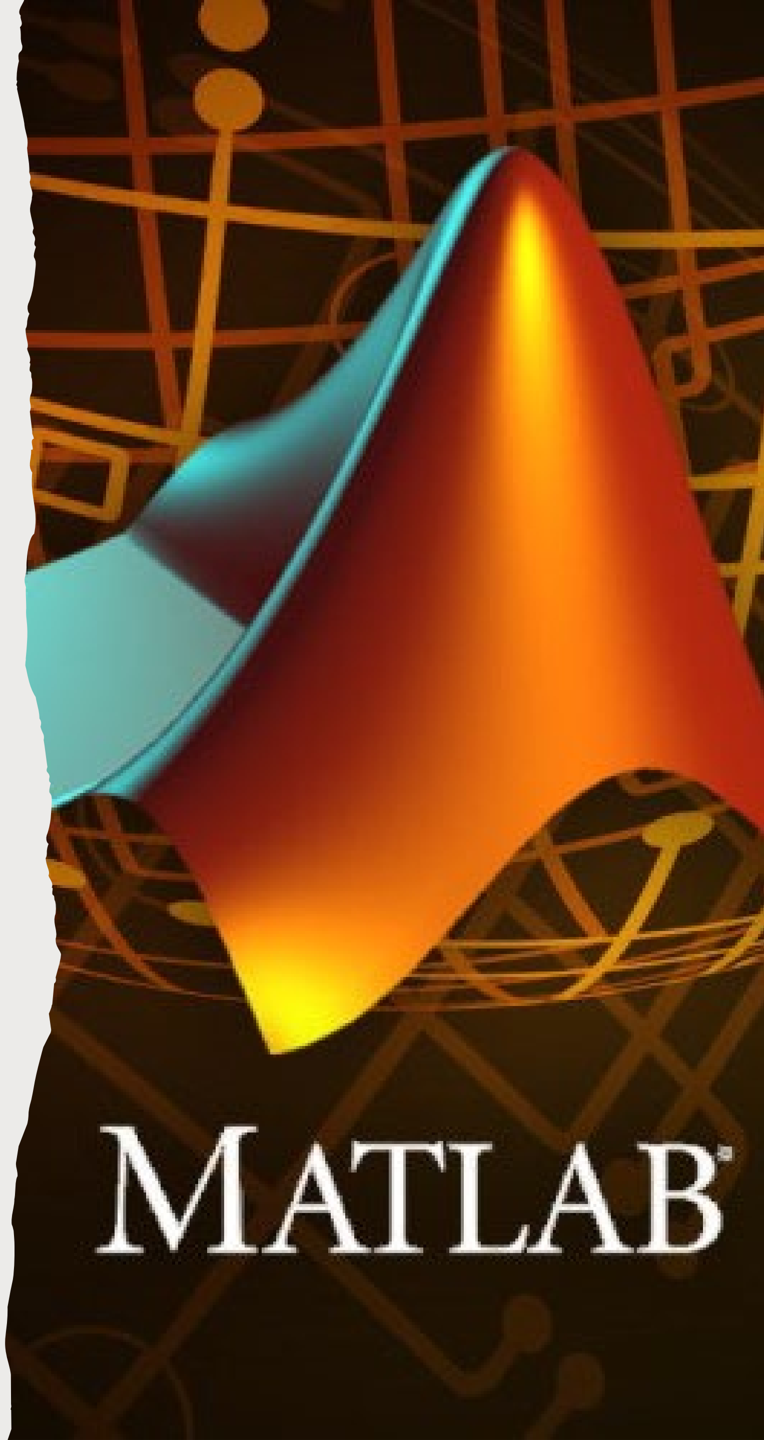
Data analysis for Geoscience

GEO8026

Dr. Matt Perks



Measurement uncertainty

Upon data acquisition, we need to be confident that the reported value represents the physical quantity of interest.

Quality assurance represents the systems in place to assure measurements are as accurate and as precise as possible. QA is a planned and systematic means for ensuring that the defined standards, practices, procedures, and methods are applied.

In lab environments, this is partly achieved through quality control measures such as using regular calibration and analysis with standards. This guards against systematic errors that may generate bias.

Random errors arise from unpredictable variations which influence the measurement procedure, and are associated with the actual measurement e.g. xrf – water content, organic content.

Lab-based procedures should be designed to account for these sources of error in their methodologies.

