



# **Developing Artificial Neural Networks for Water Quality Modelling and Prediction**

**Robert James May**

BEng (Chem) Hons, BSc

Thesis submitted to The University of Adelaide  
School of Civil, Environmental & Mining  
Engineering in fulfilment of the requirements  
for the degree of Doctor of Philosophy

Copyright© October 2009.

# Abstract

Modelling water quality within complex, man-made and natural environmental systems can represent a challenge to practitioners. Many conventional modelling tools are not capable of representing the complexities of physical and chemical processes often observed in these systems. Consequently, there has been a great deal of interest in the application of computational intelligence techniques, such as artificial neural networks (ANNs). However, “black-box” approaches, such as ANN modelling, are often criticised due to a perceived lack of transparency in the model development methodology. This research has therefore focussed on improving the tools and techniques that are used in the development of ANN models for water quality prediction and forecasting.

The body of research presented in this thesis is described by several peer reviewed articles. These articles describe the theoretical basis and practical context for the ANN model development techniques that have been proposed and applied as a part of this research. Specifically, the ANN development framework has been further enhanced by this research through the development of novel approaches to perform two key tasks: input variable selection (IVS) and data splitting.

The IVS problem is to select variables as ANN inputs from a number of potential candidates, so as to minimise the number of inputs, but maximise the predictive performance of the model. A forward-selection approach for IVS has been examined that is based on partial mutual information (PMI), which can identify an optimal set of variables to use as inputs to ANN models, given a set of candidate variables. Of particular concern is that the use of MI in place of the more traditionally used correlation, provides a more appropriate basis for the selection of inputs based on non-linear relevance. Moreover, the accuracy of MI estimates for a given sample size is difficult to determine. Quantifying the accuracy of MI estimates is necessary to determine critical values of MI, since this forms the basis for of the termination criterion that stops the forward selection process.

Novel termination criteria were developed that alternatively determine the optimum number of candidate input variables. In comparison to the existing ap-

proach, which is based on a computationally expensive, yet potentially inaccurate bootstrap approach, the alternative criteria were found to both reduce the computational requirements and increase selection accuracy of the PMI-based IVS approach, resulting in a much improved algorithm.

Data splitting is an essential part of ANN model development, as the available modelling data must be partitioned into subsets for training, testing and validation. Depending on the data splitting method employed, the data split can have a significant effect on model performance, or reduce confidence in performance assessment. A popular method based on clustering of the self-organizing map (SOM) was examined. The approach was found to be sensitive to SOM size and the manner in which samples are drawn from within the SOM units. However, despite an optimal number of partitions, the SOM can generate partitions that are non-uniformly distributed, and which differ in size and shape. Although conventional rules to increase the sampling rate within larger clusters can reduce variance, the remaining variance can still be significant.

A hybrid algorithm called SOMPLEX was developed, which combines clustering on the SOM, and the DUPLEX algorithm used to perform intra-cluster sampling. DUPLEX is a fully deterministic algorithm that generates a representative sample, regardless of the size or distribution of data within a SOM cluster. For several example applications to predicting water quality, SOMPLEX was found to generate representative data for training, testing and validation, with no variation. The hybrid SOMPLEX approach combines the strengths of the two individual data splitting algorithms, in that the clustering on the SOM reduces the operational complexity, and the DUPLEX sampling improves on random sampling of SOM units to reduce sample variability and increase the representativeness of datasets generated.

In terms of the overall ANN development framework, the outcomes of this research have been an increased understanding of how to best implement ANN techniques, and an appreciation for their place within the context of a water quality modelling toolkit, which comprises both conventional and non-conventional modelling approaches. It was also observed that although the ANN modelling paradigm is quite powerful, it is not without limitations. Many of the limitations and problems encountered with ANN model development are more indicative of the application, rather than the modelling approach itself.

# Statement of Originality

I, *Robert James May*, hereby declare that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, the Australasian Digital Theses Program (ADTP) and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed: ..... Date: .....

# Acknowledgement

Above all, I wish to express my profound gratitude towards my supervisors, Professor Graeme Dandy and Professor Holger Maier, for their invaluable guidance, support and encouragement throughout the course of this research.

The author is grateful for the financial support provided by the CRC for Water Quality and Treatment, the technical assistance from the research team for Project 2.5.0.1, and the project leadership of Mr Michael Holmes and Dr Christopher Chow. I thank United Water for their support of this research, in particular Dr John Nixon for his contribution as my industry supervisor, and Dr Stephanie Rinck-Pfeiffer, for her patience while I completed this thesis.

