

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 - OPERATIONAL EVALUATION

Submitted by the Office for National Statistics, United Kingdom¹

¹ Prepared by Jan Thomas and Pat Hostler.

I. INTRODUCTION

1. This paper should be read in conjunction with 'Neural Network Imputation - Statistical Evaluation' also presented to this work session, and it describes the operational aspects of the neural imputation trial which was carried out jointly between the Office for National Statistics, (ONS), and specialist contractors, Neural Technologies Ltd, (NTL), of Petersfield, Hants.

2. Because staff in ONS did not have neural network expertise, the Department of Trade and Industry, (DTI), was consulted which had been running a neural network awareness campaign in the U.K, for advice in finding someone to help with the evaluation. An independent expert, Dr Jim Austin, from the University of York, was appointed to give a technical and operational assessment. Dr Austin was provided with all the reports produced by NTL, and met with them to discuss issues arising.

II. HOW THE TRIAL WAS CONDUCTED

3. The following modules were developed for the trial:

(a) **Operational Imputation** - Determining the feasibility and practicality of a neural solution to imputation in an operational environment. This involved scoping the computing requirements for the 2001 Census.

(b) **Optimal Development of Neural Models** - Addressing the question of when a new neural model should be generated and how the process can be automated in an operational environment.

(c) **Priority of Fields** - Considering whether the most or least important field should be imputed first. It also considered the best neural approach for multiple field imputation and how this could be implemented operationally.

(d) **Dummy Imputation System** - This used the results from the above modules in a software prototype for the operational imputation system. Imputation is performed on real census data, with a more realistic pattern of missing values. The dummy imputation system operates on data automatically, using a pre-defined number of pre-trained neural models. It produces a complete valid dataset as output.

III. MAJOR CONSIDERATIONS

4. In order to conclude whether a neural network solution was feasible operationally, we needed to know:

- What hardware and software would be needed?
- How the neural network system would handle census data volumes?
- How long the system would take to run?
- How data would have to be stored and made suitable for neural network processing?
- How the neural solution would fit in with the IS Strategy?
- How the use of a neural solution would impact on the processing flow?
- The risks and issues.

These matters are considered in the following paragraphs.

What hardware and software would be needed?

5. The neural imputation system can be implemented on any platform which supports C/C++ code. This includes the ONS Strategic platforms of UNIX and Windows NT. Because of the amount of computation needed, UNIX workstations are preferred on performance grounds. However UNIX systems are more costly to develop due to the experience needed.

6. The neural system could extract information from the major standard database formats, including Oracle. Before the data can be shown to the neural model it has to be pre-processed into a form suitable for neural modelling. The choice of database for the 2001 Census is unlikely to create a problem because data could be extracted into a format suitable for the neural system, as happened in the trial. Additionally, for the trial, we added grid references to the data, using postcodes, prior to pre-processing. The generation of neural models depends on the geographic splitting of data.

How the neural network system would handle Census data volumes?

Scalability for 2001

7. The major recommendation Dr Austin made was that during the first phase of any contract with a supplier for a 2001 neural imputation system, we would need to undertake a phase during which the supplier would need to demonstrate the processing ability on a dataset of the same size as that likely to be used in the imputation system. This could be a county, if that is deemed to be a suitable dataset, or a smaller or larger area. Throughput is the critical factor for census processing systems, and this is the only way to be certain that a neural solution would achieve the processing rates required.

Processing Power

8. The major problem with the report of the trial is that the scalability of the approach is not proven. The estimates of processing power and time taken to process data assume that they are linearly related to the number of records but there is no proof to support this. The critical part of the neural network solution is the time it takes to train the network. The actual imputation of missing variables takes only about 1% of the training time.

How long would the system take to run?

9. NTL estimated that with the processing power of 74 distributed PCs over two weeks, or 19 distributed PCs over 2 months, would be sufficient to train the networks and impute missing variables. The nature of neural networks makes it very difficult to estimate the power and number of PCs needed. However, by 2001, it is likely that computing performance will have at least doubled, based on recent advances. For scaling up purposes, it was assumed that the 2001 Census will require 76 million, person and household, records to be processed, and 8 million imputations performed. For these volumes it was estimated that nearly 5 GB of total core data storage space will be required. There would not be likely to be a problem with disc storage given the current cost and capacity of discs.

10. A percentage processor utilisation figure was used to reflect the overhead processing encountered and the processing dependencies that occur with larger volumes of data and distributed processing (e.g. network communications overhead).

11. The hot deck system used in the 1991 Census processed data as and when they became available and this system took five elapsed months to complete the task. A neural solution could provide benefits in reducing the processing timetable and allowing for earlier publication of results.

