: Shows how Δ and computation time depend on restrictions and number of units in the fine-tune phase. JS = jump-start, FT = fine-tune. Farming activity 11.

	Seconds	Iterations	Δ, k=6	Δ, k=1	Restrictions	# in JS	# in FT
No JS	9,5	5	4,1	29,44	-	-	586
With JS	1,7	4	7,49	30,31	5 size variables	477	109
With JS	2,5	5	4,74	27,78	All size variables	450	136
With JS	2,3	4	3	29,82	All indicator variables	425	161
With JS	2,4	4	2,8	28,08	All variables	424	162

Figure 2: Shows how computation time (measured in seconds) divided on number of iterations varies with number of units in FT, for farming activity (FA) 2, 4, 10 and 11.

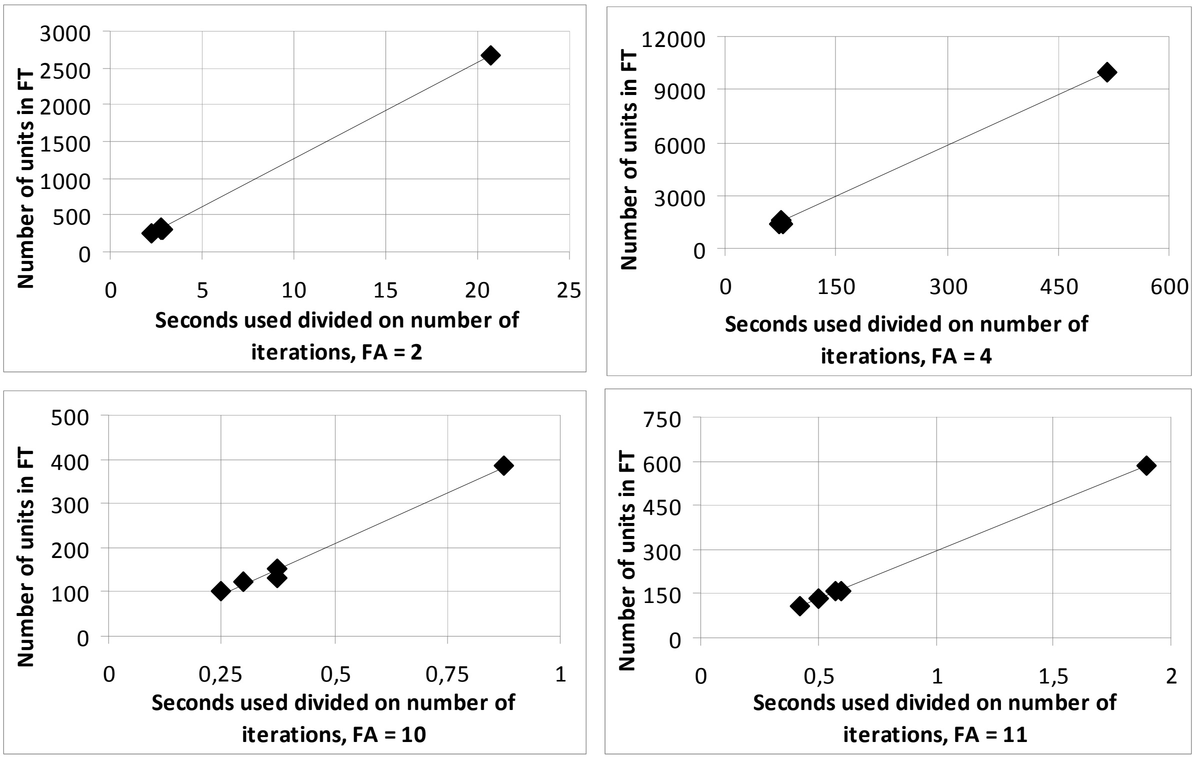


Table 6 and 7 show how the imputed totals in FA-4 and FA-10, respectively, satisfy the restrictions on the 12 most important variables listed above. All the imputations were carried out using full JS and with k =6. Five different ways of calculating Δ have been used. For the first scenario, all variables are considered equally important (that is, W=1), so that Δ is calculated as the un-weighted sum of the relative differences, and only totals at the imputation class level are considered when calculating Δ (that is, W'=0). For all other scenarios the weights for the 12 important variables are increased to W=10, i.e. these relative differences weight ten times as much as the rest in the calculation of Δ. For the third scenario, W'=0.1, which means that the restrictions for sub-populations (in this case, marginals of class and region) in calculation Δ is included to a small extent. For the fourth and fifth, the W' is increased to 0.5 and 0.9. Notice that in scenario three – five, the number of imputation restrictions is increased to 840 from 84 in the first two scenarios.

