Neural Network - Report

Arkadiusz Brunon Podkowa

February 2022

Contents

List of Figures

List of Tables

1 Introduction

We use six predictors

2 Programming Language Selection

3 Data Preprocessing

3.1 Data Interpolation

Because we deal with time series data, we need to interpolate missing or spurious data and outliers. We have usef linear interpolation. Table 1 shows rows that contain interpolated data. It also points out the reason for each data interpolation.

The following table shows the deleted rows and the reason for deletion.

Row	Erroneous Column	Erroneous Value	New Value	Reason
07/04/1993	Skelton (Mean Daily Flow)	a	number	non-numerical
01/03/1995	East Cowton (Daily Rainfall Total)	#	non-numerical data	
09/02/1996	Skip Bridge (Mean Daily Flow)	a	non-numerical data	

Table 1: Rows with interpolated values.

3.2 Outliers

I have divided outliers analysis into two parts: Mean Daily Flow columns and Rainfall columns.

I do not consider any outliers in Mean Daily Flow columns. The value that is the most distant from the mean lies more 7.2 standard deviations from the mean. However, given similar values in the Winter months (362.3 on 1995-02-02 or 337.2 on 1993-09-15) and the rising severity of the weather conditions, I consider this value to be accurate. Figure 1 shows the data for Mean Daily Flow columns.

I have identified three outliers in Rainfall columns. Figure 1 shows data for Rainfall columns.

With the scale of y-axis shown above, it is easy to detect that 3 values lie considerably far from the rest of the values. The table 2 describes outlier values and the values that they have been replaced with. Again, linear interpolated method has been used.

3.3 Data Splitting

Data has been split into three subsets:

- Training set (60% of overall data),

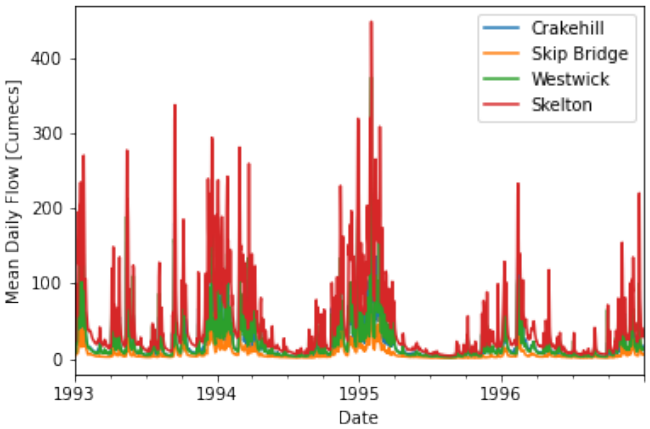

Figure 1: River Flow Data

Row	Erroneous Column	Outlier Value	Interpolated Value
1995-02-11	Arkengarthdale	5000.0	15.6
1995-02-28	East Cowton	9000.0	0.0
1996-01-10	Malham Tarn	80000.0	4.4

Table 2: Rainfall Data Outliers.

- Validation set (20% of overall data), and
- Test set (20% of overall data).

3.4 Data Standardisation

Data has been standardised using the formula below:

$$S_i = 0.8\frac{R_i - Min}{Max - Min} + 0.1$$

where R_i is a raw value and S_i is the standardised value.

Min and Max values are the minimum and maximum values of the training and validation sets combined, respectively.

4 Predictors Selection

I have identified eight predictors for the mean daily flow in Skelton:

- Mean Daily Flow in Skelton on the previous day,

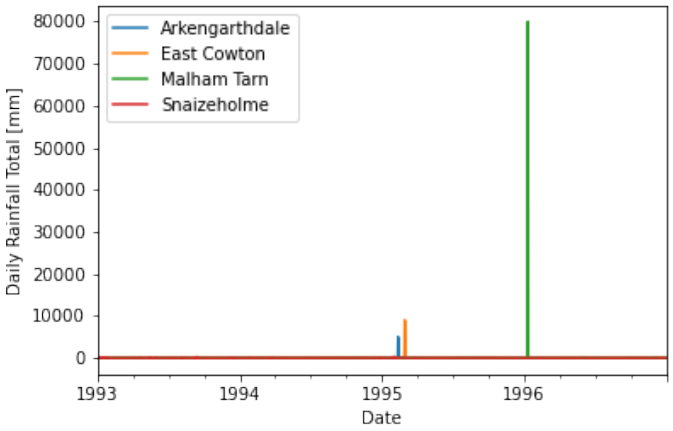

Figure 2: Rainfall Data

- Mean Daily Flow in Crakehill on the previous day,
- Mean Daily Flow in Skip Bridge on the previous day,
- Mean Daily Flow in Westwick on the previous day,
- Daily Rainfall Total in Arkengarthdale on the previous day,
- Daily Rainfall Total in East Cowton on the previous day,
- Daily Rainfall Total in Malham Tarn on the previous day, and
- Daily Rainfall Total in Snaizeholme on the previous day.

Because both mean daily flow and daily rainfall total values cannot be obtained before a particular day ends, the predictor values have to be lagged by at least one day. In essence, we use values predictors' values from the previous day to predict mean daily flow in Skelton for the current day.

I decided to lag predictor columns only by one day, as it gave the stronger correlation between a predictor and the predictand than two-day or three-day lags. Table 3 shows correlation (Pearson correlation coefficient) between a predictor and the predictand for particular lag value (in days).

Figure 3 shows correlation between a particular predictor (x-axis) and the predictor (y-axis).

Predictor	T-1	T-2	T-3
Skelton	**0.889**	0.749	0.663
Crakehill	**0.885**	0.724	0.625
Skip Bridge	**0.884**	0.735	0.643
Westwick	**0.912**	0.733	0.627
Arkengarthdale	**0.507**	0.411	0.312
East Cowton	**0.338**	0.257	0.191
Malham Tarn	**0.495**	0.411	0.333
Snaizeholme	**0.584**	0.487	0.389

Table 3: Rainfall Data Outliers.

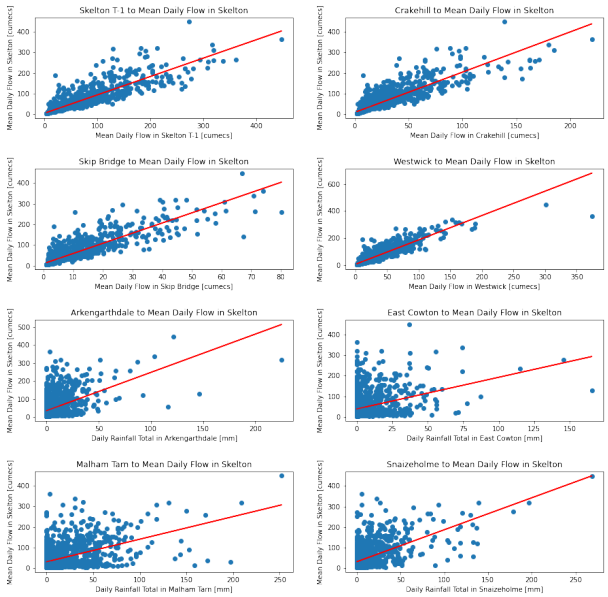

Figure 3: Correlations