H. WAHID, Q.P. HA, H. DUC. **New sampling scheme for neural network-based metamodelling with application to air pollutant estimation.** *Gerontechnology 2012;11(2):336;* doi:10.4017/gt.2012.11.02.325.00 **Purpose** A new method for the design of experiments (DOE) or sampling technique is proposed, using a distance weight function and the k-means theory. The radial basis function neural network metamodelling approach[1] is used to evaluate the performance of the proposed DOE by using an n-degree of test function, applied to the complex nonlinear problem of spatial distribution of air pollutants. A comparative study is included to analyse the performance of the proposed technique against available methods such as the n-level full fractional design method and the Latin Hypercube Design method. **Method** For one design objective and n number of input design variables, a set of input-output training dataset are $X = \{x_1^{(1)}, x_1^{(2)}, ..., x_1^{(i)}; ....; x_j^{(1)}, x_j^{(2)}, .., x_j^{(i)}) | i = 1,..,m,\ j = 1,..,n\}$ and $Y = \{y^{(1)}, y^{(2)}, ..., y^{(i)} | i = 1,2,...,m\}$, where m is the maximum number of the data points. Each data point has its own unique weight obtained from the distance factors between point p^i and a common reference point c, by using the Euclidean distance measure (i.e. $d_i(p^i, c)$). The weights represent the distinct patterns between each data point. A neighbour can be clustered as a group where the data point is taken as a candidate. To generalise the solution, the pairs of the input and output data points are combined to become the design space, given as $S = \{X; Y\}$.The solution can be simplified further if we set a common reference centre at the coordinate origin by firstly normalising the design space to $\hat{S} = [-1, 1]^{n+1}$. A list of distance weight values, $D = \{d_1, d_2, ...d_i | i = 1,2,...,m\}$, is then sorted and clustered using an available clustering algorithm. In this work, the k-means algorithm based on the Voronoi iteration[2] is used due to its fast computation especially in the 1-dimensional case. Here, the initial points are replicated randomly, so they can be expected to result in a global minimum solution. The maximum number of k corresponds to the number data points that will be sampled. **Results & Discussion** To initially validate the accuracy of the scheme, a known test function called 'Hock–Schittkowski Problem 100'[3] is used in which this nonlinear problem involving of 7 variables, 1 objective, and 4 constraints. A prepared dataset which generated randomly, is sampled at different sample size N, and then mapped using RBFNN-metamodel. An example of the estimated output for the case when the sample size is 30 percent of the full dataset is shows that the constructed metamodel is able to accurately approximate the true values at most of the points (*Figure 1*).

References

1. Wahid H, Ha QP, Duc H, Azzi M. Estimation of Background Ozone Temporal Profiles using Neural Networks. Proceedings of the 2011 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS 2011), Guangzhou; 2011; pp 292-297
2. MacKay D. Chapter 20. An Example Inference Task: Clustering. Information Theory, Inference and Learning Algorithms. Cambridge: Cambridge University Press; 2003; pp 284-292
3. Hock W, Schittowski K. Test Examples for Nonlinear Programming Codes. New York: Springer-Verlag; 1981; pp 1-177

Keywords: automation, sampling, metamodel, radial basis function neural network

Affiliation: University of Technology, Sydney, Australia; *E:* hbinwahi@eng.uts.edu.au

Full paper: doi:10.4017/gt.2012.11.02.325.710

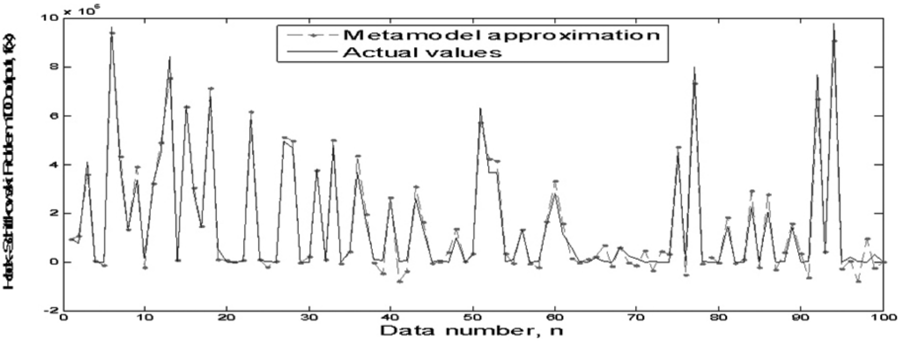

Figure 1. The estimation output for the test problem sampled at 30% of N_{full}