Published in Computational Statistics & Data Analysis 27 (1998) pp 291-309.

ESTIMATING POPULATION PROPORTIONS FROM IMPUTED DATA

Svein Nordbotten
Department of Information Science
University of Bergen
N-5020 Bergen, Norway

Abstract: *Imputation estimates based on imputed values obtained from neural network models used in an 'impute first-aggregate next' approach, have been computed from Norwegian population census and administrative register data. The imputation estimates were compared with simple unbiased estimates obtained by the traditional 'aggregate first - estimate next' approach and found to be superior for estimating proportions in small sub-groups. Predictors for predicting the accuracy of such imputation estimates were proposed. Results are promising for estimating small subgroup or area proportions.*

Keywords: *Imputation estimates, non-linear imputation models, neural networks, estimate accuracy prediction .*

Acknowledgement

This research has been carried out under a cooperation contract with the Norwegian Central Bureau of Statistics, and as part of the Statistical Information System (SIS) project at the Department of Information Science, University of Bergen. I am grateful for a number of suggestions offered by Professor Ib Thomsen, Research Director, Central Bureau of Statistics, Oslo.

1. Introduction

In Norway, an extensive amount of administrative register data for each inhabitant is available to the Central Bureau of Statistics (CBS). These were used to prepare statistics for the Population Census in 1990. In addition to the register based statistics, estimates of totals, averages, and proportions for other non-register attributes were computed from a population survey sample**.** For smaller subgroups and areas, the sample based estimates frequently failed to satisfy established accuracy criteria for publication. More sophisticated estimators could be used, for example ratio and regression estimators, taking advantage of possible functional relationships between survey variables and register variables.

The traditional estimation approach would be to *aggregate first- estimate next*, i.e. aggregate values from a sample would be used for estimating the survey attribute statistics. In the case of more advanced estimators, aggregates for the register variables from both the sample and the remaining population would be used to improve the estimates of survey attribute statistics.

However, if the nature of an assumed relationship between a survey variable and register variables is non-linear, the use of aggregates as arguments in the non-linear estimator will give estimates different from those obtained if individual imputations are computed first by means of non-linear functional relationship and then aggregated to the required statistics. The latter approach will be called the *impute first-aggregate next* approach.

Assumed relationships between dependent survey variables and independent register variables can be estimated in different ways, for example by means of methods from the statistical theory of regression or by means iterative methods from the theory of parallel distributed processing (Rumelhart 1986).

The aim of this study is to investigate if results, particularly for small subpopulations, from a population census could be made more useful by applying an *impute first - aggregate next* approach based on imputation models with parameters computed by means of an algorithm for iterative approximation.

2. Experiments

Imputation models

The paradigm of parallel distributed processing, or neural networks, have been intensively studied and improved during the last decade. The relationship between neural network theory and the theory of regression has also been highlighted [Cheng and Titterington 1994].

The possibility of using models based on Artificial Neural Networks (ANN) for imputing survey variable values for each individual not included in the sample to obtain individual data for the whole population, has been studied and reported in a recent study [Nordbotten 1996] on which the present investigation is based.

Two alternative sets of imputation models were studied, and are referred to in this paper as the *linear* and the *non-linear* imputation models. All imputation models investigated can be considered as functions imputing the individual values of unknown survey variables based on known register variables as arguments. In total, 96 survey variable functions were developed in each of the two sets of models. All functions had 96 register variables as arguments.

The linear models comprise functions of the form

$$y_{(j)}' = 1/\,(1+exp\text{-}[?w_{0k}\,?\,??\,_{i}^{96}w_{ik}*x_{(i)}\,]) \qquad\qquad for\ j=1,..,49.$$

where $x_{(1)},..,x_{(96)}$ denote the known register values and $y_{(1)}',.., y_{(49)}'$ are model predictions of the survey values $y_{(1)},.., y_{(49)}$ for an individual. The *w*-s are the parameters or weights to be computed. The functions are identical to logit regression functions [Cheng-Ming Kuan 1994]. These models are also known as simple feed-forward network models with sigmoid transfer functions[1]. The properties of these models have been extensively studied, and their limitations were early pointed out [Minsky 1969].

