CHAPTER 8

INFILLING STREAMFLOWS USING BACKPROPAGATION TECHNIQUES

8.1 INTRODUCTION

In this chapter, feedforward backpropagation (BP) artificial neural network techniques are used for streamflow data infilling. The standard BP technique with a sigmoid function (Freeman and Skapura, 1991) is used and the BP with an approximation to the sigmoid function by pseudo Mac Laurin power series order 1 and order 2 derivatives is also used. Empirical comparisons of the predictive accuracy, in terms of the directional informational transfer index (DIT) are then made. A preliminary case study in South Africa (i.e. using the Diepkloof control gauge-D1H001 on the Wonderboomspruit River and Molteno target gauge-D1H004 on the Stormbergspruit River in the River summer rainfall catchment) was then done. It should be noticed that in a paper submitted to the Water S.A. Journal and found suitable for publication, the accuracy of estimated values was investigated in terms of the root mean squared errors of predictions (Ilunga and Stephenson, 2005). Two seasons of a 6-month period each were assumed (wet-October to March and dry-April to September). Recall that Pegram (1997) found that the months of October and September could fall into early summer (e.g. wet) and dry seasons respectively. The means of seasonal values were considered as data regime and the standard BP and its variants, viz. pseudo Mac Laurin order 1 BP (McL1BP) and pseudo Mac Laurin order 2 BP (McL2BP) were applied to that data regime.

8.2 STANDARD BP WITH SIGMOID FUNCTION APPROXIMATED BY PSEUDO-MAC LAURIN POWER SERIES

Recall that the activation function most commonly used is a sigmoid, non-linear continuous function between 0 and 1, as explained in the literature review, refer to equation 2.62, Chapter 2.

$$f(x) = \frac{1}{1+e^{-x}}$$

The first derivative of this function, which is encompassed in the error term used in the update equations 2.71 and 2.72, is given by

$$f'(x) = f(x)(1-f(x))$$

Criticisms were formulated against the standard BP (which is a gradient descent method) for not guaranteeing necessarily convergence to an optimal solution (Argawal and Singh, 2001). Thus several variants of the BP such as Newton's method, Adaptive stepsize, etc were proposed. Despite these criticisms, it appears in practice that the BP leads to solutions in almost everywhere and standard multi-layer, feedforward networks are capable of approximating any measurable function to any desired degree of accuracy; as repeated by Minns and Hall (1996).

In this section, the BP is performed by approximating the sigmoid function by "pseudo" Mac Laurin power series order 1 and 2 derivatives, as shown so far in section 3.3.6.1.7 of Chapter 3. The Mac Laurin power series order 1 and order 2 derivatives approximate the sigmoid function by (see equations 3.10 and 3.12, Chapter 3)

$$f'(x) \approx \frac{1}{(2-x)^2} \approx (f(x))^2$$

$$f'(x) \approx \frac{1-x}{(2-x+\frac{x^2}{2})^2} \approx (1-x)(f(x))^2$$

These equations could be used in the error terms for weights update equations for the BP technique. Hence, the resultant BP techniques from this approximation are named pseudo Mac Laurin power series order 1 (McL1BP) and pseudo Mac Laurin power series order 2 (McL2BP) respectively.

8.3 RESULTS AND DISCUSSION

Referring to Table 8.1, DH1004 was considered as target gauge and gauge DH1001 as the control gauge. This was concluded from entropy calculations, i.e. the DIT value for the station pair D1H001-D1H004 was higher than the one for the station pair D1H004-D1H001. However, with a threshold of 30 % for DIT, both stations would have been considered as being capable of inferring information mutually.

The selected streamflow data set was complete and thus exhibited no gaps. However, for testing of the different infilling techniques, some consecutive gaps (e.g. 6.7 %, 13.3 %, 20 %, 30% of missing data, and arbitrarily starting at 1934) were created on the target streamflow gauge data series, i.e. D1H004. The ANNs were trained in a sequential mode on the concurrent parts of observed data and the weights obtained were then used to estimate the missing values. A single input-output ANN with 3 nodes in the hidden layer was used and the bias term to the input was assumed to be zero as its use is optional (Freeman and Skapura, 1991). The learning rate was set to 0.35 throughout for acceptable results compared to other values. Input and output values were scaled to fall within the range between 0.1 and 0.9 as mentioned earlier.

