New Methods of Editing and Imputation

Svein Nordbotten

University of Bergen, Norway

ABSTRACT: Editing of data collected for preparation of statistics is a time and resource consuming process. This paper presents experiments with artificial neural networks as a potential tool for increasing the effectiveness of statistical editing and imputation. To maintain accuracy in resulting statistics, the possibility of deriving reliable accuracy predictions is also discussed.

1. Introduction

Producers of statistics have always been concerned about the quality of their statistics. One hundred years ago, the name *data editors* was used on the staff who had the responsibility to inspect and control the data collected from the respondents. The appearance of programmed electronic computers 50 years ago offered opportunities to automate the editing. Already early in the 1960s, computerized editing of the U.S. Agricultural Census data was for example extensively used, but it is estimated that 20 to 40 percent of the total costs of a survey or a census are still used for editing [Granquist 1997]. To look for new opportunities to reduce the resources required for editing is therefore a continuous challenge.

The purpose of this paper is to discuss an approach to statistical editing, the main issue of which is to use available information about statistical units more effectively in control and imputation. This is no new idea. It was discussed in international meetings as functional editing more than 30 years ago [Nordbotten 1963]. Because of more efficient technical tools and more available information, functional editing is far more realistic today than at the time it was first discussed. This paper will discuss the editing and imputation process, and some aspects of neural network methods, and will conclude with illustrations based on experiments with the functional approach carried out on real life data by means of artificial neural networks. Examples of similar experiments have also been discussed by others [Roddick 1993, Teague and Thomas 1997].

2. Problem: Editing and Imputation

Different types of errors are introduced in preparation of statistics. The survey design introduces design errors, data collection introduces register, sampling, interviewer and response errors, while the aggregation and dissemination processes are the sources of processing and presentation errors. Statistical data editing plays a special role in preparation of statistics because it aims at reducing the effect of errors in other statistical processes. In editing, we distinguish between *detecting* errors and *correcting* errors. Detection of errors is carried out by a *control* process, while correction of errors is done by means of an *imputation* process.

When discussing editing, it is useful to distinguish among three versions of data compiled about a statistical unit: 1) the *raw* or collected values, 2) the *edited* values obtained by the editing process to be investigated, and 3) the *target* values which we would have obtained if an *ideal* measurement had been performed. Frequently, the values from an edit performed by subject matter experts is considered to give a satisfactory approximation to target values.

A statistical estimate, Y', is a measurement of a population target, e.g. a total, Y, based on values for individual statistical units. We shall refer to $D=|Y'-Y|$, the deviation of the estimate from the target, as the *accuracy* of the estimate. Since Y is usually unknown, the accuracy must be predicted and we denote the predicted accuracy by D'.

Consider a population of N statistical units. Without a loss of generality, we assume that each unit is characterized by a single y-variable value and a set of K x-variable values. The y-value is an unknown value of a variable *observed* with value y'. The x-values are *background* or *auxiliary* data available for the statistical unit. The source of these background values may be administrative registers, previous surveys, etc. We denote the individual *observed* values as y_i' and the *target* values as y_i, $i=1..N$, and the auxiliary values in the x-set as x_{ik}, $i=1..N$ and $k=1..K$.

The difference between observed y_i' and target y_i values is the individual *error* $e_i=y_i'-y_i$. The value of the error is assumed to be the cumulated effect of all factors influencing the outcome of an observation. We assume that e_i, $i=1..N$, can be considered as independent random variables with a common mean m_e and variance v_e. An estimate $Y'=\Sigma_1^N y_i'$ of all unedited data can therefore be considered as stochastic and be expected to vary if the observation is repeated.

Since the accuracy of Y' does not manifest itself before the statistical information is used, producers of statistics want to issue an *accuracy declaration* or a guarantee to the users about the statistical total. The declaration can, for example, be based on a predetermined *accuracy requirement* C. A narrow accuracy boundary C implies declaration of high quality estimates, and vice versa. Since the purpose of editing is to increase quality, the edited data should therefore be accompanied by some kind of accuracy predictions on which accuracy declarations for the derived statistics can be based.

If the accuracy prediction D' of an estimate Y' satisfies the condition $D' \leq C$, the estimate will be declared to have required accuracy; if not, the estimate will not be declared accurate. It is important that the accuracy predictions have a low probability Pr to be incorrect. By means of the Central Limit Theorem, bounds for such a probability can be derived if the estimate is composed of a satisfactory number of components. If not, we may rely on the Tchebycheff Inequality Theorem to specify the bounds.