Most of the objections can be escaped by non-linear imputation models

$$y_{(j)}'' = 1/\,(1+exp\text{-}[w_{0j}+?\,_{k}^{25}w_{kj}*(1/(1+exp\text{-}[?w_{0k}\,?\,??\,_{i}^{96}w_{ik}*x_{(i)}\,])]) \qquad for\ \ j=1,..,49.$$

These models are known as feed-forward networks with a layer of *K* hidden neurons. The hidden neurons are also characterized by sigmoid transfer functions. The hidden neurons correspond to latent variables in the statistical theory.

In this study, the parameters were computed from a sample by means of an iterative algorithm called Backpropagation [Rumelhart 1986]. In the case of a linear model function, experiments carried out indicate that this algorithm seems to give parameter values approaching those computed by means of the least squares method.

Following the computation of the parameters, the models were used to impute individual survey variable values for unobserved individuals which were finally aggregated together with

[1] The linear models are therefore not strictly linear.

the observed values to estimates for the population proportions.

Estimators and accuracy predictors

Three different estimators were used in the experiment. First, ordinary simple, unbiased estimates for the population proportions were computed. These estimates were referred to as the *unbiased estimates*. The second set of estimates used imputations of individual survey variable values from a linear imputation model. As pointed out, these estimates can be considered as approximations to what would have been obtained if ordinary multiple regression had been used for estimating the imputation models. These estimates are called the *linear imputation estimates*. Finally, the *non-linear imputation estimates* were computed using the individual imputation values from the non-linear imputation models.

The *accuracy* of an estimate is defined as the deviation of its value from its target. An *accuracy predictor* is a method for determining margins for the accuracy of an estimate subject to a specified risk of error. Accuracy predictions are wanted by producers of statistics for quality declaration of the estimates and by users for evaluating the appropriateness of statistics for their specific purposes. The standard error estimator for an unbiased estimate from a sample of observations, is an example of a well known method used for accuracy prediction. Without a reliable predictor for the accuracy of imputation estimates, these estimates would be of limited interest.

Cross valuation methods have been proposed for computing accuracy predictions for the individual imputations when the number of observations are small and are all needed for training [Moody 1993]. In the present application, the population is relatively large and a sample of several thousand individual observations is not prohibitive. An independent sample not overlapping the training sample, was a reasonable and straight forward solution applied for predicting the accuracy margins for the imputation estimates.

The population was assumed divided randomly into three parts. *Sample 1* was used for computing the parameters of the imputation models and *Sample 2* for developing the accuracy predictors. The members of each of these two samples were assumed to be surveyed. *Sample 3* represented the remaining population for which individual imputations were required.

The estimators and accuracy predictors used were:

a) The *unbiased estimator* P_Y of a population proportion P based on observations in *Samples 1* and *2*

$$P_X = (?\ y_1 + ?\ y_2)/(n_1+n_2)$$

where the subscript numbers refer to the corresponding samples.

The standard error of P_Y

$$S_Y=s*?\ \overline{(1-f)/\ (n_1+n_2)}$$

was used as an predictor of accuracy for this estimate where f is the sampling fraction $(n_1+n_2)/N$. The standard deviation s of y was estimated from *Sample 1* and *Sample 2* as

$$s=?\ \overline{(?\ (y_1-\ddot{y}_1)^2+?\ (y_2-\ddot{y}_2)^2)/(n_1+n_2-1)}$$

where $\ddot{y}_1$ and $\ddot{y}_2$ are then means of the variables y in *Sample 1* and in *Sample 2*. The sampling error of s was ignored.