Table 8.2 summarizes the results of performance for the three techniques, i.e. the standard BP, McL1BP and McL2BP techniques.

Table 8.1 DIT of seasonal mean flows for station-pairs.

	D1H001	D1H004

D1H001	1	0.8022
D1H004	0.3614	1

Table 8.2 Performance evaluation of standard BP, McL1BP and McL2BP

Algorithm	DIT(-)			
	6.7 %	13.3 %	20 %	30 %
Standard BP	0.838	0.6478	0.4422	0.3366
McL1BP	0.817	0.6408	0.4535	0.3319
McL2BP	0.778	0.6438	0.4385	0.3343

From Table 8.2, it follows that, generally, the DIT increases with increases in the proportion of missing values (gap size) for all three techniques. Thus, the accuracy decreases as the gap size increases. Generally, the standard BP performs just slightly better than the McL1BP and McL2BP techniques for this specific data set. This could be due to the fact that the error terms in the update equations (2.71) and (2.72), which encompass a derivative part, are slightly bigger for McL1BP and McL2BP techniques than for standard BP. However, the Mac Laurin approximation did not show any substantial negative impact on the accuracy of the estimated missing values.

Figures 8.1 (a-c) show DIT (thus accuracy) versus the gap size (% of missing values) at gauge D1H004 for the standard BP, McL1BP and McL2BP respectively. From these figures, it is seen that, for all algorithms, the bigger the gap size, the bigger the DIT, thus the accuracy becomes increasingly less. However, it is observed from these figures that an exponential function can strongly describe relationship between the gap size and DIT.

The coefficients of determination (which are very close) were found to be 0.989, 0.990, and 0.977 for standard BP, McL1BP and McL2BP respectively (refer to R-square values on figures 8.2 (d-f)). This correlates with the observation that the differences in estimated values were small for the respective techniques at different gap sizes (0-30%).
It was noticed that, increasing the number of data points between 6.7 % and 30 % (e.g. up to seven values of gap size at the target gauge D1H004) did not sensitively affect the

above-mentioned relationship (between the gap size and DIT). It was also noticed that an earlier start (e.g. at 1928) or later start (e.g. 1938) for the gaps created on the records of the subject station did not have any substantial impact on the accuracy of the estimated values.

From the results obtained here, it can be said that all three the standard BP and McL1BP and McL2BP algorithms are acceptable to fill in the missing values for gauge D1H004. This can be done within the range between 0 and 20% without any significant violation of either the accuracy of estimated values or the statistical properties such as the mean and variance of the incomplete and infilled series.

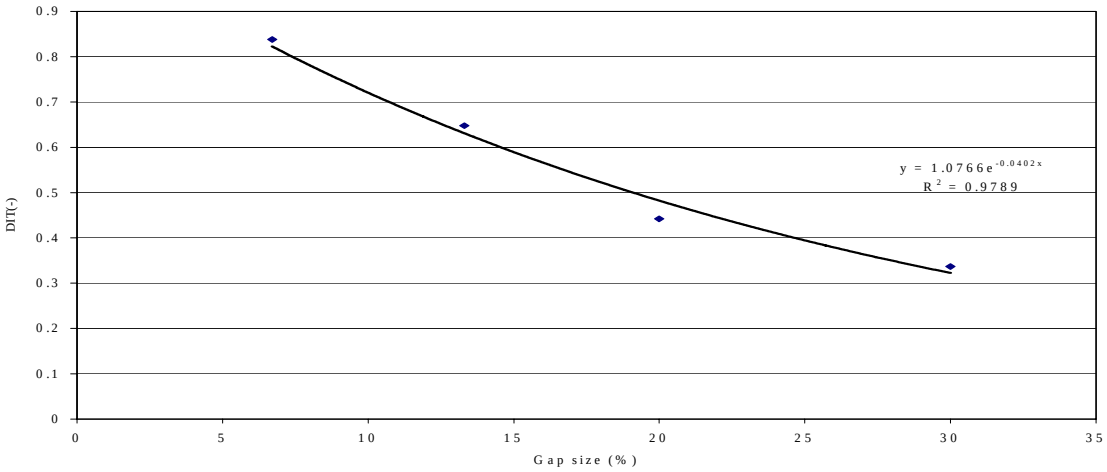

Figure 8.1a DIT versus gap size for seasonal mean flows at D1H004 (base gauge D1H001): Standard BP

