

EXPLORING THE IMPACT OF DATA SPLITTING METHODS ON ARTIFICIAL NEURAL NETWORK MODELS

WENYAN WU (1), HOLGER R. MAIER (1), GRAEME C. DANDY (1) AND ROBERT MAY (1,2)

*(1): School of Civil, Environmental and Mining Engineering, University of Adelaide,
Adelaide, 5005, Australia*

*(1,2): Veolia Water Asia-Pacific, Technical Department, Network Management Team,
Shanghai, 200041, China*

Data splitting is an important step in the artificial neural network (ANN) development process whereby data is divided into training, test and validation subsets to ensure good generalization ability of the model. In previous research, guidelines on choosing a suitable data splitting method based on the dimensionality and distribution of the dataset were derived from results obtained using synthetic datasets. This study extends previous research by investigating the impact of three data splitting methods tested in previous research on the predictive performance of ANN models using real-world datasets. Three real-world water resources datasets with varying statistical properties are used. It has been found that the relationship between different data splitting methods and data with different statistical properties obtained in previous research using synthetic data also generally holds for real-world water resources data. However, some data splitting methods produce an optimistically low validation error due to the bias created by allocating the extreme observations to the training set.

INTRODUCTION

Artificial neural networks (ANNs) are a popular approach for developing water resources models. As ANN models are often developed using limited data, a key objective of their development is to ensure the generalization ability of the trained models so that accurate forecasts/predictions on previously unseen data can be obtained. This is often achieved by dividing the available data into training, test and validation sets. The training set is used to calibrate the model (optimize model parameters), the test set is used in cross-validation during the training process to avoid over-fitting and the validation set is used to test the performance of the trained model.

Previous research has found that the data splitting method used in the development process of an ANN model has a significant impact on the performance of the final model, although this step is not addressed adequately on many studies (Maier et al., 2010). May et al. (2010) evaluated and compared the reliability of several data splitting methods reported within ANN literature, for modeling datasets with different dimensionality and statistical properties. The study benchmarked methods against uniform random sampling, and compared the reduction in the variance of model performance that was achieved using each method. In their research, the authors found that the deterministic DUPLEX data splitting

method, was a good approach for multivariate datasets that are uniformly or normally distributed and low-dimensional datasets that are skewed. For low-dimensional (univariate or bivariate) datasets that are uniformly or normally distributed, the systematic method, which draws every k th sample from a random starting point, generated lower validation errors. A stratified data splitting method based on the combination of stratification using a self-organizing map (SOM) and random sampling using the Neyman allocation rule (or SBSS-N) generated better ANN models for highly skewed datasets or datasets with high dimensions. Based on these results, May et al. (2010) proposed guidelines on selecting a data splitting method, given the properties of a dataset. However, these guidelines were derived based on a small set of synthetic examples, and have yet to be tested on real-world datasets. Consequently, the objective of this study is to test if the relationship between different data splitting methods and data with different statistical properties obtained by May et al. (2010) for synthetic datasets also holds for the case of real-world datasets.

DATASETS

Murray River (Australia) salinity data

The salinity dataset was used as a case study by Bowden et al. (2002) to forecast water quality in the River Murray, South Australia, 14 days in advance. The original dataset includes a total of 2,028 daily observations of twelve variables, including stream flow and salinity at several locations along the River Murray, for the period from August 1992 to March 1998. Up to 26 lags of each variable are used in this study, resulting in a total of 416 potential inputs.

Myponga water distribution system (Australia) chlorine data

The chlorine dataset was used as a case study by Bowden et al. (2006) and May et al. (2008a) to forecast free chlorine levels at a downstream location in a water distribution system at Myponga to the south of Adelaide, South Australia. The original data include 2,773 hourly observations of eight chlorine and temperature variables at a number of locations along the pipeline for the period from November 2003 to July 2004. Up to 48 lags for each variable are used in this study, resulting in a total of 384 potential inputs.

Upper Neckar catchment (Germany) rainfall-runoff data

The rainfall-runoff dataset investigated in this study is from the upper Neckar catchment in South-West Germany. The original data were used by B´ardossy and Singh (2008) to estimate hydrological model parameters. For the purpose of this study, 3,651 daily observations of effective rainfall and runoff for the period of 1961 to 1970 are used. The task is to forecast runoff one day in advance using previous values of effective rainfall and runoff. Up to 10 lags of each variable are used, resulting in a total of 20 potential inputs.

EXPERIMENTAL METHODOLOGY

Following the methodology in May et al. (2010), data splitting is bootstrapped 100 times with an ANN model being trained and evaluated for each split to determine the distribution of model performance for each data splitting method. The impact of data splitting on ANN model development can be determined based on the average performance and variation in model performance, which reflects the ability of a data splitting method to consistently generate a representative split. The following section provides the details of the bootstrapped ANN model development process.