The accuracy of P_Y will decrease with decreasing sample size. The estimator P_Y is not likely to be useful for small subpopulations.

b) The *linear imputation estimator* P_L of P includes three steps. First, the coefficients w_0 and w_{ij} of the linear imputation models are computed from *Sample 1*. Then individual imputations y_3' for all members of *Sample 3* are computed by means of the imputation models. Finally, the individual observed values of *Sample 1* and *Sample 2* and the imputed values for *Sample 3* are aggregated to the imputation estimate

$$P_L=(?\ y_1+?\ y_2+?\ y_3')/N$$

Resampling of the training sample will normally give different P_L estimates.

The accuracy of P_L is determined by two factors, the training Sample 1 and the individual imputation errors

$$z_3=y_3'-y_3$$

which are due to the imperfections of the imputation models.

An accuracy predictor for P_L estimates can be obtained by rewriting the estimator

$$P_L=??\ y_1+?\ y_?\,?\ ?\ y_3)/N+?\ z_3/N$$

For any estimate P_L based on imputations from a given instance of the training sample, say *Sample 1*, we can predict the accuracy by

$$S_L=RMSE_L*?\ \overline{(1-f)/\ N}$$

where *RMSE* is the root mean square error of z in the population expressed by

$$RMSE_L=?\ \overline{?\ z_3/N}$$

Since the $RMSE_L$ is measuring the deviations of the imputed values from their corresponding target values, errors due to the training sample will also be included. *RMSE* was estimated by computing imputed values y_2' for the individuals of *Sample 2* and use the imputation errrors z_2 in

$$rmse=?\ ?\ \overline{z_2^2/(n_2\text{-}1)}$$

The sampling error of the estimate $rmse_2$ was ignored.

The S_{Y} predictor has the root of *n* as a denominator, while the S_L predictor has the root of *N* as denominator. The ratio *n/N=f* will always be equal to or less than 1. Only if

$$s<rmse*?\ \overline{f}\ ,$$

will the P_Y *e*stimates thus be preferable to P_L estimates.

c) The *imputation estimator* P_E of *P,* is similar to P_{L} , but with the non-linear imputation models substituting the linear models used for P_{L}

The P_E estimator is defined by

$$P_E=??\ y_1+?\ y_?\,?\ ?\ y_3''\ )/N$$

The accuracy metric S_E for this P_E estimate is computed as

$$S_E=RMSE_E*?\ \overline{(1\text{-}f)/N}$$

where $RMSE_E$ is the root mean square error of the *z*-variables when derived from the non-linear models. $RMSE_E$ is estimated from deviations between the individual imputed values by the non-linear models and the corresponding targets in *Sample 2.*

Assuming that the estimates discussed can be considered as events from normal distributions, it can be expected that for about two of three estimates the absolute deviations from their target value are equal or less than their accuracy predictions.

Data

Data from a municipality population of *17,326* individuals were used for the experiments. This particular municipality had paid CBS to have a *100%* survey. Both survey and register variable data were thus available for all inhabitants, and well suited for experimentation. To simulate the normal situation for which survey values were collected from a sample of the population, a random survey sample of *2,007* individuals was drawn. From this sample, a subsample of *1,845* records of survey and corresponding register values, *Sample 1*, was randomly extracted. The records for the remaining *2,007-1,845=162* individuals , *Sample 2,* were used for computing the accuracy predictors. The rest, 15,319 individuals, was called *Sample 3* for which imputations were required.

The individual survey variable values for all individuals of *Sample 1* and *Sample 2* were first aggregated and then blown up to traditional unbiased estimates of the population proportions.

Ten linear and *10* non-linear imputation models were developed. Each model could simultaneously impute values for *2* to *9* survey variables, in total *49* variables, based on the values of *96* register variables. The *49* survey variables were variables observed particularly for the population census. The *96* register variables were among a larger set of register variables also used in the census processing. They were selected after a number of preliminary experiments.