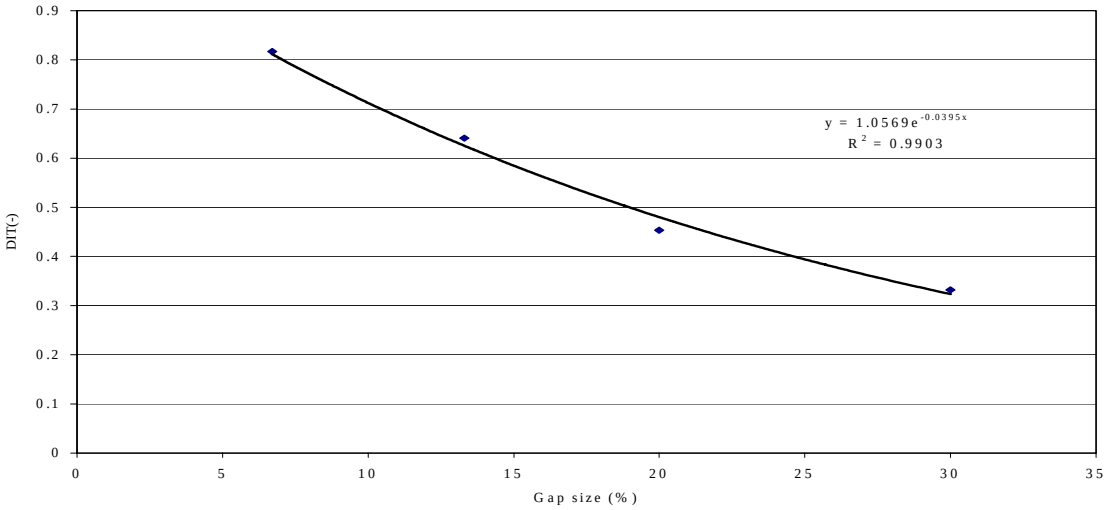

Figure 8.1b DIT versus gap size for seasonal mean flows at D1H004 (base gauge D1H001): McL1BP

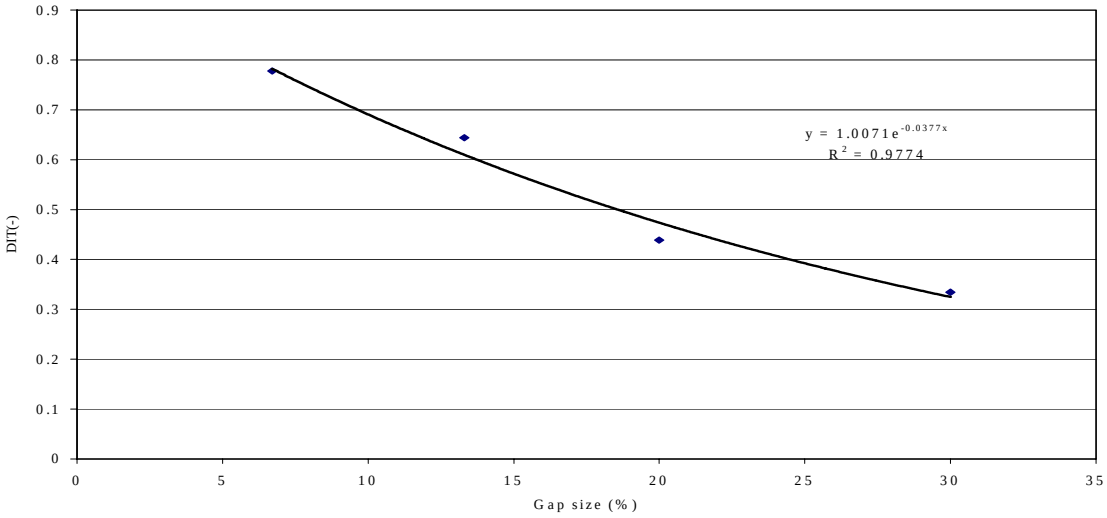

Figure 8.1c DIT versus gap size for seasonal mean flows at D1H004 (base gauge D1H009): McL2BP

