

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing  
(Prague, Czech Republic, 14-17 October 1997)

Item 4 of the provisional agenda

**NEURAL NETWORK IMPUTATION - STATISTICAL EVALUATION**

Submitted by the Office for National Statistics and University of Southampton,  
United Kingdom <sup>1</sup>

---

<sup>1</sup> Prepared by Marie Cruddas and Ray Chambers.

## **I. INTRODUCTION**

1. In the 1981 and 1991 UK Censuses a hot deck process was used to impute for missing, invalid or inconsistent observations and it is believed to have worked well - giving sufficiently accurate imputations. However, while it is simple in concept, operationally it is time consuming and resource intensive to develop and programme. The decks are programmed in a traditional way and changing them at a late stage is often very complicated and slows down the processing of the Census.
2. Neural networks (NN) were seen as an alternative to the hot deck in that they offered an automated approach to building models on which to base imputation. The advantage of the method is its ability to detect complex relationships in the data as they are received, without the need for highly complex interactive statistical modeling techniques.
3. NN technology is predominantly based on statistical modeling theory. The network is trained on valid data, learning the relationships and patterns in the data presented to it. It then forms models which link the variables together, and which are able to predict a probability distribution for each variable. The models are complicated and data dependent and the neural computer automatically decides when a new model is required, according to pre-set error limits. As output is in the form of an expected probability distribution, a random number generator can then be used to allocate a value to a missing field weighted according to the predicted probability distribution. A fuller account of the NN is given in Annex A.
4. Lacking the necessary in-house knowledge of NN, Neural Technologies Limited, (NTL), of Petersfield, Hants were chosen, through competitive tender, to work with Census Division to examine the method.
5. An initial trial was carried out to examine whether the NN could in concept impute census data. Given successful results from this trial a second trial was commissioned: an operational trial that actually imputed data into a simulation data set. This report describes the statistical aspects of the evaluation of this trial and gives some results.

## **II. THE NEURAL NETWORK TRIAL TEST DATA**

6. The data used to carry out the blind imputation test for the neural network trial was formed by knocking holes in an extract of complete data from the post-edit 1991 Census data. Thus, this is a set of data that contains no missing values and is completely consistent with edit rules.
7. Two Local Government Districts were selected referred to as County A and County B. The intention was to select a reasonably representative sample of the UK population.

Table 1.

Record type	County A	County B	Total
Households	71,459	69,908	141,367
Private persons	152,878	166,273	319,151

8. Of the complete records the following two household variables and four person variables were chosen to have 'holes' knocked in them for the neural network to then impute:

- Number of rooms (integer valued)
- Building type (8 categories)
- Age (integer valued)
- Marital Status (5 categories)
- Primary Activity Last Week (13 categories)
- Country of Birth (21 categories)

9. The holes were knocked according to a pre-defined schedule of frequencies which differed according to the combination of variables in question. The pattern of frequencies was chosen largely according to the patterns observed in the 1991 Census. Some Enumeration Districts were selected where approximately 50% of the variable were deleted to represent areas that were considered hard to enumerate.

10. Since the value of the variable was not considered when it was deleted this represented a situation where missingness was at random rather than a possibly more realistic situation where a subset of the population with particular attributes might be more likely to not respond. However this type of missingness is difficult to replicate since, by definition, the value of the items are missing.

### III. STATISTICAL EVALUATION CRITERIA

11. For census data any imputation process must

- a) retain the structure of the data and
- b) impute plausible values.

Before looking at the data it was decided for our evaluation of the NN imputation method we would examine how well marginal and joint distributions of the test data were maintained and how consistent the output was.

12. For comparison purposes the same data were fed through a process that mirrored the hot deck method used in the 1991 Census.

### a) Preservation of marginal distributions

13. For any of the six variables imputed in the trial the following type of table can be constructed by crosstabulating the actual value with the imputed value.

