Journal of Official Statistics, Vol. 12, No. 4, 1996, pp. 385–401

Neural Network Imputation Applied to the Norwegian 1990 Population Census Data

Svein Nordbotten[1]

When adequate administrative registers are available, population censuses can be carried out by supplementing the administrative data with data compiled from sample surveys. A problem is, however, to obtain reliable estimates from the survey data for small areas and groups. This article reports results from empirical experiments with neural network models for imputing individual values of survey attributes utilizing the available administrative data. The imputed values were used to prepare estimates of proportions for a population and for smaller subgroups of the population. The data used in the experiments were from the 1990 Population Census in Norway.

Key words: Population census; administrative registers; imputation; neural networks.

1. Introduction

1.1. Problem

The Norwegian 1990 Population Census was carried out by the Central Bureau of Statistics (SSB) based on *register data* obtained from administrative registers and supplemented by *survey data* compiled by mail from a sample of the population. Survey data statistics were estimated from the sample data. Collected survey attributes included both "demographic" and "housing" data, but only the demographic data were used in this study. The population was stratified by municipalities using sampling probabilities determined by the size of the population of each municipality. A few municipalities financed complete surveys, i.e., survey data were collected for all inhabitants in the municipality (Statistisk sentralbyrå 1992).

A set of rules to safeguard the level of accuracy of totals published and released for dissemination were adopted by SSB. According to these rules a number of estimated totals for smaller groups or areas within municipalities were considered too uncertain for general release. A question was raised whether an alternative approach existed which could produce more accurate estimates. It was proposed to explore the use of neural networks for producing alternative estimates.

Experimental research with Artificial Neural Network models (ANN) indicates that these might also be useful for imputing non-observed, individual survey attribute values

[1] Department of Information Science, University of Bergen, N-5020 Bergen, Norway.
Acknowledgement: This work was carried out as part of the SIS Statistical Information Systems project and under a cooperation contract with Statistics Norway. I wish to thank Ib Thomsen, Research Director at Statistics Norway for many useful discussions during the work on this project. I am also in debt to Joan C. Nordbotten and two anonymous referees who made constructive comments on earlier drafts.

based on the register attribute values for the same individual (Nordbotten 1996). An ANN is a network of interconnected, non-linear processing units for distributed, parallel processing of input values to obtain a set of output values. Individual survey attributes can be considered as dependent (output) variables related by a neural network to register attributes as independent (input) variables. The connection weights characterizing the "strength" of interconnections in an ANN correspond to the parameters in multivariate models. They can be obtained (estimated) from a sample of the population containing both survey and register data by training. Given individual register data for the population not included in the sample, the trained ANN can then be used to impute the unobserved individual survey data which in turn can be aggregated to estimates of totals, averages and proportions.

1.2. Outline of the investigation

A municipality, for which survey data were collected from all of its 17,326 inhabitants, was selected as a representative basis for empirical testing. For each person of this population, a record with values for survey and register attributes was available in an SSB file. The definition of attributes will be discussed below.

Test computations were carried out on alternative, random samples of different sizes from the SSB file to calibrate the size needed for the training sample. The tests indicated that a sampling probability of about 0.08 to 0.12 would give the best results, and a *training sample* of 1,845 records was drawn at random. This corresponds in size roughly to the sample which would have been used by SSB if the selected municipality had not ordered a complete survey. The remaining population, containing 15,481 records, was used as a *test sample*. Ten ANN models were specified and trained, one for each set of survey attributes specified in the next section. After training, the models were used to impute survey data for each person in the test sample. It should be recalled that for this particular municipality, observed data were also available. The imputed data were finally compared with the observed (target) survey data for evaluation of the imputation approach. The comparison was made both for the total test sample to make the evaluation of the ANN approach as well as for a smaller area within the municipality to study the quality of estimates based on imputed values in small subgroups.

2. Organization of Data

Each population record established by SSB consisted of a number of survey and register attribute values.The total record size was 475 bytes. Disregarding attributes not used in the present investigation, the record consisted of:

10 sets of survey attributes, and
46 sets of register attributes.

The attributes are listed in Appendices A and B.