Input variable selection

In this study, the partial mutual information (PMI) based non-linear variable selection algorithm (Sharma, 2000; May et al., 2008b), combined with the Akaike Information Criterion (AIC) for stopping (May et al., 2008a), are used to select appropriate inputs for each dataset. The PMI based input selection method has been found to be an effective approach, due to its ability to capture both linear and non-linear relationships and take into account both input significance and independence (May et al., 2008b).

Data splitting methods

In this study, 60% of each dataset is used for calibration, 20% for testing and 20% for validation. This is achieved by using three different methods: the systematic method, the DUPLEX method and the SOM based stratified sampling with Neyman allocation or SBSS-N.

Systematic data splitting method

The systematic data splitting method is a semi-deterministic method, in which every k^{th} sample from a random starting point is allocated to the training, testing and validation datasets. In implementing systematic sampling in this study, the data are first ordered along the output variable dimension. Then the sampling interval is determined based on the training and test data proportions specified by the user. Thereafter, a starting point is randomly selected and training samples are drawn first, followed by the test samples. Finally, unsampled data are allocated into the validation set.

DUPLEX data splitting method

The DUPLEX data splitting method was developed by Snee (1977) based on CADEX or Kennard-Stone sampling (Kennard and Stone, 1969). The original DUPLEX algorithm was used to divide data into two sets, and proceeds as follows. The two points which are farthest apart are assigned to the first dataset, and the next pair of points that are farthest apart in the remaining data are assigned to the second dataset. Subsequently, data are sampled alternately into each sample in a pair-wise manner, such that they are farthest from any points already in the respective sample. This process is repeated until both datasets are filled. May et al. (2010) employed a modified form of the DUPLEX algorithm to generate

three datasets to be used as training, testing and validation datasets for ANN model development. DUPLEX is deterministic and therefore, it only produces a unique split for any given dataset.

SBSS-N data splitting method

The SBSS-N approach is a two-step data splitting method. In the first step, a SOM is used to partition the data into M relatively homogeneous groups or strata by learning the optimal distribution of the weight vectors in the SOM. In the second step, the Neyman allocation rule is used to select samples within each stratum for training, testing and validation. The number of samples to be taken from stratum m based on the Neyman allocation rule is expressed as:

$$n_m = \frac{N_m \sigma_m}{\sum_{i=1}^M N_i \sigma_i} \frac{n}{N} \quad (1)$$

where, N is the size of the dataset, n is the required sample size, N_i is the size of stratum i , σ_i is the intra-stratum standard deviation of stratum i . Based on this allocation rule, the sample allocation will be increased for strata of large size or having high variance. When training the SOM for each of the datasets investigated in this paper, a heuristic SOM grid size formula suggested by Vesanto (1999) and the SOM parameters used by May et al. (2010) are used. For full details of the SBSS algorithm refer to May et al. (2010).

ANN architecture and training

In this paper, the general regression neural network (GRNN) introduced by Specht (1991) is used. Compared to multilayer perceptrons (MLPs), which have been used more commonly in ANN applications in water resources (Maier et al., 2010), the architecture of GRNNs is fixed and there is only one parameter (the bandwidth) that needs to be obtained by calibration. Therefore, a GRNN model is much faster to develop (May et al., 2008a), which suits the purposes of this study. In this study, all of the GRNN models are trained using Brent's method. For more details on the GRNN used in this study, refer to May et al. (2008a).

Model performance evaluation

Three commonly employed statistical error and goodness-of-fit measures are used, in order to determine how well the predicted response matches different aspects of the measured response (e.g. peak, average etc.) and compare the models developed using different data splitting methods. These are namely: the root mean square error (RMSE), the mean absolute error (MAE) and the square of Pearson r (r^2).

RESULTS AND DISCUSSION

Statistical properties of the datasets

The selected inputs and the output for each dataset, as well as the statistics of these variables, are summarized in Table 1. As can be seen, two inputs are selected for the Murray River salinity dataset and the Neckar catchment rainfall-runoff datasets, and ten inputs are selected for the Myponga WDS chlorine dataset using the PMI algorithm. The salinity data are approximately normally or uniformly distributed, with very low skewness and near zero or negative kurtosis. For the Myponga chlorine data, apart from the chlorine data at the Myponga water treatment plant (WTP), which are slightly negatively skewed and peaked, the distributions of the majority of the inputs and the output are close to a uniform or normal distribution. In contrast, the Neckar catchment flow data are highly skewed and peaked. In addition, the variability of the flow data is also very high, with the standard deviation of the data being higher than the mean value.