*Diagram showing the comparison of marginal distributions*

		Imputed value			Total
		1	2	3	
Actual value	1	10	5	5	20
	2	20	20	10	50
	3	15	5	10	30
Total		45	30	25	100

The margins of the tables have been shaded to illustrate the marginal distributions we wish to compare.

14. If the categorical variable has  $p+1$  categories, the last being a reference category, we use the following statistic to test for marginal homogeneity on a  $p \times p$  table

$$W = \left[ \frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{y}}_i - \mathbf{y}_i) \right]' \left[ \frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{y}}_i - \mathbf{y}_i)(\hat{\mathbf{y}}_i - \mathbf{y}_i)' \right]^{-1} \left[ \frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{y}}_i - \mathbf{y}_i) \right]$$

where

$\mathbf{y}_i$  =  $p$ -vector indicating which of the first  $p$  categories the  $i^{\text{th}}$  individual falls into ( $y_{ik}=1$  indicates the  $i^{\text{th}}$  individual falls into category  $k$ , otherwise  $y_{ik} = 0$ ).

$\hat{\mathbf{y}}_i$  = the imputed value of  $\mathbf{y}_i$ .

This is Stuart's (1955) extension of McNemar's statistic for a  $2 \times 2$  table to the case of a general square table. It is a WALD statistic, and, if the hypothesis is true, should have a (large  $n$ )  $\chi^2_p$  distribution (ie large values of the test statistic give evidence against the hypothesis of preservation of distributions). For continuous variables, such as age, the approach was to treat them as categorical (collapsing the tail categories) and use the above approach.

15. A secondary criteria is how well the individual values are correctly imputed. As census output is largely tabular this was felt to be a lesser concern. How well an imputation method demonstrates this ability is observed by clustering of the data along the leading diagonal of a table such as the one above. Obviously if the method gives perfect imputations all the results would be along the diagonal and the marginal distributions would also be identical.

#### **b) Preservation of joint distributions**

16. More important than the preservation of the univariate distributions is how well the distributions between variables are preserved. In this case we examine how well the imputation method preserves the margins of the table of the crosstabulation of actual values with that of imputed values (where at least one of the variables is imputed). The Wald statistic  $W$  extends easily to this case by creating a new variable  $z$  which can take on the  $p \times d$  values formed by combining the classifications of the two variables (one with  $p+1$  categories, the other with  $d$ ) and then using the statistic above, distributed as  $\frac{\chi^2}{pd}$ .

#### **c) Consistency**

17. Consistency here means that any imputed values should obey the pre-defined edit rules.

For example in the 1991 data :

- When TYPE OF ACCOMMODATION is a ONE ROOMED FLAT, ROOMS must equal 1.
- When AGE is under 16, MARITAL STATUS must be SINGLE.
- When AGE is over 16, ACTIVITY LAST WEEK cannot be 'NO CODE REQUIRED'.

### **IV. RESULTS OF THE STATISTICAL EVALUATION**

18. Initial analysis of the imputations carried out by NTL showed that for two of the six variables the NN preserved distributions better than the hot deck, see Table 2. However there were a number of problems with the consistency of the NN imputations, the degree of the inconsistency was worrying and led us to doubt that the NN was actually detecting relationships in the data. Because of this we agreed that the NN imputation method should be examined by NTL and the system run again. The results described in this section are from the second 'improved' NN imputation.

**Table 2 . MARGINAL DISTRIBUTIONS ORIGINAL NN DATA**

	ORIGINA L NN	HOT DECK
BUILDING TYPE	1015.7*	7.3
NUMBER OF ROOMS	1341.6*	47.6*
AGE	41.0*	28.1*
MARITAL STATUS	11.3*	89.8*
PRIMARY ACTIVITY LAST WEEK	921.2*	395.7*
COUNTRY OF BIRTH	18.2	46.7*

\* - indicates significance at the 5% level

### Preservation of marginal distributions

19. Each of the six imputed variables were examined to see how well the NN and hot deck methods preserved marginal distributions for the whole dataset and by area. Examples of the results are given in Tables 3a&b and 4a&b, Table 5 summarises the  $\chi^2$  results.

