International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

A method for comparing data splitting approaches for developing hydrological ANN models

Wenyan Wu [a], Robert May [a,b], Graeme C. Dandy [a] and Holger R. Maier [a]
a School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, 5005, Australia. wenyan.wu@adelaide.edu.au, graeme.dandy@adelaide.edu.au, holger.maier@civeng.adelaide.edu.au
b Veolia Water Asia-Pacific, Technical Department, Network Management Team, Shanghai, 200041, China. robert.may@veoliawater.cn

Abstract: Data splitting is an important step in the artificial neural network (ANN) development process whereby data are divided into training, test and validation subsets to ensure good generalization ability of the model. Considering that only one split of data is typically used when developing ANN models, data splitting has a significant impact on the performance of the final model by potentially introducing bias and variance into the model development process. Therefore, it is important to find a robust data splitting method which results in an ANN model that represents the underlying data generation process of a given dataset. In practice, ANN models developed using different data splitting methods are often assessed based on validation results. In previous research, however, it has been found that validation results alone are not adequate for assessing the performance of ANN models. Data splitting methods have the potential to bias the validation results by allocating extreme observations into the training set and therefore, the test and validation sets contain fewer patterns compared to the training set. Consequently, the generalization ability of the model may be compromised and the trained model cannot be adequately validated. This paper introduces a method to compare different data splitting methods for developing ANN models fairly. The methodology is applied to compare a number of well-known data splitting techniques in the context of some hydrological ANN modeling problems.

***Keywords**:* Artificial neural networks, data splitting.

1 INTRODUCTION

Artificial neural networks (ANNs) have become a popular approach for environmental modeling in the last two decades. An important step in the ANN model development process is data splitting, which divides available data into training, test and validation datasets (Maier et al., 2010). The training set is used to optimize model parameters (train the model); the test set is used for cross-validation during training to avoid over-fitting; and the validation set is used to assess the performance of the trained model. Thus, the generalization ability of the trained model is tested in a rigorous fashion (May et al., 2010). Generally, data are divided using a variety of methods, including ad-hoc methods, random methods, stratified methods and optimization based methods, etc. (Maier et al., 2010; May et al., 2010). Often, a different split and hence different performance for training, testing and validation are obtained when different data splitting methods are used. This is not only because of the random nature of many data splitting methods (e.g.

simple random sampling), but also due to the variability that is generally exhibited by environmental data.

The variability of ANN performance due to data variability was explored by LeBaron and Weigend (1998). In their study, the authors randomly generated 2523 bootstrap instances of training, test and validation datasets for a given dataset and trained an ANN model for each instance. A histogram of the model performance distribution was used to examine the variability that data splitting can bring into the ANN model development process. The authors found that the variability in model performance due to data splitting is greater than the variability due to model structure. Consequently, if the performance distribution generated using the random instances of data splits has a high average error and a high standard deviation, there is little that can be done in terms of adjusting model structure in order to significantly improve predictive performance.

The potential variability in ANN model performance due to the way the data are split into their respective subsets makes it difficult to compare the performance of models developed using different data splitting methods, as a single split generated using a particular method is generally used for developing ANN models (e.g. Bowden et al., 2002). In addition, this high degree of variability makes it difficult to assess the impact of other aspects of the ANN model development process on model performance, such as the choice of model inputs and the selection of model architecture or calibration method. Consequently, there is a need to develop an approach that enables the impact of different data splitting methods on model performance to be compared in a rigorous and unbiased fashion. Such an approach is introduced and tested in this paper.

The remainder of this paper is organized as follows. The proposed approach for comparing different data splitting methods is introduced in the next section, followed by the application of the approach to two rainfall-runoff case studies, as part of which the performance of three different data splitting methods is compared. The results and discussion of the results are presented in the subsequent section, followed by conclusions.

2 PROPOSED METHOD FOR COMPARING DATA SPLITTING APPROACHES

The proposed method for comparing data splitting approaches consists of two main steps. In the first step, an estimate is made of the *expected* model validation performance, given the variation in the available data. This can be done by partitioning the available data *H* times using simple random (SR) sampling from *H* different starting points (i.e. random seeds) and developing an ANN model for each partition of data, resulting in a distribution of validation performance values. The mean of this distribution provides an indication of *expected* model performance, given the available data, provided *H* is sufficiently large. Consequently, this provides an unbiased benchmark against which the performance of other data splitting methods can be compared.