The results indicate that there is a limit on the number of restrictions that can be satisfied in any given imputation class. Moreover, the variables with skewed distributions are more likely to have problems. Here a skewed distribution is indicated by a small number (or proportion) of donors with positive values. It follows that in practice one may need to give priority to certain imputation restrictions, for instance the variables of the most interest or the restrictions that have the lowest variances.

Table 6: Table of results from five different scenarios of RENI compared to restriction, for the important variables. The second column shows number of donors within the imputation class that has a value larger than 0. W = variable weight, W' = marginal weight. Farming activity 4.

Variable	Donors > 0	Restriction	W=1, W'=0	W=10, W'=0	W=10, W'=0.1	W=10, W'=0.5	W=10, W'=0.9
leasing	615	107 725 189	0,00	0,00	0,00	0,00	0,00
leasing 1	201	141 312 208	0,00	0,00	0,00	0,00	-0,01
leasing 2	26	9 510 617	-0,01	0,00	0,00	0,01	0,11
m_build	2745	245 711 424	0,00	0,00	0,00	0,00	0,02
i_fixed	388	160 950 792	0,00	0,00	0,00	0,00	-0,02
i_build	351	328 141 992	0,00	0,00	0,00	0,00	0,00
m_tractor	2517	139 840 725	0,00	0,00	0,00	0,00	0,00
m_car	561	12 133 275	0,00	0,00	0,00	0,00	0,04
m_mach	2526	115 488 841	0,00	0,00	0,00	0,00	0,00
mach_new	707	143 487 579	0,00	0,00	0,00	0,00	0,01
mach_used	275	52 354 873	0,00	0,00	0,00	0,00	-0,02
sale_mach	257	28 362 849	0,00	0,00	0,00	0,01	0,06

Table 7: Table of results from five different scenarios of RENI compared to restriction, for the important variables. The second column shows number of donors within the imputation class that has a value larger than 0. W = variable weight, W' = marginal weight. Farming activity 10.

Variable	Donors > 0	Restriction	W=1, W'=0	W=10, W'=0	W=10, W'=0.1	W=10, W'=0.5	W=10, W'=0.9
leasing	59	4 831 568	-0,02	0,00	0,00	0,00	0,23
leasing 1	20	5 166 386	0,05	0,06	0,05	0,12	0,27
leasing 2	4	622 092	-0,04	-0,04	-0,04	-0,04	-0,04
m_build	198	7 107 608	0,00	0,00	0,00	0,00	-0,04
i_fixed	20	8 041 181	-0,04	-0,05	0,09	0,11	0,14
i_build	28	16 960 322	0,09	0,00	0,03	0,07	0,19
m_tractor	215	5 356 895	0,00	0,00	0,00	0,02	0,20
m_car	91	1 083 723	-0,11	0,00	0,00	0,00	0,00
m_mach	206	4 449 857	0,00	0,01	0,00	0,07	0,15
mach_new	58	6 385 295	0,00	-0,01	0,00	0,14	0,17
mach_used	18	1 078 862	0,05	0,01	-0,01	0,00	0,01
sale_mach	33	738 748	-0,26	0,00	0,05	0,21	0,25

3.2 Co-variances of the imputed values

It is essential that the imposed restrictions do not damage the consistency property of the NNI for the estimation of the finite population distributions. In particular, with regard to the second goal, the imputed data should contain correct co-variances among the survey variables.

We now compare the variances and co-variances in the statistical register obtained by the RENI and the corresponding estimates by the weighting method. Table 8 shows that the correlations by and large are very close under the two approaches in FA-10, and the situation is very similar in the other imputation classes for which the details are omitted due to the limit of space. The same conclusion holds in terms of the coefficient of variation under the two approaches (Table 9).