The linear models included from *194* to *873* parameters to be computed, while the non-linear models had the same *96* independent variables as input variables, but included in addition *25* latent variables which implied from *2477* to *2660* parameters to be determined.[2] Based on the sample of *1,845* individuals, the parameter sets for each of the *20* models were computed by means of a standard Backpropagation algorithm.

The linear and the non-linear imputation models were applied to impute alternative values for individuals in *Sample 2* and *Sample 3*. The individual observed values for individuals in *Sample 1* and *Sample 2* plus the imputed values for the individuals in *Sample 3* were then aggregated to form *imputation estimates* of population proportions for survey variables.

Finally, all sets of estimates were compared with target proportions computed from the available, but not used, observed survey variable values for the whole population.

Total population:	*N=17,326*
Sample 1:	$n_1 = 1,845$
Sample 2:	$n_2 = 162$
Sample 3:	$n_3 = 15,319$
Sample a:	$n_a = 18$ from *Sample 1+2*
Sample b:	$n_b = 144$ from *Sample 3*.
|*Sample 1a-1e:*	*n about 1845*

Box 1: Summary of population and samples used.

[2] As for the intercepts in regressions, the imputation models required auxiliary variables with constant values 1. The number of parameters for a non-linear model is for example:
Parameters=Latent variables*(Register variables +1)+ Survey variables*(Latent variables +1).

In addition to the above samples, the population of a small census tract was extracted to test the validity of the estimates for an area with few inhabitants. This subpopulation included *162* individuals (it is a coincidence that this tract has the same size as *Sample 2*) of which *18* individuals belonged to *Sample 1+2* and for which survey variable values therefore were assumed observed and available. These *18* individuals were referred to as *Sample a* while the remaining *144* of the census tract were denoted *Sample b*.

y_1, y_2 and y_3: individual observed survey variables in *Samples 1, 2* and *3*.

$\ddot{y}_1$, $\ddot{y}_2$ and $\ddot{y}_3$: means of y_1, y_2 and y_3 in *Samples 1, 2* and *3*.

y_2' and y_3': individual imputed values in *Samples 2* and *3*.

z_2 and z_3: individual differences $(y_2' - y_2)$ and $(y_3' - y_3)$ in *Sample 2* and *3*.

x_1, x_2 and x_3: individual observed register variables in *Samples 1, 2* and *3*.

Box 2: Notations used for variables.

To demonstrate the effects of a random training sample on the imputation estimates, *5* additional random and mutually exclusive samples of size about *1800* individuals were finally selected from *Sample 3*. These are referred to as *Sample 1a-1e.* For each of these, independent imputation models were computed for a selection of survey variables.

3. Empirical analysis and tests

Preliminary tests

A number of empirical tests were carried out. First, the assumption that the estimated standard deviations s_y from *Sample 1+2* are acceptable estimates of the corresponding standard deviations for the whole population, was investigated. Table 1, Columns (1) and (2) display the standard deviations for *Sample 1+2* and for *Sample 3*. The figures do not reveal any significant differences, and the standard deviations *s* estimated from *Sample 1+2* were accepted for the following analysis.

A second question raised was whether the imputation models produced biased estimates. For this purpose, averages of the *z* variables were computed from *Samples 2* and *3* and reported in Columns (3),) and (4) of Table 1. The average biases are small for most variables. There is no good correspondence among the averages of *z* in *Samples 2* with the averages of *Sample 3*. It seems not possible to make useful predictions for biases in *Sample 3* from computations based on *Sample 2* data.

The accuracy predictor for imputation estimates required the *RMSE* of the deviations *z*. Comparison of the $rmse_1$ from *Sample 1* in Column (5) with the $rmse_3$ Column (7) from *Sample 3* in , confirms previous experience that estimates from a training sample underestimate the $rmse_3$. Column (6) shows the $rmse_2$ estimated from predictions and target values of *Sample 2*. These estimates are close to the $rmse_3$ from *Sample 3*, and were therefore used in the computation of accuracy measures for the estimated proportions. A comparisons of the *rmse* in Columns (6) and (7) with the standard deviations for the unbiased estimates in Columns (1) and (2) show that *rmse* are significantly smaller indicating better accuracy of the imputation estimates.