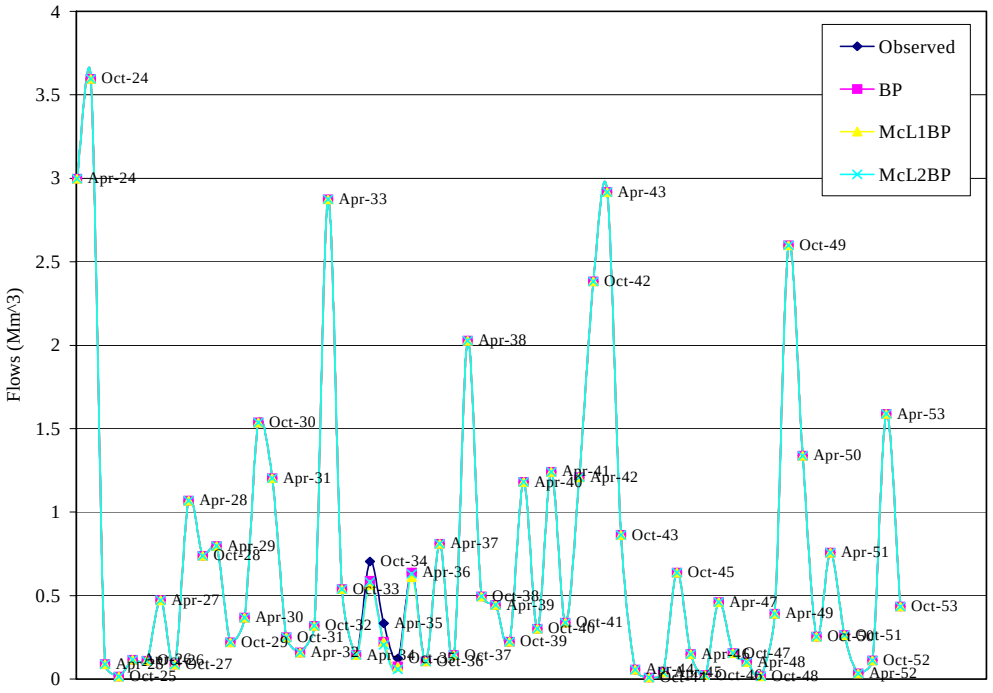

Figure 8.2a Comparison of ANNs in terms of hydrographs at D1H004 (6.7 % missing seasonal mean flows from 1934) using base gauge D1H001