This is particularly important when the distance between measurements may be large e.g. core sampling where error detection may be more problematic (more later)

Measurement uncertainty

However, what about when a sensor/measurement device is deployed?

Systematic errors can be evaluated and accounted for by sensor (re-)calibration e.g. sensor drift.

But what about random errors? These are difficult to control and account for outside of a lab e.g. environmental interference.

Deployed sensors typically sample at high frequency/resolution so we can often use other sources of information to identify and ‘flag’ spurious measurements.

Therefore, we may opt for a Type A and/or Type B evaluation:

- Statistical analysis of repeat measurements, or calibrations
- Previous measurement data;
- Experience with, or general knowledge of, the behavior and property of relevant materials and instruments;
- Manufacturer's specifications.

Handling datasets

Importing files

Often, files will be ASCII format e.g., DAT, .CSV, and .TXT.

Delimited using comma, spaces, tabs, etc.

However, some may use proprietary formats e.g. xlsx

We will explore how to handle these common data formats and to assess the quality of the data within

netCDF is an increasingly popular format for scientific variables

All of these can be easily imported into MATLAB

For ascii

```
C = textscan(fileID,formatSpec)
```

For ascii, xslx

```
C = readtable(filename)
```

For netCDF

```
C = netcdf.open('example.nc')
```

Running quality control

Assessing data quality and flagging issues: some examples

Flag (numeric value)	Description
Pass = 1	Data have passed QC tests
Not evaluated = 2	Data have not been evaluated
Suspect or high interest = 3	Data are considered suspect or high interest. Flagged to draw attention to users
Fail = 4	Data have failed one or more QC check. Not of acceptable quality
Value changed = 5	Data have been adjusted following correction
Interpolated = 8	Missing/bad data have been infilled by interpolation methods
Missing data = 9	Data are missing

Running quality control

Timing or gap tests

A test to determine whether the most recent data point was measured at the expected time

In some cases, data may not report at regular intervals so this should be considered

The gap check is not a solution for all timing errors.

Data could be measured or received earlier than expected.

This test does not address all clock drift/jump/synchronization issues.

Example Flags	**Condition**
Fail = 4	Reported value does not arrive on time
Pass = 1	Test meets the condition

Running quality control

Gross range test

All measurements are made by sensors that have a limited output range, and this can form the most rudimentary gross range check.

No values less than a minimum value or greater than the maximum value that the sensor can output are acceptable.

Additionally, the operator can select a smaller span based upon local knowledge or a desire to draw attention to extreme values.

For example, we know that measurements of many environmental variables must be non-negative, such as water depth, and concentrations of major anions and cations in environmental samples.

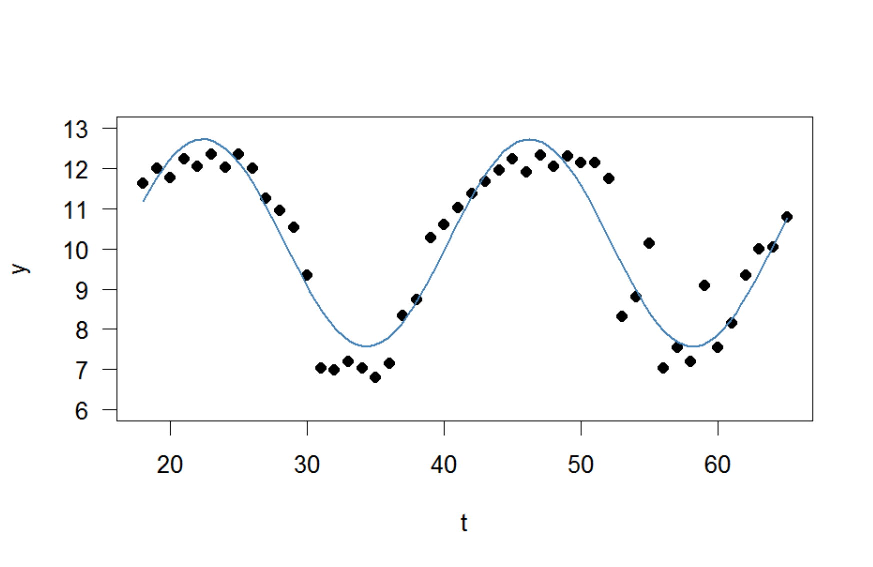
Example Flags	**Condition**
Fail = 4	Reported value is outside of sensor span.
Suspect = 3	Reported value is outside of operator-selected span.
Pass = 1	Within measurement range

Running quality control

Seasonality test

This is a variation of the gross range check and uses thresholds that are applied over differing intervals (e.g. monthly, seasonal) to identify whether measurements are within a physically reasonable range

The upper and lower thresholds may be set based on user-knowledge or use secondary variables to identify a likely range. For example, in Winter a physically probable temperature range may be -20 to +10 °C, whereas in Summer this may shift to -5 to +30 °C.