Experimental computations of estimates for *RMSE* by means of the cross-validation for non-linear models proposed by Moody, were also carried out for some selected variables on observations in *Sample 1* (Moody 1993). A 10-fold cross validation was used. The method implied the division of *Sample 1* into *10* non-overlapping random subsamples of approximately equal sizes. Ten different sets of parameter estimates for each imputation model were trained by leaving one different subsample out each time. For each parameter estimate set, the mean square error of *z* were computed for the left out subsample. The *10* sets of mean square errors were averaged, and finally the cross-validation estimates $rmse_{cross}$ computed.

The results of the cross-validation computations for the first imputation model, shown in Box 3, indicated that the $rmse_2$ estimates from *Sample 2* gave results closer to the $rmse_3$ of *Sample 3*, than did the cross validation $rmse_{cr}$ estimates from *Sample 1*. Still, with few observations available, cross validation can obviously be a useful method for estimating accuracy predictions.

Variable	Sample 1 $rmse_{cross}$	Sample 2 $rmse_{2}$	Sample 3 $rmse_{3}$
Cohabitance			
Nobody	0.248426	0.249405	0.250639
Spouse	0.142619	0.163140	0.163947
Cohabitant	0.187668	0.199472	0.200459
Children	0.315753	0.269644	0.270978
Parents	0.227954	0.187937	0.188867
Siblings	0.238513	0.222370	0.223470
In-laws	0.120515	0.053572	0.053837
Grandparents/-children	0.110045	0.061185	0.061488
Other	0.214668	0.174427	0.175290

Box 3: Root mean square error estimates from Sample 1 by means of non-linear cross-validation and from Sample 2 and 3 by ordinary computation.

Estimates for the municipality population

Column (1) of Table 2, give the target proportions from Population Census data for the total municipality, while Columns (2), (3) and (4) present the unbiased, the linear and the non-linear imputation estimates of the proportions, respectively, for the whole population. Estimates from all three estimators gave values very close to the target proportions.

Columns (5) to (7) of Table 2 give the accuracy predictions for the three sets of estimates while the following Columns (8) to (10) give the corresponding relative predictions. Inspection of the figures reveals higher absolute as well as relative predicted accuracy for the imputation estimates than for the unbiased estimates. However, for a population and a sample of the sizes used, all estimators will in general give good results.

Estimates for the census tract

Table 3 displays the application of the three estimators on a census tract area with only *162* inhabitants of which *18* were identified belonging to *Sample a* for which observations were assumed available. The four Columns (1)-(4) show the target proportions, the unbiased, the linear and the non-linear imputation estimates. The target proportions were computed from the sums of all *162* observations in *Sample a + b*. The unbiased estimates were based on the observations in *Sample a,* while the linear and the non-linear imputation estimates were aggregated from the sums of observations from *Sample a* and the sums of the *144* individual imputations made for *Sample b*. Inspection of the estimates shows that the linear imputation estimates are in average much closer to the target proportions than the unbiased estimates and the non-linear imputation estimates are in general much closer to the target proportions than the linear imputation estimates. As could be expected because of the small size of *Sample a*, the unbiased estimates were very unreliable.

Predictions of the relative estimate accuracies were computed based on the predictors discussed above, and are reported in Columns (5) to (7) in Table 3. The corresponding actual, relative deviations between estimates and target proportions are given in Columns (8) to (10).

Boxes 4, 5 and 6 demonstrate the validity of the accuracy predictors for the unbiased, linear and non-linear imputation estimates, respectively. Assume that the publication principle is to publish only estimates which are predicted to deviate less than +/- *20*% from the target value with the risk that one out of three predictions is incorrect.