The general SSB population files have one complete record for each individual. By drawing records at random from the SSB file for the selected municipality with a probability 0.1, a sample was established as a *training* file. The remaining records were kept in a *test* file.

We distinguished three types of attributes in the original SSB file and called them:

binary attributes,
categorical attributes,
numeric attributes.

A *binary attribute* had one data field that could be filled in or left blank in the population questionnaires and records. A set of binary attributes was used when the corresponding categories were not mutually exclusive, for example in the attribute set of cohabitation. A *categorical attribute* comprised a set of two or more mutually exclusive categories. Each category could therefore be represented in a single field by a specific integer, for example in the range 0 to 9, or the field could be left blank. A *numeric attribute* had a non-negative integer value range, but also for this type of attribute blank responses occurred.

The ANN models were constructed on entities differing from the SSB attributes. The SSB file records therefore required reformating. To distinguish the model entities from the SSB record *attributes*, the reformated entities for ANN were referred to as *variables*. Each binary attribute in the SSB format was represented in a single field in the ANN format with a '0' variable value if the attribute was "blank" and with a "1" for any other value in the SSB recording. An SSB categorical attribute was expanded in the ANN format to a set of binary variables, one for each category starting always with a 0 category even if it was unused. In the reformated form, the SSB category value k was represented by a "1" in the $(k+1)$th ANN binary variable leaving all other variables of the set equal to 0. Blank categorical attribute responses were represented by "0"s in all binary variables of the corresponding set. The sum of values for such a set of binary variables must therefore either be "1" if a category symbol was recorded or "0" if the category was left blank. The numeric variable values in the SSB records were kept unchanged except for blank fields. Blank numeric attribute fields in the SSB file were changed to 0 for the corresponding variables in the ANN format.

The reformating resulted in

49 survey variables, and
97 register variables

for each individual in the ANN format.

For each of the 10 sets of survey variables, a separate imputation model was established. The 10 models aimed at simultaneous imputation of 9, 2, 3, 2, 6, 7, 3, 5, 6, and 6 survey variables, respectively. Appendix A lists which variables were included in the different models. The 97 register variables were used as independent variables in all models.

3. Imputation Models

A generic imputation model was designed as a feed-forward neural network model with a single layer of hidden units. In a feed-forward neural network, no feedback connections exist among its units. Hidden units are processing units which do not receive exogenous input or deliver final output, but receive input from and pass on output to other units in the network. This type of model had been used in previous experiments for editing survey data

with partial attribute non-response and errors to which the reader is referred for further explanations of neural network models (Nordbotten 1995). Models with more than one layer with hidden units can be specified. Experience and tests indicated that for our problem, no gain could be expected from models with more than one layer of hidden units. More complicated neural networks, such as recurrent networks might also be interesting to study as tools for imputation.

For the purpose of the present article, the generic imputation model can be regarded as a set of non-parametric, non-linear, multivariate regression equations

$$y_i = f\left(b_i + \sum_{j}^{25} d_{ij} * f\left(a_k + \sum_{k}^{96} c_{jk} * x_k\right)\right) \text{ for } i = 1..N \tag{1}$$

where y_i and x_k are dependent and independent variables, respectively. The N represents the number of dependent variables included in the model. As explained in the previous section, the number of dependent variables varied in the ten models. The a's, b's, c's, and d's are all parameters which must be estimated. The f symbolizes the non-linear function

$$f = 1/(1 + e^{-z}) \tag{2}$$

The z represents the expressions appearing as arguments in the parentheses following the f symbols in (1). Alternative functions might have been used, but this sigmoid function has been used in a large number of applications and is considered to work well.

The size of M was determined by experiments with 10, 15, 25, 40, and 60 units in the layer of hidden units. A value of $M = 25$ gave the best results for most models, and the loss in quality was minimal by using 25 hidden units also for the few models for which a better alternative existed. The set of dependent variables in the individual models corresponded to those associated with separate questions in the survey questionnaire. Experiments with a single imputation model imputing all 49 variables simultaneously were also carried out. A single model may produce results which compete in quality with the 10 models discussed. Such a large model seems to require both larger training samples and more training cycles than the 10 smaller models used.

