

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (iii): New methodological and technological developments in statistical data editing

ERROR IDENTIFICATION AND IMPUTATIONS WITH NEURAL NETWORKS

Submitted by Statistics Denmark¹

Contributed paper

I. INTRODUCTION

1. Most real data sets include missing values and data values in error. An often used way of handling data sets with missing values is to complete the missing values using some kind of imputation algorithm (see e.g. Little and Rubin (1987)). The advantage of any imputation method is that it allows to obtain complete data sets and complete data statistics. However, the disadvantage is that the algorithm often underestimates the variance.

2. In some cases, the possible value of the variable for which data is missing is not known - e.g. if all people with a smoking problem do not answer questions like, 'how many cigarettes do you smoke per day?'. In such a case any automatic imputation algorithm will have problems. If some other variable indicates that the person has a smoking problem and some people with smoking problems have answered the question, then it will be possible to make an automatic imputation.

3. The most frequently used algorithms for input are the following:

- **Mean Imputation:** The sample mean of a variable is used to replace any missing data for that variable. This is the simplest method of imputation. This can also be done in separate groups.
- **Hot-Deck Imputation:** Missing values are replaced with values taken from matching respondents, i.e. respondents that are similar with respect to some variables observed for both respondents.
- **Last Value Carried Forward or LVCF** - the last observed value is used to fill in missing values at a later point in the study.

The advantages of the above-mentioned algorithms for imputation are that they are simple to implement and easy to understand. The disadvantage is that they can introduce serious bias into the data.

4. This article provides the results of testing imputation and error-localisation using neural networks. The advantages and disadvantages of this method are that it produces probabilities for each of the possible values of a class variable and the imputations can be repeated, so that the variance from the imputations can be estimated. However, the method is complex and not easy to understand.

¹ Prepared by Bjorn Steen Larsen and Birger Madsen.

II. DATA

5. For this paper, data from the Danish agricultural survey in 1997 is used. In this survey, the farmers report the amount of land used for each type of crop and number of livestock. The data set consists of 52,000 records (one from each farm) before a manual error checking procedure and one record for each farm after the manual error checking procedure. When an error is found, the production value of the farm has been changed. It is this group of records the neural network is going to identify. In the manual error check, 3012 errors were found, of which 2180 were in 'beets for sugar production', 'fodder sugar beets and other roots for fodder', 'pulses' and 'grass and green fodder in rotation'.

6. SAS Enterprise Miners Neural Networks (MLP's with one hidden layer) were used for analysis. Of the observations, 15% have been used as training data, another 15% have been used as validation data, and 70% as test data. Logistic regression is used as a benchmark.

III. RESULTS

III.1 Identifying farms with at least one error using a neural network and logistic regression

7. The two following tables show the distribution of the estimated error probabilities. It is desirable to have many observations with either a very large error probability or a very small error probability. The third column in tables 1 and 2 show the estimated number of errors if the error probability is estimated correctly. Column four shows the actual number of errors in each group. If the error probabilities are estimated correctly, column three and four will be closely related. The fifth and sixth columns show the cumulative number and percentage of errors over a certain error level.

Table 1. Error probability estimate using logistic regression.

Error probability	Records	Estimated no. of errors	Actual no. of errors	Cumulative no. of errors	Cumulative percentage of errors
90%-100%	702	692,8	586	586	27,2%
80%-<90%	84	71,8	58	644	29,9%
70%-<80%	93	69,7	61	705	32,7%
60%-<70%	67	43,5	45	750	34,8%
50%-<60%	93	51,5	48	798	37,1%
40%-<50%	129	57,4	66	864	40,1%
30%-<40%	147	51,3	60	924	42,9%
20%-<30%	324	78,1	107	1031	47,9%
10%-<20%	1397	189,0	290	1321	61,4%
0%-<10%	22917	835,8	832	2153	100,0%
Total	25953	2140,9	2153		

Table 2. Estimation of error probabilities using a small neural network (3 neurons)

Error probability	Records	Estimated no. of errors	Actual no. of errors	Cumulative no. of errors	Cumulative percentage of errors
90%-100%	712	683,6	602	602	27,96%
80%-<90%	192	163,2	134	736	34,18%
70%-<80%	142	106,4	90	826	38,37%
60%-<70%	260	168,0	157	983	45,66%
50%-<60%	193	105,9	104	1087	50,49%
40%-<50%	230	104,0	108	1195	55,50%
30%-<40%	310	107,6	133	1328	61,68%
20%-<30%	777	184,3	237	1565	72,69%
10%-<20%	730	107,2	125	1690	78,50%
0%-<10%	22407	370,4	463	2153	100,00%
Total	25953	2100,6	2153		

8. Comparing columns three and four in tables 1 and 2, both logistic regression and the neural network estimate the error probabilities well. Tables 1 and 2 show that the neural network performs slightly better than the logistic regression.

9. Table 3 shows that there were 586 records with errors and 116 without errors in the observations, to which logistic regression has given an error probability of more than 90%. The corresponding numbers for small neural networks are 602 and 110.