3. New Methods

3.1 ANN Summary

Artificial neural networks (ANN) were first presented in a paper by McCulloch and Pitts with the aim to explain neuro-physiological phenomena [McCulloch 1943]. Later, a variety of ANN models have been formulated for different problems, learning algorithms for numerical computation of the parameters of the networks have been developed, and a variety of neural network applications have appeared in different fields.

The type of ANN used in connection with the problems discussed in this paper is so-called "feed-forward", multi-layer networks which can be considered as logistic discriminant functions or non-linear regression functions. Input terminals receive values corresponding to explanatory variables, while an output neuron provides values corresponding to a dependent variable. In between are layers of hidden neurons. Each input terminal and hidden neuron is interconnected with all neurons in the next layer, meaning its value is distributed through all connections to the receiving neurons. The individual connection is characterized by a parameter value called the weight by which the input value

is multiplied. A neuron transforms the sum of all received weighted input values to an output value according to an activity function.

The development of an artificial neural network consists of two steps: 1)*the choice of network topology and activity functions for the neurons*, and 2) *the determination of the connection weight parameters of the network*. The first step must be done according to the problem for which the network should be applied, the success of which depends to a large extent on the developer's experience and creativity. The second step is carried out by means of a learning algorithm. The learning algorithm for the type of networks discussed in this paper is known as the Backpropagation algorithm.

The Backpropagation algorithm is an iterative, numerical algorithm based on stepwise adjustment of parameters according to the gradient decent principle and became generally available about 10 years ago [Rumelhart 1986]. It requires a set, called the training set, of examples of input and output value pairs from the domain to which the network will be applied. The learning algorithm searches for a set of weight values characterizing the connections among the active nodes of the net such that the net will reproduce the output values with the minimum mean square error given the input values of the set of examples. Modern implementations of this and other training algorithms offer several options for the training, such as how to initialize the weights, the size of learning steps, when to stop the iteration, etc.

3.2 Relationship to Statistical Methods

A single layer neural network which uses nodes with sigmoid activity functions can, subject to certain conditions as to how it is trained, be considered as expressing a conditional probability, a logistic discriminant function or a multivariate logarithmic regression function, and can be presented as:

$$y_i' = 1/(1 + exp\ -[w_0 + \sum_j w_j * x_{ij}\ ])\quad for\ i = 1..N, \tag{1}$$

where the y' is the probability or dependent variable corresponding to the output of the net and the x's are independent variables corresponding to the inputs to the net. The w's are parameters known as connection weights of the net which have to be determined.

To obtain better results, the network is designed as multi-layered. A network with two layers of weights connecting the inputs to a set of hidden neurons can, for example, be presented as:

$$y_i' = 1/\{1 + exp\ -[w_o + \sum w_k * (1/(1 + exp\ -[w_k + \sum_j w_{jk} * x_{ij}\ ]))]\}\quad for\ i = 1..N. \tag{2}$$

It can be proved that a 3-layer network can represent a very large set of different functions.

There are several ways to draw statistical conclusions about the accuracy of statistics based on results from a neural network [Bishop 1995]. One simple way is to draw a new random sample independent of the training sample from the population, obtain its true values, compute predicted values of the variables by means of the network, and estimate the statistical characteristics of the deviations. This approach was used for the second experiment presented in this paper. Obviously, this requires that resources are available for observing both samples. Use of jack-knife methods is another approach which has the advantage that only one sample is needed, but it requires significantly more computing [Moody 1993].

3.3 Imputation Estimates

Trained networks give a basis to impute individual values for all units not observed, and open the use of a simple estimator for computing the estimates of totals,

$$Y' = \sum_1^n y_i + \sum_{n+1}^N y_j', \quad (3)$$

in which the first term is the sum of the target values known, and the second, the sum of the imputed values. These estimates are called *imputation estimates* [Nordbotten 1997].

Assuming the imputed values have a random error component, the root mean square error of the imputation estimate of a total is

$$RMSE = \sqrt{(N-1) * mse}\ , \quad (4)$$

where *mse* is the mean square error for the residual error variables estimated from the second sample mentioned above. As long as the squared bias component is small compared with the variance of *mse*, *RMSE* can be used in the same way as the sampling error from the sampling theory [Hansen 1953].

If the conditions permit to rely on the Central Limit Theorem, we can use the theorem as justification for a probability prediction about an upper bound for the imputation estimate,

$$PR(|Y'-Y| \leq 2*RMSE) \approx 0.95, \quad (5)$$

saying that only 1 out of 20 estimates is expected to deviate more than 2 times its *RMSE*. If the statistical agency wants to release only estimates of totals which do not deviate from the true value with more than 5 units, we should reject estimates with $RMSE > 5/2$ and expect that 1 out of 20 rejected estimates in fact was acceptable.