4. Computing Resources

Two computer programs were used:

1. TRANS

TRANS is a C++ program developed to read and inspect the SSB population file, to establish and maintain meta (attribute format and description) files, to sample from SSB files, to convert the SSB records to the ANN format, and to produce aggregates and statistics for comparing the computed data files produced by the networks with the target files.

2. BrainMaker Professional

BrainMaker Professional is a commercial neural network program developed by California Scientific Software for computers with Intel processors. It was used for training and testing the models.

All processing was done on a PC with a Pentium processor, 24 Mbyte Ram and a 1 Gbyte harddisk.

5. Training Imputation Models

The number of weights to be estimated in a model of the type used is

$$W = K^*(1 + M) + N^*(1 + K) \tag{3}$$

where M, K, and N are the number of independent variables, hidden units and dependent variables. As an example, 2,684 connection values (weights) must be estimated in the imputation model for cohabitation with 97 independent variables, 25 hidden neurons and 9 dependent variables. Because of the non-linear nature of the models for which no direct estimator exists, an iterative estimation of the weights known as the Back Propagation (BP) was used (Rumelhart 1986). This procedure is referred to as training the models.

The BP procedure uses the differences between the output variable values calculated (the computed values) and the target variable values (the observed values) from the training set as a basis for adjusting weight values in each iteration. The training process can be monitored by watching the computed *mean squared error* (MSE) of all differences after each training cycle. Optimal weight values have been obtained when all the differences between imputed values and observed values become 0 for all output variables. In actual applications, the iteration has to be terminated when no further progress in diminishing the differences between estimated and target values can be observed. The set of weight or parameter values found when the training is terminated may not represent the best or optimal set, but a set representing a so-called local minimum in the MSE surface. Still, if not the best, the trained network will in many applications be useful.

The large number of weights in the models indicates a danger for overfitting. Overfitting is adjusting too much to the training sample with a risk for losing the model's ability to generalize and making useful imputations for the population outside the sample. Several procedures for avoiding the problem of overfitting have been proposed, and also tried in connection with this investigation. The approach used was early termination of the training process. It was implemented by running the learning process twice subject to identical conditions. During the first training, MSE development was recorded. The training cycle at which MSE did not decrease any more and started to get unstable was identified. In the second final run, training was terminated before this point.

The BP procedure presented some options which were considered. Among the most important were the *learning rate*, the *momentum constant* and the *initial weights*. The learning rate determines how fast the weight connections should be adjusted to the training sample set. A large learning rate reduces the required number of adjustment iterations subject to the risk of passing the best adjustment. A small value of the learning rate reduces the risk for missing the best adjustment, but requires a higher number of iterations. After some trials, the option selected was a variable learning rate starting with 0.25 and 0.50 for the first and second weight matrix, respectively, and decreasing linearly during the training to 0.1 when the model is completely trained. This strategy should support relatively fast training when the deviations between computed and target values were large, while relatively smaller changes were made when the differences became smaller to avoid correction which made the weights bypass the minimum point searched.

The momentum constant determines how much of the difference between calculated output variables and target variables the current iteration cycle should use for adjustments and how much the difference should be permitted to influence future iterations. The mechanism can be compared to moving averages. Experience indicates that the use of momentum results in a smoother and safer adjustment. The value of this parameter was set to 0.9 according to experience reported in many studies and publications (California Scientific Software 1993).

The models must be initiated with some random weight values. The choice of the range for the initial weight values may influence the speed by which the weights converge. The initial weights in the present study were randomly drawn from a rectangular distribution with range -0.2 to $+0.2$.

The *coefficient of determination* (the squared correlation coefficient) was used as a measure of the success of training. It is defined as

$$R^2 = [N^*\Sigma o^* t - (\Sigma o)^*(\Sigma t)]^2/\{[N^*\Sigma(o^2) - (\Sigma o)^2]^*[N^*\Sigma(t^2) - (\Sigma t)^2]\} \quad (4)$$

where N is the number of individuals recorded in the training sample, o is the imputed survey value and t is the recorded, true, survey variable value; R^2 indicates on average how well the trained model reproduces the target survey values of the training set given the register values. The coefficient will have a value in the interval $0 <= R^2 <= 1$. Value 1 indicates that the model reproduces target values well while 0 indicates that the recalls of target values are bad.