For the co-variation among indicator variables, we look at their cross-tables. Table 10 show these are very close under the two approaches in FA-10. Again, the situation is very similar in the other imputation classes such that the details are omitted.

Table 8: Co-variances for RENI (above diagonal) and weighting (under diagonal). Farming activity 10.

	leasing	leasing1	leasing2	m_build	i_fixed	i_build	m_tractor	m_car	m_mach	mach_new	mach_used	sale_mach
leasing		0,59	0,24	0,25	0,44	0,37	0,30	0,25	0,37	-0,03	0,12	0,07
leasing1	0,59		0,43	0,13	0,47	0,44	0,14	0,10	0,25	-0,01	0,17	0,10
leasing2	0,25	0,44		0,12	0,05	0,16	0,11	0,09	0,21	-0,02	-0,02	-0,02
m_build	0,25	0,14	0,14		0,07	0,08	0,35	0,59	0,40	0,13	0,23	0,07
i_fixed	0,45	0,47	0,05	0,06		0,80	0,05	-0,02	0,14	-0,03	-0,02	-0,02
i_build	0,37	0,42	0,15	0,08	0,79		0,07	-0,02	0,17	-0,03	0,00	-0,02
m_tractor	0,28	0,13	0,12	0,35	0,05	0,08		0,42	0,53	0,17	0,13	0,11
m_car	0,21	0,09	0,10	0,61	-0,01	-0,01	0,47		0,49	0,11	0,15	0,11
m_mach	0,35	0,25	0,24	0,41	0,16	0,18	0,57	0,52		0,28	0,18	0,30
mach_new	-0,01	-0,02	-0,02	0,15	-0,03	-0,03	0,20	0,15	0,32		0,11	0,60
mach_used	0,11	0,18	-0,01	0,27	-0,02	0,00	0,13	0,16	0,17	0,15		0,06
sale_mach	0,09	0,11	-0,02	0,07	-0,01	-0,02	0,12	0,13	0,32	0,56	0,06	

Table 9: Coefficient of variation for RENI and weighting, for all farming activities.

	FA = 2		FA = 4		FA = 10		FA = 11	
	R-NNI	Weighting	R-NNI	Weighting	R-NNI	Weighting	R-NNI	Weighting
leasing	0,0020	0,0019	0,0003	0,0003	0,0097	0,0097	0,0047	0,0047
leasing1	0,0029	0,0029	0,0006	0,0006	0,0191	0,0208	0,0098	0,0095
leasing2	0,0153	0,0149	0,0018	0,0018	0,0836	0,0823	0,0229	0,0226
m_build	0,0010	0,0010	0,0002	0,0001	0,0084	0,0082	0,0025	0,0025
i_fixed	0,0046	0,0046	0,0009	0,0008	0,0173	0,0164	0,0090	0,0094
i_build	0,0030	0,0030	0,0009	0,0009	0,0164	0,0166	0,0079	0,0073
m_tractor	0,0007	0,0007	0,0001	0,0001	0,0043	0,0043	0,0023	0,0023
m_car	0,0012	0,0012	0,0004	0,0004	0,0091	0,0091	0,0048	0,0045
m_mach	0,0010	0,0010	0,0001	0,0001	0,0053	0,0053	0,0028	0,0031
mach_new	0,0025	0,0026	0,0004	0,0003	0,0132	0,0117	0,0088	0,0081
mach_used	0,0036	0,0034	0,0005	0,0005	0,0237	0,0250	0,0066	0,0067
sale_mach	0,0031	0,0031	0,0006	0,0006	0,0225	0,0227	0,0126	0,0134

Table 10: Cross-table for the indicator variable of the important variables for the RENI (above the diagonal) and weighting method (under the diagonal). Farming activity 10.