The sum in the first column of Box 4 shows that the unbiased estimator provided *14* estimates which satisfied the requirement for publication. The linear and the non-linear imputation estimators gave *13* and *26* estimates which satisfied the publication criteria according to the sums in the first columns of Boxes 5 and 6, respectively.

		Observed:		
		<0.2	>=0.2	Sum
Predicted:	<0.2	0	2	2
	>=0.2	14	23	47
	Sum	14	25	49

Box 4: Predicted and observed relative accuracy of the 49 unbiased estimates.

It was pointed out above, that the practical value of an estimator, however, depends on the possibility to predict the accuracy of the estimates produced. The first three boxes illustrate the success of the accuracy predictors to predict which estimates should be published.

Only *2* unbiased estimates were predicted to satisfy the publication condition and both predictions were incorrect when compared with the actual targets, Fourteen unbiased estimates were predicted to be too inaccurate for publication while they in fact satisfied the publication requirement.

The predictor for the linear imputation estimates identified *14* estimates which should deviate less than *20* % of which *3* were incorrect predictions. On the other side, only 2 estimates which should have been accepted were rejected by the prediction.

For the non-linear estimates, the predictions implied that *20* estimates should pass the publication criterion of which *2* were incorrect when confronted with the actual target values. Among the *29* non-linear imputation estimates predicted to have a relative accuracy not satisfying the requirement for publication, *8* proved to be acceptable when compared with the actual targets.

These results indicate that the non-linear imputation estimates for the proportions of the census tract are significantly closer to the target proportions than both the unbiased and the linear imputation estimates. Also the accuracy predictions for the imputation estimates seem to be significantly more reliable than those for the unbiased estimates.

		Observed:		
		<0.2	>=0.2	Sum
Predicted:	<0.2	11	3	14
	>=0.2	2	33	35
	Sum	13	36	49

Box 5: Predicted and observed relative accuracy of the 49 linear estimates.

		Observed:		
		<0.2	>=0.2	Sum
Predicted:	<0.2	18	2	20
	>=0.2	8	21	29
	Sum	26	23	49

Box 6: Predicted and observed relative accuracy of the 49 non-linear estimates.

The above analysis has in part been repeated for the *55* other census tracts in the same municipality. The analysis for these areas supported the results reported above. To test if the imputation models could be used *outside* the municipality population from which the training sample was drawn, a second municipality from another part of the country and with a different socioeconomic structure, was studied. This second municipality was also surveyed *100%* in 1990. It comprised *44* census tracts and had only *230* inhabitants in average per tract. In the experiment, it was assumed that no sample survey had been carried out in this second municipality. No unbiased estimates could therefore be computed, and all individual values for the survey variables had therefore to be imputed. The non-linear imputation models developed for the first municipality were used for this purpose. A relative high number of the estimates for the small areas in this second municipality also satisfied the publication requirements. The accuracy predictions seemed to be almost as promising as those reported in Box 6 above for the first municipality. A detailed report on these small area experiments will be published in a future paper.

Effects from the random selection of the training sample

It was pointed out above that the accuracy predictions used in the previous paragraphs include the errors due to random selection of the training sample. However, resampling of the training sample can produce more or less accurate imputation estimates with corresponding accuracy measures.

Box 7 illustrates the variations in the P_E estimates for the Cohabitance proportions in the census tract based on imputation models with parameters computed from *Sample 1* and *5* other mutually exclusive random training samples.

Cohabitance	Sample						Target
	1	1a	1b	1c	1d	1e	
Nobody	0.109	***0.117***	0.093	***0.117***	0,111	***0.117***	0.142
Spouse	0.472	***0.475***	0.469	0.481	0.481	0.463	0.475
Cohabitant	0.109	0.093	***0.105***	0.086	0.080	0.086	0.105
Children	0.323	0.370	0.358	0.370	***0.296***	0.315	0.296
Parents	0.135	0.160	0.148	0.148	***0.130***	0.142	0.130
Siblings	0.073	0,049	0.080	0.031	0.043	***0.062***	0.062
In-laws	***0.031***	0.025	0.025	0.25	0.056	0.019	0.037
Grandpar./-child	0.030	0.025	0.025	0.025	***0.037***	0.019	0.037
Other	***0.029***	0.025	0.031	0.031	0.025	0.012	0.049

Box 7: Estimates for census tract based on different training samples.