After training was stopped, the coefficients of determination for all variables were computed. Table 1, column 1, presents values for R^2 for all the survey variables included. For cohabitation as an example, the coefficient of determination varied from the high 0.9419 for recalling cohabitation with spouse to low 0.4250 for recalling living with inlaws. When using this model, quite reliable results could be expected for recalling whether the current persons live with a spouse, while recalls of the three last output variables would probably give results with more frequent errors.

The coefficient of determination indicates how well a model has been trained to reproduce individual survey variables in the training file when presented with the set of register data for the individual. However, the coefficient of determination may not be a reliable indicator for the accuracy of imputing attribute values for the non-observed population.

6. Testing Trained Models

A main objective of this study was to investigate whether the trained ANN models can contribute to improved statistical estimates for a population compared with traditional estimates when both are based on the same population sample. Estimates for small subgroups were of particular interest.

6.1. Coefficients of determination

Traditional sample survey methodology provides objective measures for evaluating the precision of the population estimates such as the standard error of estimated means, proportions, totals, etc. It would be of considerable interest to find similar accuracy measures for estimates based on imputation. The coefficients of determination were

used above to measure the models' capability to reproduce the individual survey values of the training sample. Can these coefficients give guidance to the users about the risk of using imputed values for estimating population aggregates? One way to obtain an answer is to compute the corresponding coefficients for the test sample, i.e., from the imputed and true attribute values for individuals of the population which was not comprised by the training sample.

The test sample available comprised the 15,481 records. The records contained both *register* variables and *target* survey variables. For each record, imputed values of the survey variables were computed by means of the trained network models and the available register variables. The results were recorded in an *output* file for imputed survey variables.

The coefficients of determination for the test sample were computed for each survey variable from the target value and imputed value files, and are shown in the last column of Table 1. These coefficients indicate how well the individual imputed values correspond to the target survey values.

Comparing the coefficients of determination from the training sample and from the test sample indicates, as suspected, that the former overestimate significantly the imputation capability of the models. The coefficients from the test sample indicate, however, that the correlations between imputed and target values in the test sample are still high for a number of the output variables. The imputation models may therefore be expected to provide useful individual survey variable values for the population not included in the training sample.

6.2. *Estimates and true proportions for the population not included in the sample*

Simple unbiased estimates of proportions were used as benchmarks for evaluating the imputation in our study. In traditional surveys, the sampling errors are frequently used measures for the precision of proportion estimates. The estimator

$$p = f/n \tag{5}$$

where f is the absolute frequency observed in the sample and n is the size of the sample, gives a simple, unbiased estimate of the proportion P in a finite population with N units. The results from the use of this estimator will be referred to as the simple unbiased *estimates* of proportions in contrast to the *imputed* estimates obtained from aggregation of individual imputed values obtained by the network models.