As part of the second step, the validation performance of ANN models developed using different data splitting methods is compared with the *expected* value of model validation performance obtained above. As many data splitting methods are not deterministic and are likely to result in different data splits each time they are implemented, the mean value of model validation performance also has to be calculated for each of the data splitting methods considered. This can be done by implementing each data splitting method *H* times and developing an ANN model for each partition of data, resulting in a distribution of validation performance values, from which the mean value can be calculated.

This mean value can then be compared with the *expected* value of model validation performance obtained previously. If these two values are identical, the data splits obtained using the data splitting method under consideration result in models that provide a true indication of the predictive capability of the model over the full range of the data available for model development (Figure 1). If the values are different, then the data splitting method under consideration results in biased model performance. If the mean validation performance value obtained using a particular data splitting method is better than the *expected* value, the predictions obtained using the resulting model are likely to be *optimistic*, where the predictive performance of the model over the full range of data is overestimated (Figure 1). In contrast, if the mean validation performance value obtained using a particular data splitting method is worse than the *expected* value, the predictions obtained using the resulting model are likely to be *pessimistic*, where the predictive performance of the model over the full range of data is underestimated (Figure 1). An over-pessimistic model cannot produce forecasts with a desired level of accuracy; whereas, an over-optimistic model often results from under-represented sparse data corresponding to extreme cases in test and validation datasets, which does not guarantee the underlying data generating process is fully represented or tested (LeBaron and Weigend, 1998; Wu et al., 2012).

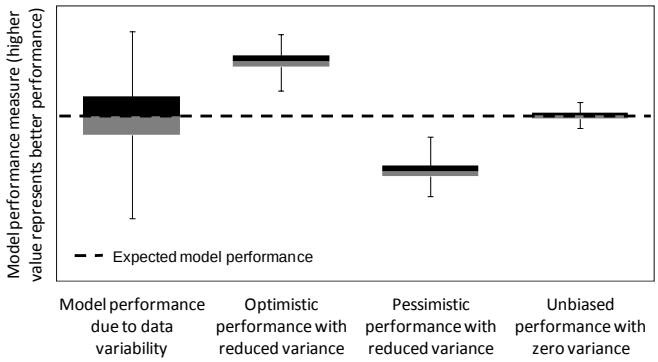

Figure 1 Illustration of the performance distribution of a dataset due to data variability generated using SR method and the optimistic, pessimistic and unbiased model performance with reduced variance due to use of other data splitting methods

In practice, an ANN model is generally developed using a single split of data generated using a particular method. Consequently, the variance of the distribution of model validation performance for different data splitting methods is also extremely important. Although the SR data splitting approach, when repeated sufficient times, leads to an unbiased estimate of model performance, the high variance of the method makes it unsuitable for practical purposes (i.e. when only one data split is used). As the way data are sampled for the various subsets is more structured in other data splitting methods, one would expect the resulting variance to be reduced. However, some methods can still result in significant variation in validation performance (see May et al., 2010), which may lead to validation performance of the actual model developed being far from the mean performance of the data splitting method. In addition, biases (either optimistic or pessimistic) are introduced when these methods are used. Therefore, an ideal data splitting method is one that leads to an unbiased model (i.e. neither optimistic nor pessimistic) and exhibits no or very low variance due to the split of data (the fourth plot at the right end in Figure 1).

3 CASE STUDIES

In this section, the proposed approach for comparing different data splitting methods introduced in the previous section is applied to two real-world hydrological datasets with varying statistical properties. Three representative data splitting

methods used in previous ANN applications, including the systematic method, the self-organizing map (SOM) based stratified sampling (SBSS) method with Neyman sample allocation rule (SBSS-N) and the DUPLEX method are used for this purpose.

3.1 Datasets

3.1.1 Kentucky River Catchment (USA) Rainfall-runoff Data

This dataset is from the Kentucky River catchment for the period of 1960 to 1972. The dataset includes 4,749 daily observations of effective rainfall and runoff. The dataset was used by Jain and Srinivasulu (2006) to train an ANN model to forecast runoff one day in advance. In this paper, up to 10 lags for each of the two variables are used, resulting in a total of 20 potential inputs.

3.1.2 Upper Neckar Catchment (Germany) Rainfall-runoff Data

This dataset 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 effective rainfall and runoff values. Up to 10 lags of each variable are used, resulting in a total of 20 potential inputs.