If the estimate is based on a sample which is too small for relying on the Central Limit Theorem, the Tchebycheff Inequality Theorem,

$$PR(|Y'-Y| \leq 3*RMSE) < 9/10 \approx 0.9, \quad (6)$$

can be another possibility for accuracy predictions. This theorem tells us that about 1 out of 10 estimates are expected to deviate more than 3 times the *RMSE*. If only estimates which deviate with 5 or less from the true total can be considered acceptable, we should reject estimates with $RMSE \geq 5/3$ and be prepared that 1 out of 10 rejections should have been accepted.

4. Empirical Example I

4.1 Experiments

The first illustration of the methods described above uses data from the Norwegian Agriculture and Forestry Surveys carried out by Statistics Norway (SSB). The experiment demonstrates the use of an ANN to discriminate between acceptable and deficient records in editing control. A decennial agricultural census was taken in 1989. Using this as a frame, a stratified sample was selected for annual surveys. The data collected for the census and the annual surveys have been edited by agricultural subject matter specialists.

In 1996, SSB decided to evaluate a computerized control compared with human editing of the annual surveys. Unit data collected for two counties were recorded in three alternative versions: the *raw* data

records from the questionnaires, the *corrected* data records based on the computerized control rules supplemented by human correction of records failing the control, and the *edited* data records from the questionnaires completely edited by the specialists. Evaluation reports concluded that the data controlled by a rule-based computerized method followed by a manual correction of records rejected by the control gave data which, when aggregated, deviated insignificantly in some fields, but not in all, from aggregates on data both controlled and corrected by specialists.

Experiments were initiated to investigate: 1) if it is possible from a sample of edited records to train a neural network to function like a human expert and detect the same deficiencies in raw records as human editors do, and 2) if a neural network can be trained on a sample of deficient raw records and the corresponding edited records to perform the same corrections as the human editors do. Some results from the first part will be reported below, while the second part of the experiment is still being worked on.

The experiments were based on records for $n+N=3030$ units. From these, $n=1030$ records were randomly selected for training a two-layer ANN, while the remaining $N=2000$ were reserved for testing. Each record comprised 593 bytes distributed to 154 variables. The control task to be simulated was to classify each raw record as either acceptable or suspicious.

Since the ANN used was based on sigmoid output functions, the output value would be within the range $0<f(x_{i1}\ldots x_{iK}\,;\,w)<1$. The value can be considered as an estimate of the probability that the record is correct. The following decision rule was therefore established based on the outcome of the ANN:

$$\begin{array}{lll} f(x_{i1}\ldots x_{iK}\,;\,w) & >= & a = 0.5 \text{ then the record is classified as correct} \\ & < & a = 0.5 \text{ then the record is classified as suspicious} \quad \text{for } i=1..N. \end{array} \tag{7}$$

In this case, the f can be considered a discriminant function and 0.5 a cutting point defining the classification regions. The a reflects the cutting point with a minimum risk for misclassification if the loss of rejecting a record which should have been accepted is equal to the loss of accepting a record which should have been rejected. If the statistical producer has a conservative strategy and tends to think that it is better to inspect too much than too little, a should be set > 0.5, or vice versa.

4.2 Training the Control Function

The ANN used consisted of 154 input variables, two layers of weights connected by 154 hidden neurons and one output neuron, the decision variable. In total, this model included 24,025 weights to be determined. To avoid the problems of multiple local minima, a problem frequently met with this type of model, the weights were given small initial random values and a learning rate which decreased linearly in value from 0.5 to 0.1 as the learning improved was used. The training function learned to correctly predict the values of the decision variable for the 1030 records of the training set during 2035 iterations.

4.3 Testing the Control Function

The practical use of the trained function depends, however, on how well it performs on an independent set of records. It was, therefore, applied to predict the decision variable values for and classify the 2000 test records not used in the training. When the classified records were compared with records edited by the editors, it was found that the control function had failed to predict 514, or 26 percent of the raw records correctly. Two types of misclassification appeared. Type 1 misclassification errors

included correct records which were predicted to be erroneous. Type 2 errors included deficient records which were incorrectly predicted as acceptable by the control function. We assumed that all records classified as incorrect should be inspected by the editors who would identify the 252 Type 1 errors. Among the 2000 test records, 262 records were classified as correct by the control function. These 13 percent erroneous records would not be detected by the editors.