The relative variance of a simple, unbiased estimate is

$$S^2/P^2 = (1 - n/N)^*(1 - P)/n^*P \tag{6}$$

Table A. Standard deviations for estimates of P

P	S	S/P
0.01	0.002	0.219
0.05	0.005	0.096
0.10	0.007	0.066
0.25	0.010	0.038
0.50	0.011	0.022

Table 1. Coefficients of determination for training and testing

	Coefficients of determination	
	Training sample	Test sample
With whom are you living?		
1. Nobody	0.7706	0.4628
2. Spouse	0.9419	0.9060
3. Cohabitant	0.8190	0.6209
4. Children	0.8187	0.5985
5. Parents	0.8975	0.6157
6. Siblings	0.8961	0.4553
7. Inlaws	0.4250	0.0085
8. Grandpar.,-child.	0.5620	0.0135
9. Other people	0.5574	0.0129
Did you have paid work for 100 hours or more?		
1. Yes	0.9821	0.7690
2. No	0.9782	0.7127
Type of employment?		
1. Employed	0.9668	0.6146
2. Self-employed	0.8573	0.0720
3. Family employee	0.8458	0.0065
Income from work in week Oct. 27–Nov. 2, 1990?		
1. Yes	0.9966	0.9952
2. No	0.8830	0.2123
Usual weekly work-time?		
1. 1–9 hours	0.6813	0.0998
2. 10–19 hours	0.7176	0.1282
3. 20–29 hours	0.8236	0.1798
4. 30–34 hours	0.4921	0.0272
5. 35–39 hours	0.8034	0.3527
6. 40 hours or more	0.7482	0.0979
Paid working hours during week Oct. 27–Nov. 2, 1990?		
1. 1–9 hours	0.6390	0.0711
2. 10–19 hours	0.6401	0.1321
3. 20–29 hours	0.6588	0.1040
4. 30–34 hours	0.3948	0.0106
5. 35–39 hours	0.7182	0.2526
6. 40 hours or more	0.7045	0.1324
7. No paid work	0.9404	0.8762
Working location during week Oct. 27–Nov. 2, 1990?		
1. Usual place	0.7664	0.6071
2. A different place	0.2760	0.0738
3. Different places	0.2712	0.0097

Table 1. Continued

	Coefficients of determination	
	Training sample	Test sample
Travel trips during week Oct. 27–Nov. 2, 1990?		
1. None, at home	0.7309	0.0499
2. Not at home	0.5903	0.0114
3. One trip	0.5435	0.0056
4. 2–3 trips	0.7196	0.0720
5. 4 or more trips	0.8600	0.3682
Time for one way travel to work during week Oct. 27–Nov. 2, 1990?		
1. < 15 min.	0.6726	0.1566
2. 15–29 min.	0.5361	0.0277
3. 30–44 min.	0.4165	0.0114
4. 45–59 min.	0.4152	0.0067
5. 60–89 min.	0.0431	0.0030
6. 90 min. or more	0.1621	0.0008
Means of transportation to work during week Oct. 27–Nov. 2, 1990?		
1. Car	0.7943	0.3676
2. Bus	0.5010	0.0025
3. Train	0.6096	0.0017
4. Boat	0.9908	0.2114
5. Bicycle	0.6096	0.0467
6. Other	0.4409	0.0006

where S is the standard deviation. In our investigation, $(1 - n/N)$ is approximately 0.9 for $n = 1,845$ and $N = 17,326$. Table A gives the approximate values for the standard deviations and the relative standard deviations for estimates p for a few values of a true proportion P.

With reasonably sized population and sample, the usual interpretation of these figures is that if p is a simple, unbiased estimate of the proportion P, then the confidence interval $(p + / - 2^*S)$ will include P with a probability about 0.95. SSB considered an estimate suited for publication if its relative standard deviation was less than or equal to 0.3. For sample and population sizes of about 1,800 and 18,000, the relative standard deviation would be less than or equal to 0.3 for all proportions $0.008 < P < 0.992$.

To facilitate comparison, all estimates and target values refer to the proportions of the test sample and not to the complete population. The estimator used for computing the estimates based on imputed values was therefore

$$p' = f'/(N - n) \tag{7}$$

were f' is the absolute frequency imputed in the test sample.

The proportion estimates based on imputed values for the unobserved population of 15,481 individuals cannot be expected to give results with reliability higher than the traditional unbiased estimates because of the sizes of the population and the sample.

Table 2. Estimates for municipality and census tract. Per cent of population

	Municipality					Census tract				
	Imputed (1)	Unbiased (2)	Target (3)	(1)–(3) (4)	(2)–(3) (5)	Imputed (6)	Unbiased (7)	Target (8)	(6)–(8) (9)	(7)–(8) (10)
With whom are you living?										
1. Nobody	9	10	9	0	1	9	16	13	−4	3
2. Spouse	40	38	40	0	−2	50	33	49	1	−16
3. Cohabitant	8	8	8	0	0	8	27	8	0	19
4. Children	30	29	30	0	−1	35	27	29	6	−2
5. Parents	12	13	13	−1	0	11	27	11	0	16
6. Siblings	8	8	8	0	0	5	16	4	1	12
7. Inlaws	0	1	0	0	1	2	11	2	0	9
8. Grandpar.,-child.	0	0	0	0	0	1	11	2	−1	9
9. Other people	1	2	2	−1	0	1	11	4	−3	7
Did you have paid work for 100 hours or more?										
1. Yes	50	51	50	0	1	54	50	54	0	−4
2. No	27	27	28	−1	−1	25	33	28	−3	5
Type of employment?										
1. Employed	45	46	45	0	1	49	44	49	0	−5
2. Self-employed	3	3	4	−1	−1	5	11	5	0	6
3. Family employee	0	1	0	0	1	0	5	0	0	5
Income from work in week Oct. 27–Nov. 2, 1990?										
1. Yes	43	46	43	0	3	49	38	50	−1	−12
2. No	6	5	6	0	−1	8	16	4	4	12
Usual weekly work-time?										
1. 1–9 hours	1	2	2	−1	0	0	5	4	−4	1