From the figures we can seethat the different tra9ining samples resulted in different final results. The estimates with the smallest deviation from their target proportions are marked. The figures indicate that *Sample 1*did not produce significantly better or worse estimates than the other *5* training samples.

Box 8 shows similar information about the Means of transportation estimates from the different training models. Also these figures support the assumption that different random traing samples produce different imputation models, but with sample of the size used, the variations in the final imputation estimates are moderate.

Means of transportation:			Sample				Target
	1	1a	1b	1c	1d	1e	
Car	***0.311***	0.358	0.340	0.259	0.191	0.364	0.309
Bus	0.031	***0.025***	***0.025***	0.031	***0.025***	0.019	0.025
Train	***0.043***	0.025	0.025	0.031	0.025	0.025	0.037
Boat|	0.024	0.019	***0.025***	0.019	***0.025***	***0.025***	0.025
Bicycle	0.108	***0.086***	0.031	0.062	0.031	0.031	0.080
Other	0.028	***0.025***	***0.025***	***0.025***	***0.025***	***0.025***	0.025

Box 8: Estimates for census tract based on different training samples.

4. Conclusions

The main objective of this study was to evaluate the *impute first-aggregate next* estimators for proportions based on imputed values for small areas with few inhabitants. Individual imputations were obtained by means of models estimated from data for the individuals in a sample supplemented with administrative data for all individuals. The models used were ANN feed-forward models.

The investigation indicated that imputations used in the *impute first-aggregate next* estimators, improve the results compared with those obtained by means of ordinary estimators. Accuracy predictors of these estimates were developed. In the same way as the standard error is used as an accuracy predictor for unbiased estimates, predictors for imputation estimates can be based on the root mean square error for individual imputations. The empirical study indicated that satisfactory predictors for the accuracy of imputation estimates can be based on a small sample independent of the sample used for estimating the parameters of the models. The reliability of accuracy predictions of imputation estimates seems to justify practical use.

Empirical computations for a small census tract illustrate that linear imputation estimates were significantly closer to the target proportions than corresponding unbiased estimates. The non-linear imputation estimates were even better. The use of non-linear imputation estimates should permit a substantial increase in the number of statistics which could be published, and provide a new readiness to prepare statistics on request.

The effects of the random training sample on the accuracy of the estimates were investigated by resampling and recomputing the models. For training samples as large as the one used in the experiments, the sampling should not be expected to have a significant influence on the results. The predicted accuracy measures will always reflect the error effects due to the training sample.

5. References

Bing Cheng and Terrington, D.M. (1994): Neural Networks - A Review from a Statistical Perspective, Statistical Science. Vol. 9. No. 1. pp 2-54.

Cheng-Ming Kuan and White, H. (1994): Artificial Neural Networks: An Econometric Perspective. Econometric Review, Vol. 13, No. 1, pp. 1-91.

Minsky, M. and Papert, S.(1969): Perceptrons- An Introduction to Computational Geometry. MIT Press. Cambridge, Mass.

Moody, J.E. (1993): Prediction risk and architecture selection for neural networks, In Charkassy, V., Friedman, J.H. and Wechsler, H. (Eds.): From Statistics to Neural Networks. Theory and Pattern Recognition Applications. Springer. Berlin.

Nordbotten, S. (1996): Neural Network Imputation Applied on Norwegian 1990 Population Census Data Utilizing Administrative Registers. Journal of Official Statistics. Vol. 12, No. 4, pp.385-401.

Rumelhart, D.E. and McClelland, J.L. (1986): Parallel Distributed Processing. The MIT Press. Cambridge, Mass.