20. *ROOMS* - Tables 3a & 3b give the actual and imputed values for the two imputation methods for the number of rooms variable, collapsing rooms of 8 and above together. Examination of the bodies of  $\chi^2$  shows that while the NN is most likely to imputed a value close to the true one it was not as good as hot deck at imputing the correct value.

21. The ability to impute the correct distribution rather than the true value is our main criteria and here again the hot deck performed better than the NN, as shown by the relative sizes of the  $\chi^2$ s (note these have 9 df rather than 7 as they were calculated on a fuller table). While the value for the hot deck was smaller than the NN value it was large enough to indicate that the method was not preserving the marginal distribution very well - however it may be expecting lot from an imputation method that it does this.

22. *MARITAL STATUS* - The results for the NN and hot deck methods are given in Tables 4a & 4b. For this variable the NN had a much lower  $\chi^2$  value than that of the hot deck method so it outperformed the hot deck, although both methods produced significant test statistics.

23. The original NN imputation had a  $\chi^2$  that was only just significant at the 5% point, indicating some preservation of the original distribution, however it was not as good as the hot deck at predicting the true value.

24. The overall results in Table 5 show that for the household variables: BUILDING TYPE and ROOMS, the hot-deck method preserves the marginal distributions better than the neural approach (Note the improved neural approach performs better than the previous neural imputation) The imputation of the person variables now generally matches the performance of the hot-deck, with one notable exception, AGE which is a key variable, here there has been a dramatic reduction in its ability to preserve marginal distributions. This could be the cost of the improvement in the consistency of the neural approach.

25. Examination of the cross-tabulations of the actual and imputed values for each variable shows that the ability of the neural network to predict the true values has improved since the previous results but it is still not as good as the hot-deck. On this evidence the hot deck performed better than the NN, however it still did not preserve the marginal distributions very well.

**Table 3a. NUMBER OF ROOMS - IMPROVED NEURAL NETWORK IMPUTATION**

		IMPUTED VALUE								Total
		1	2	3	4	5	6	7	8+	
ACTUAL VALUE	1	239	133	119	93	84	36	20	15	739
	2	173	204	304	246	196	85	46	49	1303
	3	248	340	515	518	403	253	103	291	2456
	4	221	363	611	738	656	465	264	299	3508
	5	137	231	417	657	780	632	363	171	3508
	6	59	92	242	406	468	545	359	100	2470
	7	13	36	68	119	169	181	110	41	867
	8+	17	20	49	70	87	95	130	36	645
Total		1107	1419	2325	2847	2843	2292	1395	1268	15496

$$\frac{2}{9} = 640.7$$

**Table 2b. NUMBER OF ROOMS - HOT DECK IMPUTATION**

		IMPUTED VALUE								Total
		1	2	3	4	5	6	7	8+	
ACTUAL VALUE	1	342	133	120	88	33	9	6	8	739
	2	84	427	371	237	115	52	9	8	1303
	3	81	311	921	683	294	112	31	23	2456
	4	49	195	702	1410	710	310	83	49	3508
	5	21	92	306	708	1379	746	175	81	3508
	6	10	31	112	323	698	896	277	123	2470
	7	4	8	33	82	183	244	169	144	867
	8+	1	10	19	51	96	133	129	206	645
Total		592	1207	2584	3582	3508	2502	879	642	15496

$$\frac{2}{9} = 47.7$$

**Table 4a. MARITAL STATUS - NEURAL NETWORK IMPUTATION**

		IMPUTED VALUE					Total
		Single	Married	Remarried	Divorced	Widowed	
TRUE VALUE	Single	2207	467	50	99	59	2882
	Married	581	1216	133	127	152	2209
	Remarried	88	161	22	24	23	318
	Divorced	157	150	31	41	26	405
	Widowed	64	191	35	30	173	493
Total		3097	2185	271	321	433	6307