3.2 Data Splitting Techniques

3.2.1 Systematic Data Splitting Method

The systematic data splitting method (Baxter et al., 2000) is a semi-deterministic method, in which every k^{th} sample from a random starting point is selected to form the training, test and validation datasets. In implementing systematic sampling in this study, the data are first ordered in increasing values 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.

3.2.2 SBSS-N Data Splitting Method

The SBSS approach is a two-step data splitting method. In the first step, multivariate stratified random sampling is performed to partition the data into *M* strata, where clustering is performed using a self-organizing map (SOM). In the second step, uniform random intra-cluster sampling is applied to generate the data split. In this study, the Neyman allocation rule for determining the number of training and test points drawn from each stratum is used. The number of samples to be taken from stratum *m* based on the Neyman allocation rule is expressed as:

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

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 multivariate standard deviation of stratum i. Based on this rule, samples are taken from each stratum based on the global proportions, but with increased sampling for wider clusters. The aim of this rule is to add more data into training and test sets where data are increased in variability, and less where data are less variable or abundant. SBSS-based data splitting methods were implemented by Bowden (2002) and Kingston (2006). In this study, the SOM algorithm was implemented following the methodology defined in May et al. (2010).

3.2.3 DUPLEX Data Splitting Method

The DUPLEX data splitting method was developed by Snee (1977) based on one of the earliest data splitting algorithms called CADEX or Kennard-Stone sampling

(Kennard and Stone, 1969). DUPLEX draws samples based on Euclidean distances. When applying DUPLEX, the two points which are farthest apart are assigned to the first dataset. The next pair of points that are farthest apart in the remaining list are assigned to the second dataset. This process is repeated until both datasets are filled (Snee, 1977). The original DUPLEX algorithm was used to divide data into two sets. May et al. (2010) modified the original DUPLEX algorithm to generate three datasets based on the proportion specified by the user. Thus, DUPLEX can be used to generate the training, test and validation datasets for ANN model development. DUPLEX has been found to generate representative subsets of data by Despagne and Massart (1998).

3.3 ANN Model Development

In this study, the partial mutual information (PMI) based non-linear variable selection algorithm (Sharma, 2000) combined with the Akaike Information Criterion (AIC) for stopping (May et al., 2008) are used to select appropriate inputs for each dataset. After applying the PMI algorithm, two inputs are selected for both datasets. The statistics of the selected inputs and the outputs of the two datasets are summarized in Table 1. As can be seen, the variability of both datasets is very high, with the standard deviation of the data being higher than the mean value. In addition, the skewness and peakedness of the Neckar dataset are much higher than those of the Kentucky dataset.

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

Datasets and variables		Lags	Mean	S.D.	Skew	Kurt.
Kentucky (Daily)	Input Variables (2)					
	Flow (ft^3/s)	t,t-1	5174	8436	3.72	18.99
	Output variable					
	Flow (ft^3/s)	t+1	5174	8436	3.72	18.99
Neckar (Daily)	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

Sixty percent of each dataset is used for training, 20% for test and 20% for validation. This is achieved using SR sampling and the three data splitting approaches presented in the previous section. In this paper, the general regression neural network (GRNN) is used. Compared to multilayer perceptrons (MLPs), which have been used more commonly in ANN applications in hydrological modeling (Maier et al., 2010), the architecture of GRNNs is fixed and there is only one parameter (the bandwidth) that needs to be optimized. Therefore, a GRNN model is much faster to develop (May et al., 2008), which suits the purposes of this study. In addition, it assists with isolating the effects of the different data splitting methods on model performance. All of the GRNN models developed in this study are trained using Brent's method. For model validation, the root mean square error (RMSE) and the square of Pearson R (R^2) are used in order to assess model performance and compare the models developed using different data splitting methods.

3.4 Comparison of data splitting methods

In this study, 100 bootstrap instances (e.g., H=100) of training, test and validation samples are generated using the SR, systematic and SBSS-N methods and an ANN model is developed for each instance to develop the expected performance distribution of the dataset and the performance distribution of the systematic and SBSS-N methods. As the performance of DUPLEX is deterministic (i.e. there is no random component in the method), no bootstrapping is required for this method.

Box-and-whisker diagrams are used to compare the validation performance distribution generated using different methods, as they do not make any assumption about the underlying statistical distributions of the performance distributions obtained using different data splitting methods.