Thanks to the all the staff within the School of Civil, Environmental & Mining Engineering, in particular Dr Stephen Carr for his assistance with software development. The author also wishes to acknowledge that a significant part of this research was aided by the facilities provided by the South Australian Partnership for Advanced Computing.

Many thanks to all of my fellow post-graduate students within the School, for the Friday afternoon drinks and carpet bowls—I blame you all for making the experience so enjoyable that I lost all motivation to finish. Thanks especially to Ms. Gayani Fernando, for sharing a mutual love of neural networks; Dr Matthew Gibbs and Mr Darren Broad, for the more than occasional Tuesday drinks; and I am truly indebted to Mr Michael Leonard for a year of Saturdays spent writing up, and many other enjoyable discussions over the years.

Finally, I acknowledge that credit for any of my past, current and future achievements must be shared with my parents, brother, sister, and my partner, Shannon; for I could accomplish very little without their love and support.

*Robert May  
Adelaide, 14 October 2009*

# Contents

<b>Abstract</b>	<b>iii</b>
<b>Statement of Originality</b>	<b>v</b>
<b>Acknowledgement</b>	<b>vii</b>
<b>Contents</b>	<b>ix</b>
<b>Publications</b>	<b>xv</b>
<b>List of Figures</b>	<b>xvii</b>
<b>List of Tables</b>	<b>xix</b>
<b>List of Abbreviations</b>	<b>xxi</b>
<b>List of Symbols</b>	<b>xxiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Research Background . . . . .	1
1.2 Artificial Neural Networks . . . . .	3
1.3 Research Objectives . . . . .	6
1.4 Thesis Structure . . . . .	12
<b>2 Input Variable Selection</b>	<b>15</b>
2.1 Introduction . . . . .	15
2.2 The Input Variable Selection Problem . . . . .	16
2.3 Strategies and Algorithms . . . . .	20
2.3.1 Optimality Criteria . . . . .	21
2.3.2 Search Strategies . . . . .	26
2.3.3 Dimensionality Reduction . . . . .	30
2.3.4 Wrappers . . . . .	33
2.3.5 Filters . . . . .	36
2.4 Comparison of Approaches . . . . .	44

<b>3</b>	<b>Data Splitting</b>	<b>49</b>
3.1	Introduction . . . . .	49
3.2	Generalisation and Over-fitting . . . . .	50
3.2.1	Cross-validation . . . . .	51
3.2.2	Ensemble Training . . . . .	54
3.2.3	Regularisation . . . . .	55
3.3	The Hold-out Bias and Variance Dilemma . . . . .	56
3.4	Sampling Techniques . . . . .	59
3.4.1	Probability Sampling . . . . .	59
3.4.2	Non-probability Sampling . . . . .	68
3.5	Comparison of Approaches . . . . .	74
<b>4</b>	<b>Synopsis of Publications</b>	<b>79</b>
<b>5</b>	<b>Publication 1: Critical values of mutual information</b>	<b>87</b>
5.1	Introduction . . . . .	93
5.2	Preliminaries . . . . .	94
5.2.1	Estimation of Mutual Information . . . . .	94
5.2.2	Distribution of Mutual Information . . . . .	95
5.3	Determining Critical Values of Mutual Information . . . . .	97
5.3.1	Methodology . . . . .	97
5.3.2	Approximate Distribution of Mutual Information . . . . .	98
5.3.3	Critical Values . . . . .	98
5.4	Example Application . . . . .	100
5.4.1	Selection Algorithm . . . . .	100
5.4.2	Dataset . . . . .	101
5.4.3	Selection Results . . . . .	102
5.5	Concluding Remarks . . . . .	104
<b>6</b>	<b>Publication 2: Non-linear IVS for ANNs Using PMI</b>	<b>107</b>
6.1	Introduction . . . . .	113
6.2	Theoretical Overview . . . . .	114
6.2.1	Input variable selection techniques . . . . .	114
6.2.2	Estimation of partial mutual information . . . . .	116
6.2.3	Description of the PMIS algorithm . . . . .	121
6.3	Formulation of Alternative Termination Criteria . . . . .	123
6.3.1	Modified bootstrap . . . . .	123
6.3.2	Tabulated critical values . . . . .	124
6.3.3	AIC-based criterion . . . . .	125
6.3.4	Hampel test criterion . . . . .	127
6.4	Experimental Methods . . . . .	129
6.4.1	Comparison to IVS based on the correlation coefficient . . . . .	131