$$\frac{2}{4} = 41.2$$

**Table 3b. MARITAL STATUS - HOT DECK IMPUTATION**

		IMPUTED VALUE					Total
		Single	Married	Remarried	Divorced	Widowed	
TRUE VALUE	Single	2160	256	137	229	200	2882
	Married	479	1442	185	60	43	2209
	Remarried	73	184	38	9	14	318
	Divorced	221	42	9	76	57	405
	Widowed	154	33	1	51	254	493
Total		3087	1957	270	425	568	6307

$$\frac{2}{4} = 89.8$$

**Table 5. MARGINAL DISTRIBUTIONS IMPROVED NN DATA**

	IMPROVED NN	ORIGINAL NN	HOT DECK
BUILDING TYPE	31.91*	1015.7*	7.3
NUMBER OF ROOMS	640.6*	1341.6*	47.6*
AGE	610.0*	41.0*	28.1*
MARITAL STATUS	41.2*	11.3*	89.8*
PRIMARY ACTIVITY LAST WEEK	391.8*	921.2*	395.7*
COUNTRY OF BIRTH	49.7*	18.2	46.7*

\* - indicates significance at the 5% level

### Consistency of imputed values

26. The original neural imputation approach suffered from the problem of imputing different values of household variables for members of the same household. This problem has now been solved.

27. While the above aspect of the inconsistency was relatively easy to fix, a more important problem with the original neural imputation process was that consistency with the edit rules was not preserved in the imputed data. This is still a problem with the improved method for example the neural network is likely to impute 2 or more rooms when the type of accommodation shows there should only be one room (Table 6).

**Table 6. NEURAL NETWORK IMPUTATION OF ROOMS**

	Part of converted/shared accommodation				Not part of converted/shared accommodation	Total
	1 room, self contained	1 room not self contained	2+ rooms, self contained	2+ rooms, not self contained		
ROOMS 1	52	142	225	33	655	1107
2	29	67	248	37	1038	1419
3	23	60	349	35	1858	2325
4	22	33	367	32	2393	2847
5+	28	52	525	53	7140	7798
Total	154	354	1714	190	13084	15496

28. The previous neural imputation results showed that there was a great deal of inconsistency between the age and activity last week variables. The degree of inconsistency has been markedly reduced in the improved imputation although there are still 13% of cases where activity last week is 'no code required' who

have an age of 16 years and over imputed, this is shown in Table 7. The hot-deck imputation method is designed to provide consistent imputations so in this respect it outperforms the neural approach.

**Table 7. CONSISTENCY - AGE IMPUTED BY NEURAL NETWORK, ACTIVITY LAST WEEK PRESENT.**

29. Shaded areas indicate inconsistencies.

		ACTIVITY LAST WEEK				No code required	Total
		Working F/T	Working P/T	Retired	Other activity		
AGE GROUP	0- 9	27	4	0	59	552	642
	10-15	62	3	0	98	200	363
	16-19	100	7	2	94	51	254
	20-29	396	39	5	278	51	769
	30-59	756	167	87	545	16	1571
	60+	81	54	528	149	1	813
Total		1422	274	622	1223	870	4411

### Preservation of joint distributions

30. This property of imputation is closely connected to that of consistency and given the poor performance of the NN method when evaluating consistency the examination of how well joint distributions were preserved was not carried out in depth.

31. The approach was to calculate the usual  $\chi^2$  test of association between each of the imputed variables and all the other variables in turn, to rank according to the size of the  $\chi^2$  and then to choose a range of good, medium and poor associations to test.

A number of cases were examined and the results indicate that for the household variables the hot deck preserved the relationships better than the NN and for the person variables, apart from those concerning marital status.