Table 2. Continued

	Municipality					Census tract				
	Imputed (1)	Unbiased (2)	Target (3)	(1)–(3) (4)	(2)–(3) (5)	Imputed (6)	Unbiased (7)	Target (8)	(6)–(8) (9)	(7)–(8) (10)
2. 10–19 hours	4	4	4	0	0	5	5	5	0	0
3. 20–29 hours	4	5	4	0	1	4	5	4	0	1
4. 30–34 hours	2	3	2	0	1	0	5	3	−3	2
5. 35–39 hours	21	20	20	1	0	25	27	27	−2	0
6. 40 hours or more	9	10	9	0	1	6	16	7	−1	9
Paid working hours during week Oct. 27–Nov. 2, 1990?										
1. 1–9 hours	1	2	2	−1	0	1	5	3	−2	2
2. 10–19 hours	3	4	4	−1	0	3	5	6	−3	−1
3. 20–29 hours	3	5	5	−2	0	5	5	6	−1	−1
4. 30–34 hours	1	3	2	−1	1	0	5	2	−2	3
5. 35–29 hours	19	17	17	2	0	22	27	24	−2	3
6. 40 hours or more	12	11	10	2	1	8	16	9	−1	7
7. No paid work	0	0	0	0	0	0	5	0	0	5
Working location during week Oct. 27–Nov. 2, 1990?										
1. Usual place	35	38	36	−1	2	43	33	42	1	−9
2. A different place	2	3	3	−1	0	0	5	4	−4	1
3. Different places	4	3	2	2	1	5	11	4	1	7
Travel trips during week Oct. 27–Nov. 2, 1990?										
1. None, at home	3	4	4	−1	0	4	5	11	−7	−6
2. Not at home	2	1	1	1	0	1	5	3	−2	2
3. One trip	0	1	1	−1	0	0	5	2	−2	3
4. 2–3 trips	4	5	4	0	1	6	11	6	0	5
5. 4 or more trips	26	29	27	−1	2	29	26	26	3	0

Table 2. Continued

	Municipality					Census tract				
	Imputed (1)	Unbiased (2)	Target (3)	(1)–(3) (4)	(2)–(3) (5)	Imputed (6)	Unbiased (7)	Target (8)	(6)–(8) (9)	(7)–(8) (10)
Time for one way travel to work during week Oct. 27–Nov. 2, 1990?										
1. < 15 min.	15	20	19	−4	1	14	16	22	−8	−6
2. 15–29 min.	5	7	7	−2	0	4	11	7	−3	4
3. 30–44 min.	1	3	3	−2	0	0	16	1	−1	15
4. 45–59 min.	2	3	2	0	1	3	5	2	1	3
5. 60–89 min.	0	0	0	0	0	0	5	2	−2	3
6. 90 min. or more	0	0	0	0	0	0	5	1	−1	4
Means of transporation to work during week Oct. 27–Nov. 2, 1990?										
1. Car	25	29	27	−2	2	29	38	29	0	9
2. Bus	0	1	0	0	1	2	11	1	1	10
3. Train	0	1	0	0	1	3	11	2	1	9
4. Boat	0	0	0	0	0	1	11	1	0	10
5. Bicycle	7	7	6	1	1	8	11	7	1	4
6. Other	0	1	0	0	1	1	11	1	0	10

This is also demonstrated by the figures in Table 2, columns 1-5, which show the estimated proportions from imputed survey attribute values, the corresponding simple, unbiased estimates of the proportions from the training sample and the true target proportions from the test sample.