6.5	Results and Discussion . . . . .	132
6.5.1	Selection Accuracy . . . . .	133
6.5.2	Computational efficiency . . . . .	142
6.5.3	Linear versus non-linear input variable selection . . . . .	144
6.5.4	Effect of sample size . . . . .	145
6.6	Conclusions . . . . .	146
<b>7</b>	<b>Publication 3: Application of PMI to ANN Water Quality Forecasting</b>	<b>149</b>
7.1	Introduction . . . . .	155
7.2	Background . . . . .	157
7.3	Methodology . . . . .	160
7.3.1	Model architecture . . . . .	160
7.3.2	Input variable selection . . . . .	162
7.3.3	Data sampling . . . . .	164
7.3.4	GRNN training . . . . .	165
7.3.5	Performance criteria . . . . .	166
7.4	Cherry Hills–Brushy Plains WDS Example . . . . .	168
7.4.1	System Description . . . . .	168
7.4.2	Synthetic data generation . . . . .	169
7.4.3	Selected input variables . . . . .	170
7.4.4	Model performance . . . . .	170
7.5	Myponga WDS Example . . . . .	175
7.5.1	System Description . . . . .	177
7.5.2	Data collection and pre-processing . . . . .	177
7.5.3	Selected input variables . . . . .	178
7.5.4	Model performance . . . . .	181
7.6	Discussion . . . . .	182
7.6.1	Model parsimony . . . . .	182
7.6.2	Comparison of developmental frameworks . . . . .	184
7.6.3	Interpretability of forecasting models . . . . .	185
7.7	Conclusions . . . . .	187
<b>8</b>	<b>Publication 4: Data Splitting Using SOM-based Stratified Sampling</b>	<b>189</b>
8.1	Introduction . . . . .	195
8.2	Data Splitting Methods . . . . .	196
8.3	SOM-based Stratified Sampling . . . . .	199
8.3.1	Choice of Variables . . . . .	199
8.3.2	Location of Strata Boundaries . . . . .	200
8.3.3	Sample Allocation . . . . .	201
8.3.4	Number of Strata . . . . .	203
8.3.5	Proposed SBSS Algorithm . . . . .	207



8.4	Experimental Study Design . . . . .	207
8.4.1	Datasets . . . . .	208
8.4.2	Bias and Variance Estimation . . . . .	209
8.4.3	Neural Network Training . . . . .	210
8.4.4	Data Splitting Algorithms . . . . .	210
8.5	Results . . . . .	212
8.6	Discussion . . . . .	219
8.6.1	Factors influencing data splitting performance . . . . .	219
8.6.2	Selecting a suitable data splitting approach . . . . .	220
8.6.3	Specification of SOM Parameters . . . . .	222
8.6.4	Effect of SOM initialisation . . . . .	223
8.7	Conclusions . . . . .	223
<b>9</b>	<b>Publication 5: SOMPLEX: A hybrid SOM-DUPLEX data splitting algorithm</b>	<b>225</b>
9.1	Introduction . . . . .	231
9.2	Data Splitting Methods . . . . .	232
9.2.1	Uniform random . . . . .	233
9.2.2	Stratified . . . . .	234
9.2.3	Convenience . . . . .	235
9.2.4	Judgement . . . . .	236
9.2.5	Systematic . . . . .	236
9.2.6	Kennard-Stone . . . . .	236
9.2.7	Search-based . . . . .	237
9.2.8	Multi-stage . . . . .	237
9.3	The SOMPLEX Algorithm . . . . .	238
9.4	Methodology . . . . .	240
9.5	Datasets . . . . .	243
9.5.1	Pre-processing . . . . .	243
9.5.2	Coagulation . . . . .	244
9.5.3	Salinity . . . . .	244
9.5.4	Chlorine . . . . .	246
9.6	Results and Discussion . . . . .	246
9.7	Software . . . . .	252
9.8	Conclusions . . . . .	252
<b>10</b>	<b>Publication 6: Development of ANNs for Water Quality Modelling</b>	<b>255</b>
10.1	Introduction . . . . .	259
10.2	Applications in Water Quality Modelling . . . . .	260
10.2.1	Prediction and Forecasting . . . . .	260
10.2.2	Process control . . . . .	261