## V. DISCUSSION OF THE STATISTICAL EVALUATION

32. The hot deck is superior to the NN method in maintaining the distributions and consistency of the data. The inconsistency of the NN is particularly worrying because as a model based method is supposed to be capable of recognising complex interdependencies present in the data. The test data is fully consistent with the edit rules but it does not appear to be recognising these fairly obvious

relationships at all. If only a few inconsistencies were present in data and the NN was most likely to impute a consistent value, the inconsistencies could probably be dealt with by applying the edit rules and re-imputing the data. However given the large number of edit failures this is not a practical alternative as there is no guarantee the method would ever converge to provide a consistent data base.

33. On the other hand, the NN does give fairly consistent imputations of marital status with age and this does point to an ability to recognise relationships in the data, thus the inconsistencies observed are surprising. Perhaps the data was not sufficiently rich in the associations for the NN to recognise the relationships - again surprising given the amount of data consistent with the edit rules.

34. On the basis of the statistical evaluation the neural network imputation method is inferior to the 1991 hot deck approach and we cannot recommend that it be considered as an option for the 2001 Census. We will devote our resources into improving and developing the hot deck/donor method; a prototype system shows considerable improvement on the 1991 system - even maintaining marginal distributions.

35. We still believe there is merit in the neural approach and have included further work within a proposed imputation research program to be carried out jointly with the University of Southampton, aimed at providing general theoretical and practical guidelines for imputation.

## REFERENCES

Brant J. D. & Chalk S. M. The use of automatic data editing in the 1981 Census. *JRSSA*, **148**, 126 - 146.

Census Report for Great Britain, 1991 (Part 1).

Cheng B & Titterton, D.M (1994) Neural Networks: A review from a statistical perspective. *Statistical Science*, **9**, 2-54.

Fellegi, I.P & Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, **71**, 17 - 35.

Neural Technologies Limited (1996). Final Report from Phase 1 of the Neural Imputation Trial.

Neural Technologies Limited (1997). Final Report from Phase 2 of the Neural Imputation Trial

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, **42**, 412.

## ANNEX

### WHAT ARE NEURAL NETWORKS

Professor Ray Chambers, University of Southampton

1. Neural networks are a class of highly non-parametric regression methods that have proved useful in classification and prediction problems where standard parametric methods are inappropriate. Although originally patterned after the processes assumed to underpin human learning, neural networks are essentially a class of adaptive statistical methods for computer-based pattern recognition which attempt to emulate (at least in theory) the way humans recognise patterns.
  
2. The basic idea behind the neural network (NN) is the creation of a ‘fuzzy’ mapping from an input data set consisting of a collection of vectors  $\{x_1, \dots, x_n\}$  to an output data set  $\{y_1, \dots, y_m\}$ , where the input variable  $x_i$  represents the information available for classifying the  $i^{\text{th}}$  item of a ‘training set’ and the output variable  $y_i$  represents the ‘true’ classification of that item. The mapping is ‘fuzzy’ because it does not recover the actual value  $y_i$  in general. Instead, if we represent the outcome of the NN mapping by the  $\hat{y}(x_i)$ , then the NN is constructed so that the differences between the  $\hat{y}(x_i)$  and the  $y_i$  in the training set are ‘small’ according to some appropriately chosen criteria. Once constructed the value  $\hat{y}(x_j)$  obtained by applying the NN to a value  $x_j$  for which  $y_j$  is unknown can then be used to predict this unknown value.
  