The absolute deviations of estimates and true proportions are shown in columns 4 and 5. The simple, unbiased estimates seem to be slightly closer to the true proportions than the estimates from imputed values, in particular for the variables in the lower part of Table 2. The most serious deviation of the estimates based on imputed values is for the proportion of the population that has less than 15 minutes' travel to work. This proportion based on imputed values underestimated the true proportion of 13% by 4 percentage points. On the other hand, the simple, unbiased estimates overestimated the true proportion of 43 % of the population answering they had income from work in the selected week by 3 percentage points. The remaining deviations were 1 or 2 percentage points for both types of estimates.

6.3. Estimates and true proportions for small areas

The figures of Table A assumed a population of 18,000 and a sample of 1,800. For a subgroup of 180 and a sample of 18, the standard deviations for the proportions of the table will increase by a factor of about 10. Only simple, unbiased estimates of proportions between 0.4 and 0.6 would then be expected to satisfy the SSB rules for publication and be released. Objection can also be raised to the basic assumptions for reliability considerations based on such small samples. For such small subgroups, a number of estimates must therefore be suppressed.

To investigate how the imputed values compare with the true values in a small subgroup, a census tract with only 142 people was selected. This census tract was represented with 16 people in the training sample. For each of the 128 individuals not included in the sample, imputed values were computed from the ANN models. For comparison, simple, unbiased estimates were also computed for all variables based on the records of the 16 individuals in the sample. The results are presented in Table 2, columns 6–10, together with the true proportions of the 128 individuals of the census tract.

The figures in column 7 of Table 2 confirm that simple, unbiased estimates of proportions for the population in the census tract are of poor quality. The imputed proportions shown in column 6 indicate, on the other hand, a higher quality. A comparison of the absolute deviations of imputed and estimated proportions from the true proportions shows that on average the estimated proportions deviated from the true proportions with more than 2.5 the deviations of the imputed proportions. The greatest deviation for imputed proportions from true proportions was also for this area the proportion of the population that had less than 15 minutes travel to work in the selected week. The true proportion of 22% was underestimated by 8 percentage points by the imputed proportion. Out of the 49 proportions imputed, 15 were correct and another 15 deviated by +/− 1 from the true proportion in the population outside the sample of the census tract. The corresponding figures for the unbiased estimates were 3 correct estimates and 5 estimates which deviated by 1 percentage point.

As pointed out above, the estimates and the true proportions refer to the test sample

only. If the observed values for the sample population were added to obtain the proportions for the *total* population, the observed accuracy of proportions based on imputed values would have been even higher while the unbiased estimates would be unchanged.

In a real application, no true target observations would exist for the population not observed in the sample. It would be crucial for both providers as well as for users of the estimates to obtain some indication of the quality of the imputed proportions. The prediction risk connected to the use of ANN models, i.e., the uncertainty of imputed values, has recently been investigated by several authors (Moody 1993). Methods for predicting the uncertainty of imputed values are extremely important for practical use of imputation models. Methods for this purpose are being explored (Nordbotten 1996).

7. Summary and Conclusions

1. The objective of this study was to investigate neural network models for estimating proportions in a population for which register data were also available. A random sample of about 10% from a population of about 18,000 for which both survey and administrative data were available was used to develop and evaluate imputation models based on neural networks.
2. The 10 imputation models used were feed-forward neural networks with a single hidden layer trained by the Back Propagation algorithm. Each model was designed to impute individual survey attribute values for 2-9 variables simultaneously based on administrative data for the individual.
3. Training each model required up to 2,000 iterative cycles through the sample, or one to three hours computing time for the final training of each model. The training was terminated when the development of the MSE indicated the possibility of overfitting.
4. On the other hand, imputing attribute values for 15,481 individuals not included in the sample required about 30 seconds from each trained imputation model.
5. The learning results varied by attribute and category and are reflected in varying values of the computed coefficients of determination. Improved learning may be obtained by a careful limitation of register variables used to those which were well correlated to the survey variables.
6. The training sample included 1,845 records. Records identical or very similar could be excluded from the training process to speed up the learning since the purpose was to learn recognized patterns, not their relative importance.
7. There were many records in the SSB file with blank attribute fields. Better learning results could probably be obtained if the sample records had been preprocessed to eliminate the ambiguity of some blank fields before the training of the imputation models.
8. The test experiments carried out indicated that proportion estimates aggregated from individual imputed values competed in reliability with traditional unbiased estimates for large groups. The proportion estimates based on imputed values seemed to be superior compared with simple, unbiased estimates when applied to population in small areas or groups.
9. Practical use of imputation models will require a method for predicting the accuracy of estimates based on imputed values.