Neural Building Blocks

12. The key factor in determining the processing power is the building block for creating neural models. The estimates were based on the time taken to build a neural system to impute six variables for two local government districts. The data were initially split into small batches, equivalent to that for a ward, and networks were created for each small batch. This initial division was primarily to provide a starting point, so that the model could merge or subdivide these areas if appropriate. If ward is chosen as the block to be the imputation building brick in 2001 and if these two wards are fairly representative, the estimates are probably soundly based and reflect the processing power required in 2001. Additionally, the estimates were based on trickle training, i.e. training as data arrives. If the final imputation system trains in batch, NTL estimated that the training would be 3-4 times faster and hence reduce the processing power and time needed. It was recommended that if we were to go ahead with the procurement of a neural system, then there must be a first phase which demonstrates the processing ability on data sets for the 2001 building brick or, if undecided, potential building bricks.

System performance with different thresholds

13. The neural models are built according to geographic areas and the setting of threshold parameters. One parameter determines when there is sufficient data in an area to start training, and another determines when the network no longer fits the data and needs to be split into smaller geographic areas. If these parameters are incorrectly set you could end up with systems which do not model the problem properly, or which require much splitting and re-training, resulting in a slower system.

14. The interaction between the census database and the neural network system would need very careful design as the transfer speed is critical to the speed of operation of the whole system. The likelihood is that for census processing, imputation would be delayed until all, or practically all, of the data for a building brick has been processed. In which case, the setting of thresholds determining when training is to start will be simple and will lead to little, if any, retraining. This could be confirmed during the first phase of the procurement.

System performance with multiple missing inputs

15. The second trial used data with up to 6 variables missing. Since the trial we have analysed 1991 data for the two local government districts and found that 75% of data had only one variable missing, and less than 10% had more than 3 variables missing, with less than 1% having more than 6 missing. If this pattern is maintained in 2001 the performance is not

likely to be adversely affected by multiple missing inputs and the estimates from the trial are probably realistic. Again, this would need to be confirmed during the first phase of any procurement.

How data would have to be stored and made suitable for neural network processing?

16. Before the data can be shown to the neural model it has to be pre-processed into a form suitable for neural modelling. The choice of database for the 2001 Census is unlikely to create a problem because data could be extracted into a format suitable for the neural system, as happened in the trial. Additionally for the trial we added grid references to the data, using postcodes, prior to pre-processing, as the generation of neural models depends on the geographic splitting of data.

How would the neural solution fit in with IS Strategy

17. The suggested solution is compatible with the IS Strategy, however, software support would be crucial throughout the processing cycle. Any contract with a supplier would need to include an agreement whereby we would get the source code and documentation if that supplier went out of business. However it is very unlikely that the software could be picked up and used by people unfamiliar with its specialist nature.

How the use of a neural solution would impact on the processing flow

18. The neural network imputation system would be a 'black box' into which all data was sent, with complete data being used for training or validation of the model, and incomplete data being imputed. The difficulties would be in controlling the flow of data to and from the system, and in monitoring what was happening.

The risks and issues for Neural Imputation

19. **Transparency** - The contractors use a proprietary method and are very reluctant to disclose how their the neural software engine operates. This would give problems for the ONS in explaining to our customers how imputation was performed.

20. **Flexibility** - One of the criticisms of the 1991 Hotdeck system was that it was inflexible and changes were difficult to make particularly at late stages. It may be necessary to make such changes once the 2001 data are being processed for imputation. A neural solution would allow for retraining of the models to accommodate unexpected changes in the data and, in addition, a neural solution could process data either in blocks or in a trickle feed mode, one record at a time.

21. **Tendering the contract** - Any contract for the 2001 Census would go to open tender. The DTI consider it likely that other larger companies would move in at this point, although Dr Austin considers that there are only a few companies capable of doing the job. Any other supplier would probably not accept NTL results, which are based on their proprietary method, and would insist on their own trials. This would increase the cost and ONS would then have to ensure that any neural model developed, delivered what was needed. In addition, there are very few known solutions to the multiple imputation problem and ONS would have to look very closely at solutions proposed.

IV. CONCLUSIONS

22. The results of the operational evaluation did not give any clear indication that the neural network approach was not viable operationally. The main problem lies with whether the volume of data examine in the trial was enough to show the proposed system could cope with Census data. Any procurement exercise would have to involve a demonstration of the neural approach to cope with a larger data set.

23. Dr Austin was impressed by the neural imputation solution and thought it performed well in the tests in the report and believed that it would be possible to get the system to the required performance, but that additional work would be needed to achieve this.

24. However on the statistical side, the system failed to demonstrate that it is superior to the 1991 hot-deck method. The statistical evaluation criteria were developed to objectively compare the hot-deck and neural imputation approaches and NTL were awarded the contract for this work on the basis of a competitive tendering process. On this basis we could not recommend that the neural imputation approach be adopted for the 2001 Census.

References

Neural Imputation Operational Report (1997) - Dr Jim Austin - University of York.

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

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

For further details please contact:

Jan Thomas
Office for National Statistics,
Segensworth Road,
Titchfield,
Fareham,
Hants.
U.K.
PO15 5RR.

Telephone 00 44 (0)1329 813296.
E-mail: jan.thomas@ons.gov.uk.