3. Unlike standard statistical classification algorithms, which typically make assumptions about the conditional distribution of  $y$  given  $x$ , in order to arrive at an ‘optimal’ predictor, NN methods proceed in an essentially nonparametric fashion, building up the classification from the information about the relationship between  $y$  and  $x$  contained in the training data set. In general, the resulting classifier can be represented as a nested sequence of ‘coupled’ linear and nonlinear transformations of the input data:

$$\hat{y}(x_j) = F_{\text{out}} \left( \sum_{k_L=1}^{K_L} u_{k_L} F_{k_L} \left( \dots \sum_{k_2=1}^{K_2} u_{k_2} F_{k_2} \left( \sum_{k_1=1}^{K_1} u_{k_1} F_{k_1} \left( \sum_{i=1}^n w_{ij} F_{\text{in}}(x_i) \right) \right) \right) \right)$$

That is, a NN consists of an initial transformation of the training data ( $F_{\text{in}}$ ), followed by a set of linear transformations of the resulting values characterised by a set of weights  $\{w_{ij}\}$  which depend on the ‘new’ value  $x_j$ , followed by a set of nonlinear transformations ( $F_{k_1}$ ) corresponding to the first ‘hidden layer’ of the NN, followed by a second set of linear transformations characterised by weights  $\{u_{k_1}\}$  followed by a second nonlinear transformation ( $F_{k_2}$ ) corresponding to the second hidden layer’ of the NN and so on. In general there is a sequence of  $L$  such hidden layers. At the end, there is a final output transformation  $F_{\text{out}}$  which gives the NN ‘prediction’ of  $y_j$ .

4. The number of hidden layers ( $L$ ), the number and type of transformations involved in the  $l^{\text{th}}$  hidden layer ( $K_l$ ), the values of the weights  $\{u_{k_l}\}$  used to modify the results of these transformations and the weights  $\{w_{ij}\}$  applied to input data can all (in theory) be determined by a numerical search procedure which attempts to optimise the classifying performance of the NN with respect to the optimality criterion. This process is usually referred to as ‘training’ the NN. The initial transformation of the input data is usually necessary to ensure that the NN has acceptable operating characteristics (typically referred to as pre-processing), as is the post-processing transformation  $F_{\text{out}}$  which ensures that the output  $\hat{y}(\mathbf{x}_j)$  typically is a probability distribution, corresponding to the NN prediction of the conditional density of  $y_j$  given  $x_j$  and the training data.

5. The structure of a NN mimics the connectivity of the neuron structure in a human brain, and consequently enables it to identify highly complex nonlinear patterns in the  $x - y$  relationship. Furthermore, these patterns are identified in a rather automatic way, in the sense that the intrinsic complexity of the NN allows it to ‘represent’ many more distinct patterns than conventional statistical methods, and consequently provided the training data set is sufficiently rich in such patterns, the numerical optimisation methods used to ‘train’ a NN will create ‘pathways’ within the NN for each unique pattern. Finally, because unique  $x$ -patterns will typically be associated with different values of  $y$ , these optimisation methods allow the identification of those  $y$  values ‘most likely’ to arise from a particular  $x$ -pattern, resulting in the development of a nonparametric estimate of the conditional distribution of  $y$  given  $x$ .

6. Unfortunately, current developments in NN technology do not allow the completely ‘automatic’ prediction process implied in the preceding paragraphs. A considerable amount of human intervention is still required, in order to specify the types of nonlinear transformations used in pre and post-processing of the training data, and in the specification of the type and number of transformations used in the NN’s ‘hidden layers’. This means that NNs are still some way off offering a completely automatic solution to pattern recognition and prediction problems. However, with the types of hidden layer transformations available at present (e.g. perceptions, radial basis functions) they are capable of handling quite complex data structures, and so provide a viable alternative to statistical methods for pattern recognition which either assume some underlying parametric structure (e.g. multinomial logistic models), or are based on alternative paradigms for identifying structure (e.g. binary segmentation or tree-based models).

7. For the application of interest in this paper (imputation for missing data in the 2001 Census). Neural Technologies Ltd (NTL) were commissioned to create and train a NN using their proprietary AMAN software to impute for the types of data typically found in the census records. In this case, the training data consisted of a set of ‘complete’ census data records and the aim was to develop a NN which could identify the relationships between the records and hence to impute the values of missing items in a ‘test’ data set consisting of comparable records but with missing values for different fields.