Appendix A. Survey data

Attribute name	Variables	Type
Model 1		
Single habitant	1	bin
Living with spouse	1	bin
Living with cohabitant	1	bin
Living with daughter/son	1	bin
Living with mother/father	1	bin
Living with siblings	1	bin
Living with inlaws	1	bin
Living with grandparents/grandchildren	1	bin
Other	1	bin
Model 2		
Income from work	2	cat
Model 3		
Type of work association	3	cat
Model 4		
Income in week	2	cat
Model 5		
Number of usual weekly work-time	6	cat
Model 6		
Number of work hours in week	7	cat
Model 7		
Work location in week	3	cat
Model 8		
Number of trips to work in week	5	cat
Model 9		
Normal time required for each trip in week	6	cat
Model 10		
Car to work in week	1	bin
Bus to work in week	1	bin
Train to work in week	1	bin
Boat to work in week	1	bin
Bicycle, walk to work in week	1	bin
Other means of transport	1	bin

Appendix B. Register data

Attribute name	Variables	Type
Education A	10	cat
Household type	2	cat
Age	1	num
Sex	2	cat
Marital status	5	cat
Labor activity	3	cat

Head of family	2	cat
Type of family	5	cat
Family reference person	2	cat
No. family members	1	num
Persons >=17 years in family	1	num
Persons>17 years in family	1	num
No. family persons labor act. in week	1	num
No. parents labor active in week	1	num
Household reference person	2	cat
Age of youngest household member	1	num
No. of families in household	1	num
No. of persons in household	1	num
Persons>=17 years in household	1	num
Persons>17 years in household	1	num
No.household members labor act. in week	1	num
No. full time active in week	1	num
No. part time active in week	1	num
Employment status	1	bin
Labor force status	4	cat
Student	3	cat
Cohabitant code	6	cat
Labor activity status of spouse	4	cat
No. of children living at home	1	num
Age of youngest child	1	num
Age of oldest person in household	1	num
No. of persons with income in family	1	num
Highest income in household	2	cat
No. of persons with income in household	1	num
Highest income in household (1980)	3	cat
No. with income in household (1980)	1	num
Driver's licence	3	cat
Education B	10	cat
Income 1990	1	num
Taxable income 1990	1	num
Disposable income 1990	1	num
Family income 1990	1	num
Household income 1990	1	num
Disposable income 1980	1	num
Family income 1980	1	num
Household income 1980	1	num

8. References

California Scientific Software (1993). BrainMaker Professional, User's Guide and Reference Manual, 4th Edition. Nevada City, CA.: California Scientific Software.

Moody, J.E. (1993). Prediction Risk and Architecture Selection for Neural Networks. In Cherkassy, V., Friedman, J.H. and Wechsler, H. (eds): From Statistics to Neural Networks, Theory and Pattern Recognition Applications. Berlin: Springer Verlag.

Nordbotten, S. (1995). Editing Statistical Records by Neural Networks. Journal of Official Statistics, 11, 391-411.

Nordbotten, S. (1996). Predicting the Accuracy of Imputed Proportions. Department of Information Science, University of Bergen, Bergen, Norway.

Rumelhart, D.E. and McClelland, J.L. (1986). Parallel Distributed Processing - Explorations in Microstructure of Cognition, Vol.I. Foundation. Cambridge, Mass: MIT Press.

Statistisk sentralbyrå (1992): Folke-og boligtellingen 1990. Oslo, Norway. (In Norwegian).

Received March 1996
Revised January 1997