		I_leasing		I_leasing1		I_leasing2		I_m_buil		I_i_fixed		I_i_build		I_m_tractor		I_m_car		I_m_mach		I_mach_new		I_mach_used		I_sale_mach	
		1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
I_leasing	1			14	42	2	54	50	6	8	48	12	44	56	0	33	23	48	8	6	50	5	51	1	55
	0			0	328	0	328	244	84	17	311	17	311	265	63	88	240	249	79	46	282	14	314	30	298
I_leasing1	1	11	45			2	12	14	0	5	9	7	7	14	0	7	7	12	2	2	12	4	10	1	13
	0	2	323			0	370	280	90	20	350	22	348	307	63	114	256	285	85	50	320	15	355	30	340
I_leasing2	1	2	54	2	12			2	0	0	2	2	0	2	0	2	0	0	2	0	2	0	2	0	2
	0	0	325	0	368			292	90	25	357	27	355	319	63	119	263	297	85	52	330	19	363	31	351
I_m_build	1	47	9	13	0	2	0			21	273	28	266	267	27	108	186	262	32	45	249	17	277	26	268
	0	252	73	286	82	298	82			4	86	1	89	54	36	13	77	35	55	7	83	2	88	5	85
I_i_fixed	1	8	48	5	8	0	1	22	277			9	16	22	3	16	9	17	8	1	24	1	24	0	25
	0	17	309	20	348	25	355	3	79			20	339	299	60	105	254	280	79	51	308	18	341	31	328
I_i_build	1	12	44	8	5	1	0	29	270	12	13			29	0	7	22	26	3	3	26	0	29	1	28
	0	19	306	23	345	30	349	2	80	20	337			292	63	114	241	271	84	49	306	19	336	30	325
I_m_tractor	1	56	0	13	0	2	0	266	33	24	1	31	0			112	209	262	59	44	277	14	307	31	290
	0	263	62	306	62	317	62	53	29	295	61	288	62			9	54	35	28	8	55	5	58	0	63
I_m_car	1	26	30	5	8	2	0	89	210	12	13	7	24	97	222			102	19	13	108	6	115	4	117
	0	77	248	98	270	101	278	14	68	91	265	96	254	6	56			195	68	39	224	13	250	27	236
I_m_mach	1	47	9	12	1	0	1	259	40	22	3	28	3	264	55	90	13			48	249	19	278	26	271
	0	248	77	283	85	295	85	36	46	273	83	267	83	31	31	205	73			4	83	0	87	5	82
I_mach_new	1	8	48	2	11	0	1	54	245	5	20	3	28	55	264	21	82	60	235			3	49	21	31
	0	56	269	62	306	63	316	9	73	59	298	61	289	9	53	42	236	3	83			16	316	10	322
I_mach_used	1	3	53	4	9	0	2	14	286	2	23	2	29	14	305	5	98	16	279	6	58			3	16
	0	13	312	11	357	16	364	2	80	14	342	14	336	2	60	10	268	0	86	10	307			28	337
I_sale_mach	1	4	52	2	11	0	2	23	276	0	25	1	30	30	289	6	97	26	269	21	42	4	11		
	0	26	299	27	341	30	350	6	76	29	327	29	321	0	62	23	255	3	83	8	309	25	340		

4 Summary and discussion of future works

Application of the RENI to the Norwegian Agriculture Survey 2006 suggests that the method is capable of fulfilling the triple-goal criterion for the construction of statistical registers.

The most problematic issue that we have encountered concerns how well the imputed totals satisfy the imposed restrictions. A large number of restrictions can easily be obtained by method of weighting. However, not all the weighted totals are equally reliable. One may choose those with small coefficients of variation and/or those that are considered important. It may also be helpful to examine systematically the effective choice of imputation restrictions. For instance, given the number of restrictions to be imposed, which ones should be chosen in order to obtain in addition the closest agreement of the rest totals?

The definition of the distance metric is another issue one might look into. For instance, what is a suitable balance between the absolute and relative differences?

Partial missing/non-response is as common as unit missing/non-response. In general, a donor may not have exactly the same values as the observed ones for the receivers, when observations are missing partially. It seems natural in such a situation that the observed values of the donors should be adjusted before they are imputed for the receiver. Otherwise, the covariance between the variables may be distorted. This requires extending the NNI to partially missing data. More generally, the donor and receiver may not exactly match on the x-values, either. Thus, the extension may have a more general bearing on the methodology.

5 References

Zhang, L. C. and Nordbotten, S. (2008). Prediction and imputation in ISEE: Tools for more efficient use of combined data sources. *Proceedings from UN/ECE Workshop Session on Statistical Editing in Vienna. Geneva 2008.*

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