# Preprocessing - Cleaning

Rows deleted and justification

* Row 79: -999,0.483,0.979,0.0339,40.88,0.58,42.8,1145,63.141
  + ‘-999’ value in column 1
* Row 114: 87.41,0.499,0.979,0.0837,17.88,a,39.1,891,26.158
  + ‘a’ value in column 6
* Row 181: 53.88,0.461,-999,-999,14.1,0.54,43.8,1316,19.135
  + ‘-999’ value in column 3
  + ‘-999’ value in column 4
  + Note: This row may be used if columns 3 and 4 prove to be inconsequential
* Row 538: 164.46,0.419,0.997,0.0414,36.96,0.59,53.1,,114.109
  + Missing value in column 8
* Row 548: 56.18,0.968,0.983,0.1161,21.65,0.24,30.6,688,-999
  + ‘-999’ value in column 9
* Row 587: 79.86,0.561,0.975,0.0316,21.35,,38.3,949,19.489
  + Missing value in column 7
  + Note: This row may be used if column 7 prove to be inconsequential
* Row 340: 1380.04,0.506,0.976,0.0468,107.51,1.74,38,1108,436.809
  + ‘1.74’ value in column 6 doesn’t fit the range for this column (~0.2 to ~0.8)
  + <http://nrfa.ceh.ac.uk/feh-catchment-descriptors>
    - “PROPWET values range from over 80% in the wettest catchments to less than 20% in the driest parts of the country.”
  + Not yet deleted programmatically

Total rows deleted: 6

Total rows to be deleted: 7

# In Excel

1. Create Correlation Matrix of each column against the flood index.
   1. Sort by correlation, note columns that have low correlation (<0.2)
2. Remove columns that have a low correlation from the data set.
3. Plot validation matrix to check for spurious values.
   1. Manually clean data until validation matrix returns no spurious values.
      1. Sort columns with suspect columns to find spurious values and remove.
4. Plot line graph for each column.
   1. Investigate any outliers and remove if appropriate.
5. Save dataset as csv.