Table 1: Records' Prediction Compared with True Status

	True		
Predicted	Correct	Incorrect	Sum
Correct	317	262	579
Incorrect	252	1169	1452
Sum	569	1431	2000

4.4 Comparing the Two Sets of Corrected Records

A first task was to compare the above result from ANN control with the results from the control by means of the computerized rule control used in the SSB evaluation. The number of records from this computer control which differed from the version which had been controlled and corrected by experts was 722, or 36 percent of the 2000 records. Compared with the 13 percent of the records which would not be corrected using the ANN-based control, a preliminary conclusion was that the ANN control function detected more errors than the SSB computerized control.

After eliminating six fields with data irrelevant for this purpose, 149 totals were generated from the records resulting from the SSB control, the records from the ANN control, and the records controlled by the experts. The average relative errors,

$$E = \sum \mid (t_{stat} - t_{true}) / t_{true} \mid / 149 , \qquad (8)$$

were computed and gave the results $E_{SSB} = 0.13$ and $E_{ANN} = 0.06$.

The average deviation for the totals from the ANN control were only 50 percent of those resulting from the SSB control. The ANN control rejected 1452 records out of the 2000 for inspection by the experts, implying that the experts had to inspect about 73 percent of the records, or a saving of only 27 percent. Unfortunately, it is impossible to deduce from the available data sets exactly how many records of the raw data set were rejected by the SSB computerized control for inspection by the experts.

The next question, which so far has not been answered, is if some of the 1452 records could have been processed by an ANN imputation model instead of editors.

5. Empirical Example II

5.1 Experiments

Extensive experiments with ANN have been performed on Norwegian data from the 1990 Population Census [Nordbotten 1996]. The data used in this presentation are from a municipality with 17,326 individuals distributed to 56 small census tracts, and the aim of the presentation is to illustrate mass

imputation of individual data for deriving small area statistics. To be quite correct, the areas may not be small in space, but they contain a small number of statistical units.

The municipality is located in the middle part of Norway with farming, some manufacturing and transport as its main industries. The average population of a small area was 310 inhabitants. The majority of areas had from 100 to 300 inhabitants, while some had less than 50 inhabitants.

The 1990 Census in Norway was based partly on administrative data available for each inhabitant, referred to as *x*-variables, and partly on data collected particularly for the census, referred to as *y*-variables. For most municipalities, observations of *y*-variables were collected from random samples containing about 10 percent of the inhabitants. The municipality used in the study, however, wanted statistics on the *y*-variables based on complete counts for all inhabitants and paid for the additional observations itself. The data from this municipality were thus well suited for the type of experiments we wanted to perform. In the experiments, we simulated a situation in which a normal sample survey was taken and imputed values for the *y*-variables for each individual in the remaining population were computed. Since computed totals and accuracy predictions could be compared with observed target totals and evaluated in this municipality, the data represented a unique possibility for an evaluation study.

Two sets of *y*-variables were included in the experiment, in total 15 variables, which were assumed to have been observed in the sample survey, while 97 auxiliary *x*-variables were available from administrative registers for all individuals. The 15 *y*-variables were about cohabitation and the means of transportation to work. Two separate, simultaneous ANN with together 15 non-linear imputation functions were trained and used to generate imputations for the inhabitants not included in the sample. Both models used the individual values of the 97 *x*-variables as independent variables for imputing the individual *y*-variables. In addition, both models included 25 hidden neurons. The details about design and construction of the neural network models used for imputations and the computation of the mean square errors have been described elsewhere [Nordbotten 1996].

The sample drawn from the population in the municipality was divided into two random, mutually exclusive samples. The first training sample, $S^{(1)}$, comprised 1845 individuals and was used to determine altogether 5240 parameters for the two imputation models. The second sample, $S^{(2)}$, with only 165 individuals, was used to estimate mean square errors and biases of the imputed variables.

The *observed* *y*-values for each of the 2,007 inhabitants in the two samples were used in computing the first term of formula (3) for computing the imputation estimates. The second term of the estimates for the remaining 15,319 individuals outside the samples observed was obtained as the sum of the individual *imputed* y'-values.

Eight hundred and forty Y'-totals were estimated for the 56 areas. The corresponding target *Y*-totals were aggregated as references from the "true" *y*-values for all individuals. For each estimated total, the imputation error *RMSE* was computed according to formula (4). In order to test the validity of applying the computed *RMSE*s for predicting the accuracy of the estimates, all imputation estimates with $RMSE > 2.5$ were identified.

Relying on the Central Limit Theorem, this condition is equivalent to predicting that these estimates will not deviate from targets with more than 5 individuals with a confidence of 0.95. Alternatively, if we rely on the Tchebycheff Inequality Theorem, the confidence is reduced to about 0.75 for the same accuracy predictions.