Table 1 Statistics of selected inputs and outputs for the three case studies investigated

Datasets and variables		Lags	Mean	Std.	Skew.	Kurt.
Murray River Salinity (Daily data)	Input Variables (2)					
	Mannum salinity (EC units)	t	558	204	0.59	0.02
	Waikerie salinity (EC units)	t	524	177	0.45	-0.03
	Output variable					
	Murray Bridge salinity (EC units)	t+14	602	229	0.48	-0.23
Myponga Chlorine (Hourly data)	Input Variables (10)					
	Myponga WTP chlorine (mg/L)	t	3.17	0.62	-1.56	7.27
	Myponga tank chlorine (mg/L)	t, t-17	2.01	0.26	-0.43	-0.04
	Cactus Canyon Temp. (°C)	t-13	16.78	3.52	0.18	-0.99
	Aldinga chlorine (mg/L)	t, t-1, t-3, t-24, t-27, t-47	0.75	0.33	0.49	0.14
	Output variable					
	Aldinga chlorine (mg/L)	t+24	0.75	0.34	0.47	0.08
Neckar Runoff (Daily data)	Input Variables (2)					
	Flow (m ³ /s)	t, t-2	5.42	6.94	5.42	49.01
	Output variable					
	Flow (m ³ /s)	t+1	5.42	6.94	5.42	49.00

Validation results and discussion

The mean (μ) and standard deviation (σ) of the performance measures of the trained models obtained using the validation datasets are summarized in Table 2. For the salinity dataset, which has the lowest skewness and kurtosis, the systematic method generates the lowest error measures and highest r^2 for the validation set compared to DUPLEX and SBSS-N (Table 2); while DUPLEX generates the highest RMSE and SBSS-N performed the worst in terms of r^2 . For the chlorine dataset, which is slightly skewed, the systematic method

also performs best with low validation errors, high correlation and low variance. While DUPLEX and SBSS-N perform similarly in terms of validation errors and correlation, DUPLEX does not generate any variation. In addition, for both salinity and chlorine dataset, the variance generated using the systematic method is much lower than that generated using SBSS-N. For the rainfall-runoff dataset, the systematic method performs worst with high validation errors and the lowest r^2 value. In contrast, SBSS-N leads to significantly improved validation results for this dataset compared to the systematic method and DUPLEX. However, the standard deviations of the error measures for the models using SBSS-N are much higher than those found for models that use the systematic method.

A comparison of the guidelines for selecting a data splitting method proposed by May et al. (2010) and the results obtained for the real-world datasets used in this study is illustrated in Figure 1. As can be seen, the same relationship between different data splitting methods and data with different statistical properties obtained by May et al. (2010) generally holds for the real-world datasets. In addition, it has been found that both the skewness and kurtosis of a dataset have a significant impact on the performance of the SBSS-N method; while, the dimensionality of a dataset is more important for the systematic method.

Table 2 Validated model performance for the three case studies and data splitting methods considered

Methods		Murray River salinity			Myponga chlorine			Neckar catchment		
		RMSE	MAE	r^2	RMSE	MAE	r^2	RMSE	MAE	r^2
Systematic	μ	41.27	28.64	0.968	0.071	0.033	0.957	6.53	2.08	0.283
	σ	1.07	0.71	0.002	0.008	0.005	0.009	0.17	0.10	0.038
DUPLEX	μ	59.23	38.69	0.933	0.081	0.041	0.943	6.53	2.69	0.393
	σ	NA*								
SBSS-N	μ	51.54	36.10	0.906	0.072	0.040	0.943	2.54	2.31	0.734
	σ	8.98	5.73	0.028	0.017	0.005	0.026	0.88	0.95	0.061

*NA = Not Applicable. DUPLEX only produces one data split for a given dataset.

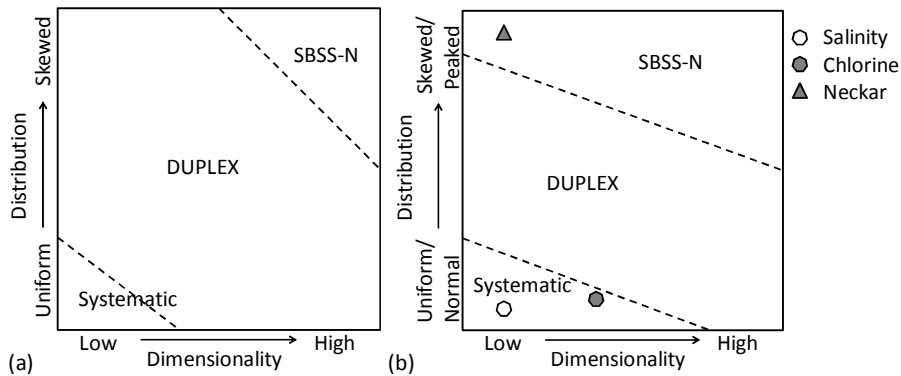


Figure 1 (a) Guidelines for selecting a data splitting method (May et al., 2010); (b) Locations of the three datasets in the dimensionality-distribution space and the corresponding data splitting method that generates the best validation results