10.2.3	Integrated Modelling . . . . .	262
10.2.4	Metamodelling . . . . .	263
10.2.5	Knowledge Extraction . . . . .	263
10.3	Neural Architectures . . . . .	264
10.3.1	Multilayer Perceptron . . . . .	264
10.3.2	Generalised Regression Neural Network . . . . .	266
10.4	Model Development . . . . .	268
10.4.1	Data Collection . . . . .	268
10.4.2	Data Pre-processing . . . . .	271
10.4.3	Input Variable Selection . . . . .	274
10.4.4	Data Subset Selection . . . . .	281
10.4.5	Training . . . . .	289
10.4.6	Model Selection . . . . .	294
10.4.7	Validation . . . . .	295
10.5	Summary . . . . .	299
<b>11</b>	<b>Conclusions</b>	<b>301</b>
11.1	Contributions of Research . . . . .	302
11.1.1	Input Variable Selection . . . . .	302
11.1.2	Data Splitting . . . . .	304
11.1.3	Water Quality Forecasting . . . . .	306
11.1.4	Field Research . . . . .	308
11.1.5	Software . . . . .	308
11.2	Research Limitations . . . . .	308
11.3	Future Research . . . . .	310
	<b>References</b>	<b>313</b>
<b>A</b>	<b>Critical Values of <i>I</i> and <i>R</i></b>	<b>329</b>
<b>B</b>	<b>IVS Performance Data</b>	<b>337</b>

# Publications

## Book chapters

- May, R. J., H. R. Maier, and G. C. Dandy, Development of artificial neural networks for water quality modelling and analysis, in *Modelling of Pollutants in Complex Environmental Systems*, edited by G. Hanrahan, vol. 1, pp. 27–62, ILM Publications, London, UK, 2009.

## Journal articles

- May, R. J., G. C. Dandy, H. R. Maier, and T. M. K. G. Fernando, Nonlinear variable selection for artificial neural networks using partial information, *Environmental Modelling and Software*, 23, 1312–1326, 2008.
- May, R. J., H. R. Maier, G. C. Dandy, and J. B. Nixon, Application of partial mutual information-based variable selection to ANN forecasting of water quality within water distribution systems, *Environmental Modelling and Software*, 23, 1289–1299, 2008.
- May, R. J., H. R. Maier, and G. C. Dandy, Data Splitting for Artificial Neural Networks Using SOM-based Stratified Sampling, *Neural Networks*, 20, 283–294, 2010.

## Journal articles under review

- May, R. J., H. R. Maier, and G. C. Dandy, SOMPLEX: a hybrid SOM-DUPLEX data splitting algorithm for ANN development, *Submitted to Water Resources Research*

## Peer-reviewed conference articles

- May, R. J., G. C. Dandy, H. R. Maier, and T. M. K. G. Fernando, Critical values of a kernel-density based mutual information estimator, in *IEEE International Joint Conference on Neural Networks*, pp. 9997–10,002, Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, 2006.

# List of Figures

1.1	Biological neuron and the mathematical perceptron . . . . .	4
1.2	Framework for ANN model development . . . . .	9
2.1	Taxonomy of IVS algorithms . . . . .	22
2.2	Wrapper and filter IVS algorithm designs . . . . .	23
3.1	Phenomenon of over-fitting . . . . .	52
3.2	Stop training (early-stopping) using test data . . . . .	53
3.3	Taxonomy of sampling methods . . . . .	60
3.4	Multivariate stratification by cut-points and clustering . . . . .	66
4.1	Contribution of publications presented within this thesis . . . . .	80
5.1	Approximate distribution of MI estimator . . . . .	98
5.2	Critical values of the MI estimator . . . . .	100
5.3	PMIS for the ADD10 model . . . . .	103
5.4	Comparative run-time for selection with and without MCS . . . . .	104
6.1	Mutual information . . . . .	117
6.2	Partial mutual information . . . . .	120
6.3	Application of PMIS termination criteria for the AR9 time-series . . . . .	134
6.4	Performance of PCIS for linear data . . . . .	135
6.5	Performance of PMIS for linear data . . . . .	136
6.6	Performance of PCIS for non-linear data . . . . .	137
6.7	Performance of PMIS for non-linear data . . . . .	138
6.8	Computational requirement of PMIS with a bootstrap . . . . .	143
7.1	Conceptual approach to time-series regression using a historical window . . . . .	159
7.2	Architecture of the GRNN . . . . .	172
7.3	Cherry Hills–Brushy Plains WDS . . . . .	173
7.4	Forecast time-series at Node 36 of Cherry Hills–Brushy Hills WDS . . . . .	176