Table 3. Cumulative number of errors and non-errors in observations with an error probability over a given level for each algorithm

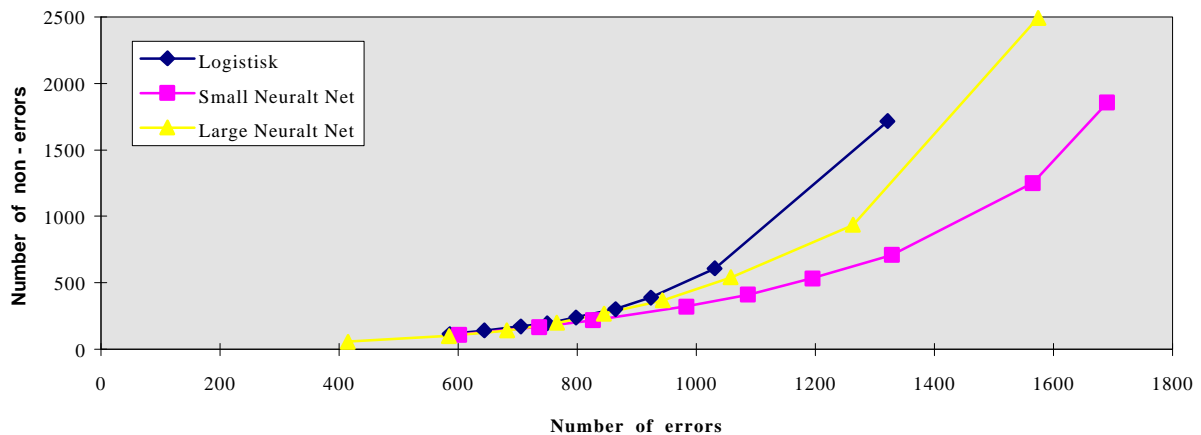
Error probability	Logistic regression		Small neural net		Larger neural net	
	no. errors	no. non-errors	no. errors	no. non-errors	no. errors	no. non-errors
>90%	586	116	602	110	415	57
>80%	644	142	736	168	585	100
>70%	705	174	826	220	682	142
>60%	750	196	983	323	766	200
>50%	798	241	1087	412	845	269
>40%	864	304	1195	534	943	367
>30%	924	391	1328	711	1058	542
>20%	1031	608	1565	1251	1263	937
>10%	1321	1715	1690	1856	1574	2493

Small network had 3 neurons.

Larger network had 10 neurons

The data in table 3 are shown in Figure 1.

Figure 1. Number of observations with and without errors for the three algorithms at 90%, 80%, 70%,...,10% error levels.



10. Figure 1 shows that the small neural network performs better than the large neural network, which is better than the logistic regression.

III.2 Identifying farms with errors in one specific variable using a neural network

Table 4. A small network identifying errors in 'beets for sugar production'

Error probability	Records	Estimated no. of errors	Actuarial no. of errors	Cumulative no. of errors	Cumulative percentage of errors
90%-100%	313	300,3	297	297	52,4%
80%-<90%	94	80,2	81	378	66,7%
70%-<80%	51	38,7	41	419	73,9%
60%-<70%	26	17	18	437	77,1%
50%-<60%	23	12,8	14	451	79,5%
40%-<50%	24	10,9	14	465	82,0%
30%-<40%	29	9,9	15	480	84,7%
20%-<30%	44	11,3	15	495	87,3%
10%-<20%	120	16,4	21	516	91,0%
0%-<10%	25229	49,3	51	567	100,0%
Total	25953	546,8	567		

11. Table 4 shows, that most of the errors in 'beets for sugar production' are found with very high accuracy. The results for logistic regression are similar.

III.3 Test of neural network ability to impute a new value for data with high error rate

12. It has been tested if a neural network can impute 'beets for sugar production' from table 4. To avoid overfitting, only observations with an estimated error rate of more than 50% have been imputed. The imputation with a neural network has been compared to estimates made with a linear regression. A small network with 3 neurons has been used. The data was split to 40% for training, 30 % for validation and 30 % for testing. A larger network was also tested but was less successful. The linear regression used forward variable selection at 5% level.

Table 5. A small network identifying errors in 'beets for sugar production'

Error probability	Mean error before imputation	Mean error with linear regression imputation	Mean error with neural network imputation.
90%-100%	610	241,7	301,7
80%-<90%	3092	733,9	1504,3
70%-<80%	2310	1384,2	1643,7
60%-<70%	2284	1430,9	1546,6
50%-<60%	1908	959,3	1520,3

13. From table 5, it can be seen that the neural network performed rather poorly, which is probably due to outliers in the data.

IV. CONCLUDING REMARKS

14. In conclusion, the following can be said:

- In identifying errors, neural networks performed better than logistic regression.
- In identifying errors in a specific variable neural networks performed well (about the same performance as logistic regression).
- In imputing a continuous variable, neural networks performed poorly (worse than linear regression).

References

Little, R.J.A and Rubin D.B. (1987): "Statistical Analysis with Missing Data". Wiley, New York.

A. C. Davison D.V. Hinkley (1997): "Bootstrap methods and their application", Cambridge.

Nordbotten, S. (1995): "Editing Statistical Records by Neural Networks". Journal of Official Statistics, Vol. 11, No. 4.

Nordbotten, S. (1996): "Neural Network Imputation Applied to the Norwegian 1990 Population Census Data". Journal of Official Statistics, Vol. 12, No. 4.