



# Using Artificial Intelligence to Improve Real World Data Estimation

## Introduction

Missing data is a common problem in large data sets and can cause serious problems in data analysis<sup>[1]</sup>. To help combat these issues statisticians use a set of mathematical techniques called data imputation.

Data imputation uses existing values to calculate a mathematical estimate for the missing value. The most basic form of these is generalised imputation which is an average of the known values in the data set.

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|-------|-------|-------|-------|-------|
| 25    | 3.5   | ?     | 5000  | -3.5  |
| ?     | 6.9   | 5.6   | ?     | 0.5   |
| 45    | 3.6   | 9.5   | 1500  | 46.5  |
| 27    | 9.7   | ?     | 3000  | ?     |

Fig. 1 Database with missing data; Enders Craig K., Applied Missing Data Analysis. Guilford Press, 2010

For example 3166.67 could be substituted for the missing value in column  $x_4$  [Fig. 1] as it's the mean average of the known values in that column.

Self Organising Maps (SOMs) are an Artificial Intelligence technique capable of finding logical groupings within data<sup>[2]</sup>.

This project will investigate if, by grouping the data into smaller data sets, the SOM is capable of reducing the amount of irrelevant data included in the generalised imputation calculation, leading to greater accuracy.

Jamie McBride P143814

University of Wales Trinity St. David, Mount Pleasant, Swansea, SA1 6ED

## Aim

Investigate the effectiveness of Kohonen Self Organising Maps in improving the accuracy of mean, mode and median generalised imputation methods for missing data in large data sets.

## Methodology

To achieve this aim anonymized MOT data will be sourced from the DVLA under the terms of the Open Government Licence<sup>[Fig. 2]</sup>.

A SOM program will be written, then trained and evaluated using this data. To evaluate the SOM's effectiveness in improving imputation the accuracy of its results will be compared against that of standard generalised imputation.

Their accuracy will be measured in terms of their difference from the correct value and, where possible, the base probability.



**Open Government Licence**  
for public sector information

Fig. 2 Open Government Licence; [Online]. Available: <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/> [Accessed 04 05 2016].

## Research Strategy

- Experimental

## Development Model

- Incremental  
[Fig 3]

## Expected Outcomes

- SOM Program
- Experimental Results
- Analysis of accuracy
- Assess potential for application to more complex data imputation methods

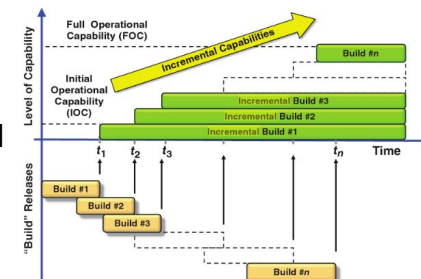


Fig. 3 Incremental development model concept; C. S. Wasson, System engineering analysis, design, and development : concepts, principles, and practices. Dec 2015

## Use

Missing data commonly affects research, possibly leading to invalid conclusions<sup>[3]</sup>. This can be disastrous particularly in fields like medicine. With increased focus on 'Big Data' and 'Data Mining' these improved techniques could add value.

## References

- [1] University of Houston, "Missing data: The Hidden Problem," [Online]. Available: <https://www.bauer.uh.edu/jhess/documents/2.pdf>. [Accessed 04 05 2016].
- [2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," ACM Comput. Surv., vol. 31, no. 3, pp. 264–323, Sep. 1999.
- [3] H. Kang, "The prevention and handling of the missing data.," Korean J. Anesthesiol., vol. 64, no. 5, pp. 402–6, May 2013.