7.5	24-hour test and validation forecasts of free chlorine residual generated by Model B for an instance of training, test and validation data. . . . .	183
8.1	Partitioning of data on a $7 \times 5$ SOM . . . . .	200
8.2	Effect of SOM size on SBSS . . . . .	217
8.3	Silhouette and quantisation error versus SOM size . . . . .	218
8.4	Guidelines for choosing a sampling technique . . . . .	221
9.1	Taxonomy of sampling techniques . . . . .	233
9.2	Codebook vectors and resulting Voronoi regions for the partitioning of the Salinity dataset by a $1 \times 90$ SOM . . . . .	249
9.3	Training, test and validating data selected from the Salinity dataset using SOMPLEX . . . . .	251
10.1	Multi-layered perceptron . . . . .	265
10.2	The general regression neural network . . . . .	266
10.3	Framework for the development of ANN models . . . . .	269
10.4	The input variable selection task . . . . .	276
10.5	Wrapper and filter approach to selecting input variables . . . . .	277
10.6	Generalisation and over-fitting . . . . .	282
10.7	Stop training (early-stopping) using test data . . . . .	283
10.8	Stratified sampling based on the SOM . . . . .	286
10.9	Error surface with multiple local optima . . . . .	291

# List of Tables

1.1	Examples of ANN applications within the water resources field . . .	7
2.1	Curse of dimensionality . . . . .	19
2.2	Comparison of Various IVS algorithms . . . . .	47
3.1	Qualitative comparison of sampling methods . . . . .	75
5.1	Critical values of the KDE-based mutual information estimator . . .	99
6.1	Critical values of the KDE-based mutual information estimator (af- ter <i>May et al.</i> (2006)). . . . .	126
6.2	Benchmark data-generating models. . . . .	130
6.3	Summary of termination criteria . . . . .	139
7.1	Historical data for the Cherry Hills–Brush Plains WDS . . . . .	171
7.2	PMIS analysis for the Cherry-Hills—Brushy Plains WDS . . . . .	172
7.3	Variables for GRNN models of the Cherry Hills–Brushy Plains WDS	174
7.4	1-hour test forecasts of chlorine in the Cherry Hills—Brushy Plains WDS . . . . .	174
7.5	Validation forecasts within the Cherry Hills–Brushy Plains WDS . .	175
7.6	Historical data collected for the Myponga WDS . . . . .	179
7.7	Input variables for the Myponga WDS . . . . .	180
7.8	Variables for GRNN models of the Myponga WDS . . . . .	180
7.9	24-hour test forecasts of chlorine within the Myponga WDS . . . .	181
7.10	Validation forecasts of chlorine within the Myponga WDS . . . . .	182
8.1	Interpretation of the silhouette coefficient . . . . .	205
8.2	SOM parameters for implementing SBSS . . . . .	208
8.3	Generalisation error for uncorrelated datasets ( $(n/N) = 80\%$ ) . . .	213
8.4	Generalisation error for uncorrelated datasets ( $(n/N) = 80\%$ ) . . .	214
8.5	Generalisation error for correlated datasets ( $(n/N) = 40\%$ ) . . . .	215
8.6	Generalisation error for correlated datasets ( $(n/N) = 80\%$ ) . . . .	215
9.1	Specifications of the SOM . . . . .	240

9.2 Algorithms included in the comparative study . . . . .	243
9.3 Summary of modelling datasets used for the comparative study of data splitting algorithms . . . . .	245
9.4 Performance of data splitting algorithms on the water resources datasets . . . . .	247
A.1 Critical values of the KDE estimate $I(x_1; y)$ . . . . .	330
A.2 Critical values of the KDE estimate $I(x_1, x_2; y)$ . . . . .	331
A.3 Critical values of the KDE estimate $I(x_1, x_2, x_3; y)$ . . . . .	332
A.4 Critical values of the KDE estimate $I(x_1, x_2; y_1, y_2)$ . . . . .	333
A.5 Critical values of the KDE estimate $I(x_1, x_2, x_3, x_4; y)$ . . . . .	334
A.6 Critical values of the KDE estimate $I(x_1, x_2, x_3; y_1, y_2)$ . . . . .	335
A.7 Critical values of the Pearson correlation coefficient . . . . .	336
B.1 Model specifications for PMIS and PCIS (50-sample datasets) . . .	338
B.2 Model specifications for PMIS and PCIS (100-sample datasets) . . .	339
B.3 Model specifications for PMIS and PCIS (500-sample datasets) . . .	340
B.4 Model specifications for PMIS and PCIS (1000-sample datasets) . .	341