5.2 Estimates and Accuracy Predictions

The last row of Table 2 shows that 658 of the 840 imputation estimates deviated from their target totals with 5 or less individuals. Other experiments showed that only 4 percent of the estimates deviated from their targets with more than 10 individuals. This indicates that the imputation estimates are quite accurate and should be useful information. The question is, are we able to predict which of these estimates satisfy our conditions that only estimates with deviation 5 or less should be released?

Table 2: Predicted and Observed Accuracy for Imputation Estimates of 840 Totals in Small Areas

	Observed $\|Y'-Y\| <= 5$	Observed $\|Y'-Y\| > 5$	Sum
Predicted $\|Y'-Y\| <= 5$	433	29	462
Predicted $\|Y'-Y\| > 5$	255	153	387
Sum	658	182	840

Table 2 also indicates that 462 estimates were predicted to satisfy the condition of a deviation with 5 or less individuals. Of these, nearly 94 percent were correctly predicted, i.e. only 29 errors of Type 2 occurred. From the assumption about normal distributions, we would expect that the number of Type 1 errors, rejecting estimates that satisfy the requirement, would be less than 42, or 5 percent of the estimates. The table shows, however, that 225 estimates were incorrectly predicted to deviate from their targets with more than 5 inhabitants. If an application is such that Type 2 errors are relatively expensive while lost opportunities represented by Type 1 errors are of little concern, this would probably be an acceptable set of predictions.

We may suspect that for so many small populations, the use of the Tchebycheff Theorem is more realistic than assuming a normal distribution. Using $RMSE > 2.5$ subject to the assumptions of this theorem means that the level of confidence for the predictions is reduced to 0.75. This corresponds to an expected number of Type 1 errors about 210, which makes the 225 observed more acceptable. If in this case we require a confidence level of 0.95, the condition for rejection of an estimate must be changed to $RMSE > 4.3$. A possible solution being investigated is to use $RMSE > 4.3$ estimates for areas with population below a certain size, for example 100 and less individuals, and $RMSE > 2.5$ for areas with population above this size.

6. Final Remarks

The first experiment demonstrated how ANN can be used for editing control of agriculture data and provide results which may compete with computer control methods. It is expected that further experiments will give answers to how well ANN can also be trained to make imputations for rejected observations.

The second experiment reported in this paper confirms that useful statistics for small areas/populations can be obtained by aggregating individually imputed variable values. The imputed variable values for units not observed were based on an ANN trained from an observed sample of the area population and

available auxiliary data. Our results indicate that individual imputed values can be aggregated to useful estimates of totals for areas with too few units for the use of traditional estimates.

Accuracy predictions for the imputation estimates can be based on mean square errors for the individual deviations between imputed and target values estimated from a small sample independent of the ANN training sample.

Acknowledgment

This paper presents research from a project SIS (Statistical Information Systems). The experiments referred to in this presentation were carried out in cooperation with and supported by Statistics Norway. The views expressed, however, do not necessarily reflect those of Statistics Norway. Further information about the SIS project is available on the World Wide Web at *http://129.177.34.238/sis/sis.html-ssi*.

References

Bishop, C.M. (1995), *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press.

Granquist, L. (1997), "An Overview of Methods of Evaluating Data Editing Procedures," *Statistical Data Editing, Vol. 2. Methods and Techniques*, Statistical Standards and Studies No. 48. UN/ECE, New York and Geneva, pp. 112-122.

Hansen, M., Hurwitz, W.N. and Madow, W.G. (1953), *Sample Survey Methods and Theory, Vol. I*, New York: John Wiley & Sons.

McCulloch, W.S. and Pitts, W. (1943), "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics* No. 5, pp. 115-133.

Moody, J.E. (1993), "Prediction Risk and Architecture Selection for Neural Networks," in *From Statistics to Neural Networks, Theory and Pattern Recognition Application*, eds. V. Charkassy, J.H. Friedman and H. Wechsler, Berlin: Springer.

Nordbotten, S. (1963), *Automatic Editing of Individual Statistical Observations*, Statistical Standards and Studies No.2. UN Statistical Commission and Economic Commission of Europe, New York.

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

Nordbotten, S. (1998), "Estimating Population Proportions from Imputed Data," to appear in *Computational Statistics and Data Analysis*.

Roddick, H. (1993), "Data Editing Using Neural Networks," *Memorandum*, Statistics Canada, Ottawa.

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

Teague, A. and Thomas, J. (1997), "Neural Networks as a Possible Means for Imputing Missing Census Data in the 2001 British Census of Population," *Survey and Statistical Computing 1996*, Association for Survey Computing.