4 RESULTS AND DISCUSSION

The average (μ) and standard deviation (σ) of the performance measures of the trained models obtained using the validation datasets, as well as the bias of the three data splitting methods, are summarized in Table 2. The box-and-whisker diagram of the model validation performance distributions generated using the repeated SR sampling approach and the three other data splitting methods considered for both datasets are presented in Figure 2.

Table 2 Comparison of validation results obtained using different data splitting methods for the case studies considered

Method		Kentucky		Neckar	
		R^2	RMSE	R^2	RMSE
Repeated SR (100)	μ	0.826	3516	0.437	5.20
	σ	0.070	795	0.090	0.91
Systematic (100)	μ	0.851	3249	0.283	6.53
	σ	0.005	50	0.038	0.17
	Bias	3.09% (O)*	7.59% (O)	35.24% (P)	25.57% (P)
SBSS-N (100)	μ	0.871	1303	0.734	2.54
	σ	0.043	230	0.061	0.88
	Bias	5.51% (O)	62.94% (O)	67.96% (O)	51.16% (O)
DUPLEX	μ	0.849	3889	0.393	6.53
	σ	0	0	0	0
	Bias	2.84% (O)	10.61% (P)	10.07% (P)	25.57% (P)

* O = optimistic; P = pessimistic

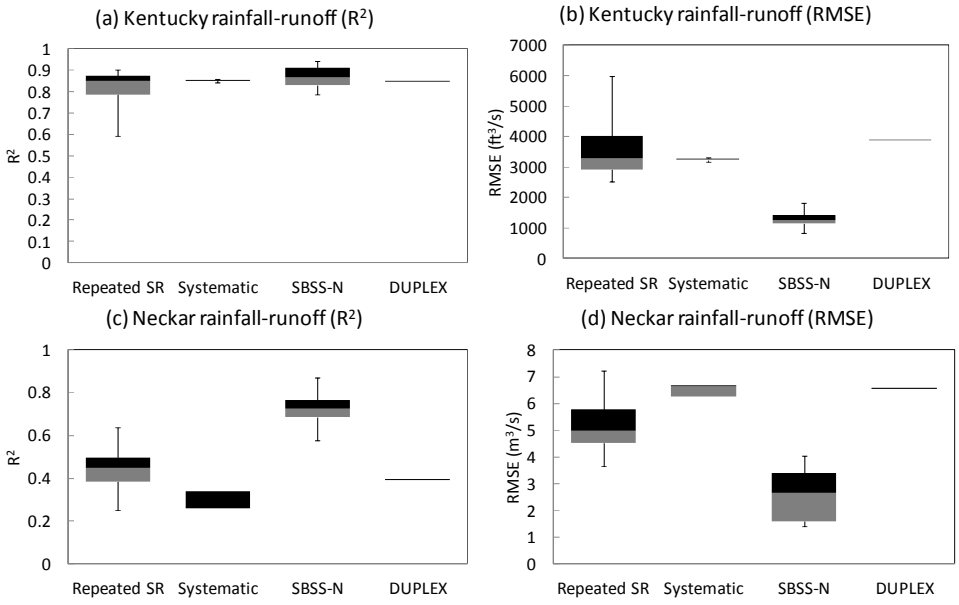

Figure 2 The *expected* model performance distribution and the performance distributions obtained using the three data splitting approaches

As can be seen in Figure 2, for all of the cases considered (e.g. combination of different datasets with different model performance measures), use of the three data splitting methods investigated can significantly reduce the variability of model performance compared with that estimated using the SR sampling approach. This is due to the more structured ways in which data are divided when these data

splitting methods are applied compared with complete random trials, as mentioned in the Section 2. However, as a tradeoff for the reduced variance, bias is introduced due to the use of these data splitting methods. The magnitude and direction of these biases depend on the data splitting method and the statistical properties of the model development data. It is obvious from Figure 2 that the biases generated by data splitting are much lower for the Kentucky dataset compared to those for the Neckar dataset. This may be due to the higher skewness and kurtosis of the Neckar dataset, as shown in Table 1. Table 2 also shows that the biases generated using the systematic and DUPLEX methods are generally much lower than those generated using the SBSS-N method. For example, the biases generated using the systematic and DUPLEX methods for the RMSE measure of the Kentucky dataset are around or under 10%. In contrast, the bias generated using SBSS-N in the same case is above 60%. The reasons for this may be the sensitivity of the SBSS-N method 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. The optimistic nature of the SBSS-N method for highly skewed datasets was also observed by Wu et al. (2012).