# List of Abbreviations

ACF	Auto-correlation function
ACO	Ant colony optimisation
AIC	Akaike Information Criterion
ANN	Artificial neural network
ARMA	Auto-regressive moving-average
ARX	Auto-regressive with exogenous inputs
BIC	Bayesian Information Criterion
BPA	Back-propagation algorithm
CV	Cross-validation
CVI	Cluster validity index
DBS	Density biased sampling
DSS	Data subset selection
EA	Evolutionary algorithm
EANN	Evolutionary neural network
GA	Genetic algorithm
GRIDA	GRNN input determination algorithm
GRNN	Generalised regression neural network
ICA	Independent component analysis
IMC	Inverse model control
IVS	Input variable selection
JMI	Joint mutual information
KDE	Kernel density estimation
MAD	Median absolute deviation from the median
MAE	Mean absolute error
MPC	Model predictive control
MCS	Monte Carlo simulation
MI	Mutual information
MIFS	Mutual information feature selection
MLP	Multi-layer perceptron
MR	Maximum relevance

*continued on next page*



mR	Minimum redundancy
mRMR	Minimum redundancy–maximum relevance
MRE	Mean relative error
MSE	Mean squared error
PACF	Partial auto-correlation function
PC	Principal component
PCA	Principal component analysis
pdf	Probability density function
PMI	Partial mutual information
PMIS	Partial mutual information-based selection
QE	Quantisation error
RI	Relative importance
RMSE	Root mean squared error
SBSS	SOM-based stratified sampling
SCE	Shuffled complex evolution
SOM	Self-organizing map
SRS	Simple random sampling
SVR	Single variable regression
UVA	Ultra-violet absorbance
VQ	Vector quantisation
WDS	Water distribution system

# List of Symbols

## General

$a(z_j)$	Activation function
$d$	Number of variable dimensions
$E$	Expectation
$f(z_j)$	Hidden node transfer function
$\hat{f}(x), \hat{f}(x, y)$	Density function estimate
$F(x)$	Process/model transfer function
$p$	Number of model parameters
$r$	Coefficient of determination
$R$	Linear correlation
$R'$	Partial correlation
$V$	Variance
$W$	ANN weight matrix
$w_i$	ANN weight
$X$	Random independent/ANN input variable
$x$	Realisation/observation of $X$
$\hat{x}$	Model estimate of $x$
$Y$	Dependent/ANN output variable
$y$	Realisation/observation of output variable $Y$
$\hat{y}$	Model estimate of $y$
$z_j$	training input vector (GRNN) / hidden node input (MLP)

## Input Variable Selection

$B$	Bootstrap size
$C$	Candidate input variable
$G$	Gaussian kernel function
$h$	Kernel bandwidth

*continued on next page*

$h_G$	Scott reference bandwidth
$I$	Mutual information
$I'$	Partial mutual information
$k$	Number of selected input variables
$K_h$	Kernel function
$p(x), p(x, y)$	Probability density function
$S$	Hat matrix defined by $\hat{y} = S(y)$

### Data Splitting

$C^{(i)}$	$i^{th}$ cluster
$\mathcal{D}$	Modelling data set
$H$	Number of strata
$k$	Number of SOM units = $m \times n$
$m$	Number of SOM rows
$N$	Length of dataset $D$
$n$	Number of sampled data / number of SOM columns
$n_S$	Number of training data
$n_T$	Number of test data
$n_V$	Number of validation data
$S$	Silhouette coefficient
$\mathcal{S}$	Test data set
$\mathcal{T}$	Training data set
$\mathcal{V}$	Validating data set

### Greek Symbols

$\varepsilon$	Error term
$\phi$	Partial auto-correlation
$\psi$	Golden Ratio ( $\sim 1.6$ )
$\sigma$	Standard deviation / GRNN smoothing parameter
$\mu$	Mean