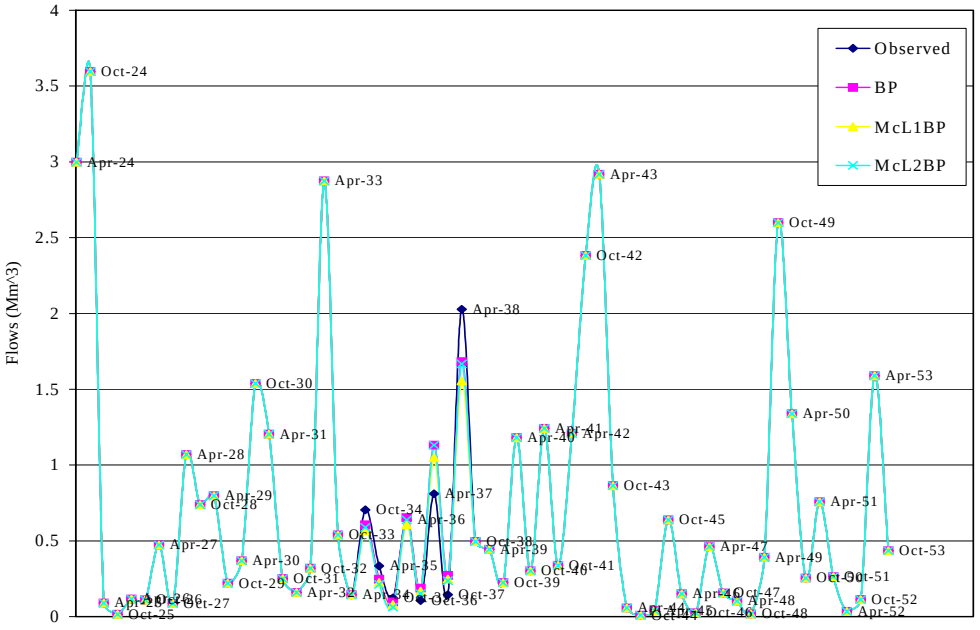

Figure 8.2b Comparison of ANNs in terms hydrographs at D1H004 (13.3 % missing seasonal mean flows from 1934) using base gauge D1H001

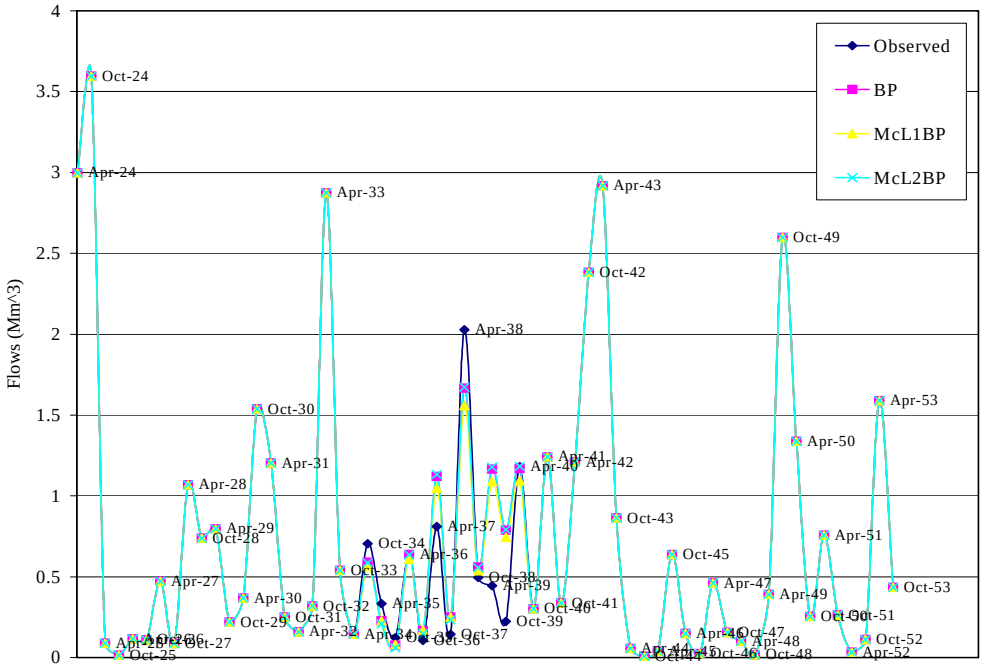

Figure 8.2c Comparison of ANNs in terms hydrographs at D1H004 (20 % missing seasonal mean flows from 1934) using base gauge D1H001

8.4 SUMMARY

Besides the standard BP algorithm, two other techniques, viz. pseudo Mac Laurin (order 1 and order 2 derivatives) BP have been introduced for scaled input and output data in the interval (0.1, 0.9). These preliminary results showed that the pseudo Mac Laurin approximation does not affect substantially the accuracy of the estimated values at gauge D1H004, when compared to the standard BP. Thus, both techniques were acceptable to fill in the missing values. However, it was observed that a decay exponential function could describe a strong relationship between the gap size and the expected DIT for the three algorithms under investigation. Recall that these techniques have been applied to mean values of seasonal streamflow data. Other flow regimes should be also tried (4-month seasons, extremes, etc.). These techniques should be also applied to streamflow series of a winter-rainfall region.