The results presented in Table 2 and Figure 2 also show that the systematic method can be either optimistic (e.g. for the Kentucky dataset) or pessimistic (e.g. for the Neckar dataset) and SBSS-N is extremely optimistic for both case studies. In contrast, the results for Duplex are slightly pessimistic for both case studies, which is in agreement with the observation by Snee (1977). However, compared to the other methods, DUPLEX performs more consistently across all four cases (e.g. combination of case study and performance criteria) in terms of the magnitude of the bias generated.

Another significant advantage of DUPLEX is that it only produces one split for a given dataset and therefore, it does not generate any variance. The variance generated using the systematic method is relatively low compared to that generated using the SBSS–N method. This is most obvious for the Kentucky dataset, where both the inter-quartile range and the data range of the RMSE and R^2 obtained using the systematic method are very small (Figures 2(a) and 2(b)). The relatively higher variability of the systematic method for the Neckar dataset is probably due to the high skewness and peakedness of the dataset, which has been reported by May et al. (2010) and Wu et al. (2012). In contrast, the validation performance obtained using SBSS-N is highly variable for both datasets compared with that obtained using DUPLEX and the systematic method.

It should be noted that the overall performance of the ANN models for the two case studies varies significantly. In general, the performance of the models developed for the Kentucky dataset is much better than that of the models developed for the Neckar dataset, which is represented by the generally higher R^2 shown in Figure 2(a) compared to those in Figure 2(c). This indicates that some important information representing the input-output relationship might not be included in the latter dataset. However, this is unlikely to have a significant impact on the comparison of different data splitting methods, which is the primary focus of this study.

5 CONCLUSIONS

Due to the high variability often exhibited by environmental data, the validation performance of ANN models developed using these data can be highly variable, depending on which data are put into the training, test and validation subsets. This makes it difficult to provide an objective assessment of the performance of models developed using different data splitting methods. By using the approach suggested in this paper, the performance of a data splitting method for a given dataset can be

obtained in terms of bias and variance. The results from the hydrological case studies indicate that different data splitting methods exhibit different levels of bias and variance compared with the expected model performance of a given dataset. This is problematic for developing ANN models, as commonly only a single split of data is generated using a particular data splitting method. However, the DUPLEX method has been found to overcome this problem to a large extent. As DUPLEX is a deterministic method, it generates zero variance. It also results in relatively low bias compared to other data splitting methods, although it leads to performance estimates that are slightly pessimistic.

ACKNOWLEDGEMENT

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

REFERENCES

B´ardossy, A. and Singh, S. K. "Robust estimation of hydrological model parameters." *Hydrology and Earth System Sciences*, 12(6), 1273-1283. 2008.

Baxter, C. W., Stanley, S. J., Zhang, Q., and Smith, D. W. "Developing artificial neural network process models: A guide for drinking water utilities." 6th environmental engineering society specialty conference of the CSCE, 376-383. 2000.

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.*, 38(2), 1010. 2002.

Despagne, F. and Luc Massart, D. "Neural networks in multivariate calibration." *Analyst*, 123(11). 1998.

Jain, A. and Srinivasulu, S. "Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques." *Journal of Hydrology*, 317(3-4), 291-306. 2006.

Kennard, R. W. and Stone, L. A. "Computer Aided Design of Experiments." *Technometrics*, 11(1), 137-148. 1969.

Kingston, G. B. "Bayesian artificial neural networks in water resources engineering," The University of Adelaide, Adelaide, Australia. 2006.

LeBaron, B. and Weigend, A. S. "A bootstrap evaluation of the effect of data splitting on financial time series." *Neural Networks, IEEE Transactions on*, 9(1), 213-220. 1998.

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*, 25(8), 891-909. 2010.

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*, 23(10-11), 1289-1299. 2008.

May, R. J., Maier, H. R., and Dandy, G. C. "Data splitting for artificial neural networks using SOM-based stratified sampling." *Neural Networks*, 23(2), 283-294. 2010.

Sharma, A. "Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 -- A strategy for system predictor identification." *Journal of Hydrology*, 239(1-4), 232-239. 2000.

Snee, R. D. "Validation of regression models: Methods and examples." *Technometrics*, 19(4), 415-428. 1977.

Wu, W., Maier, H. R., Dandy, G. C., and May, R. "Exploring the impact of data splitting methods on artificial neural network models." *the 10th International Conference on Hydroinformatics*, Hamburg, Germany. 2012.