In order to investigate the reason why SBSS-N leads to significantly improved validation performance for the highly skewed rainfall-runoff dataset, the statistical properties of the training, test and validation sets from a typical split using SBSS-N for the rainfall-runoff dataset were compared, as shown in Table 3. All of the statistical measures for both the test and validation sets are much lower than for the training set. In particular, the difference in maximum values and the kurtosis suggests that SBSS-N produces a significant degree of bias by allocating the extreme observations into the training set. Consequently, although model performance is good in regions corresponding to the predominant mean rainfall conditions (low or no rainfall), the performance for extreme conditions is actually not tested or validated and is therefore unknown. The reasons for this may be the sensitivity of the methodology to the clustering of the data onto the SOM, and the generation of strata containing a few, yet widely dispersed data, for which the Neyman allocation rule breaks down due to sample quotas exceeding the available number of points required to sample proportionally into each set.

Table 3 Statistics of the training, test and validation datasets from a typical data split using the SBSS-N method for the Neckar catchment data

Dataset	Min	Max	Mean	Std	Skewness	Kurtosis
Training	0.6	114.6	12.15	11.16	3.42	18.70
Test	0.6	18.8	6.12	3.64	0.59	0.03
Validation	0.4	17.2	3.05	1.86	1.76	5.35

CONCLUSIONS AND RECOMMENDATIONS

In this paper, three real-world water resources datasets with varying statistical properties have been used to validate guidelines for selecting a data splitting method for developing ANN models derived by May et al. (2010) for synthetic datasets. It was found that the relationship between different data splitting methods and the statistical properties of the data, which was determined by May et al. (2010) for synthetic datasets, also generally holds true for real-world datasets. However, further examination of the results suggests that some methods, such as SBSS-N may potentially under-represent sparse data corresponding to extreme cases in test and validating datasets, leading to an over-optimistic assessment of model performance.

ACKNOWLEDGEMENT

The authors would like to thank Water Quality Research Australia (WQRA) for its financial support for this study.

REFERECSNES

- Bárdossy, A. and Singh, S. K. "Robust estimation of hydrological model parameters." *Hydrology and Earth System Sciences*, Vol. 12, No. 6, (2008), pp 1273-1283.
- Bowden, G. J., Maier, H. R., and Dandy, G. C. "Optimal division of data for neural network models in water resources applications." *Water Resour. Res.*, Vol. 38, No. 2, (2002), pp 1010.
- Bowden, G. J., Nixon, J. B., Dandy, G. C., Maier, H. R., and Holmes, M. "Forecasting chlorine residuals in a water distribution system using a general regression neural network." *Mathematical and Computer Modelling*, Vol. 44, No. 5-6, (2006), pp 469-484.
- Kennard, R. W. and Stone, L. A. "Computer Aided Design of Experiments." *Technometrics*, Vol. 11, No. 1, (1969), pp 137-148.
- Maier, H. R., Jain, A., Dandy, G. C., and Sudheer, K. P. "Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions." *Environmental Modelling & Software*, Vol. 25, No. 8, (2010), pp 891-909.
- May, R. J., Dandy, G. C., Maier, H. R., and Nixon, J. B. "Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems." *Environmental Modelling & Software*, Vol. 23, No. 10-11, (2008a), pp 1289-1299.
- May, R. J., Maier, H. R., Dandy, G. C., and Fernando, T. "Non-linear variable selection for artificial neural networks using partial mutual information." *Environmental Modelling & Software*, Vol. 23, No. 10-11, (2008b), pp 1312-1326.
- May, R. J., Maier, H. R., and Dandy, G. C. "Data splitting for artificial neural networks using SOM-based stratified sampling." *Neural Networks*, Vol. 23, No. 2, (2010), pp 283-294.
- Sharma, A. "Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 -- A strategy for system predictor identification." *Journal of Hydrology*, Vol. 239, No. 1-4, (2000), pp 232-239.
- Snee, R. D. "Validation of regression models: Methods and examples." *Technometrics*, Vol. 19, No. 4, (1977), pp 415-428.
- Specht, D. F. "A general regression neural network." *Neural Networks, IEEE Transactions on*, Vol. 2, No. 6, (1991), pp 568-576.
- Vesanto, J. "SOM-based data visualization methods." *Intelligent Data Analysis*, Vol. 3, No. 2, (1999), pp 111-126.