Flags	**Condition**
Suspect = 3	Reported value is outside of operator-selected span.
Pass = 1	Within measurement range

Running quality control

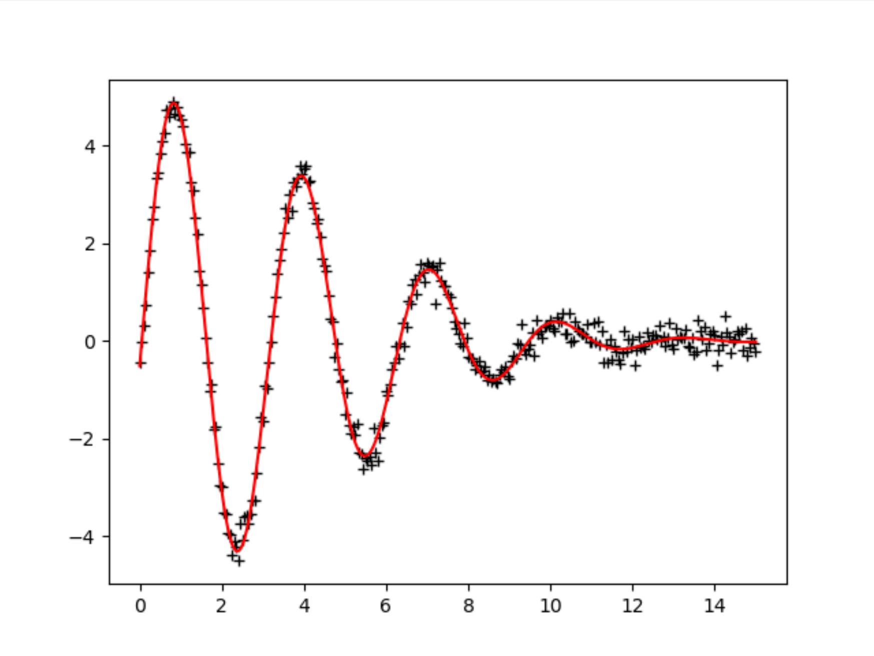
Attenuated Signal Test

In some instances, external factors can affect the performance of the sensor. This may result in an attenuated signal which is not representative of conditions. An example of this would be the production of a nearly flat line.

This response may be indicative of the sensor not sampling the desired material e.g. located too far away from object, or sensor out of water

Alternatively, in environments where we might expect cyclical signals (e.g. tides), these may be severely dampened. If these are dampened or non-existent, this may be indicative of an error.

Flags	Condition
Suspect = 3	Reported value appears to be attenuated.
Pass = 1	Sufficient change between



Running quality control

Rate of Change Tests (1)

This test inspects the time series for a rate of change that exceeds a threshold value identified by the operator. This could be an absolute value, percentage change, or standard deviation of the mean.

A balance must be found between a threshold set too low, which triggers too many false alarms, and one set too high, making the test ineffective.

Determining the excessive rate of change is based on user-experience and/or sensor sensitivity.

Flags	**Condition**
Suspect = 3	Reported change is greater than would be expected
Pass = 1	Change below the maximum expected

Running quality control

Rate of Change Tests (2)

Single value spikes are relatively easy to detect

Spikes consisting of more than one data point are difficult to capture, but their onset may be flagged by the rate of change test.

Adjacent points (n-1 and n+1) are averaged to form a spike reference. The absolute value of the spike is tested to capture positive and negative spikes. Large spikes are easier to identify as outliers and flag as failures. Smaller spikes may be real and are only flagged suspect.

The thresholds may be fixed values or dynamically established (for example, a multiple of the standard deviation over an operator-selected period).

For example, we may state that the rate of change between two temperature measurements must be less than 3σ over the previous 24-hours

Flags	**Condition**
Fail = 4	Very large spike reported.
Suspect = 3	Elevated spike reported.
Pass = 1	Within expected range

Running quality control

Rate of Change Tests (3)

$$Z = \frac{X - \mu}{\sigma}$$

BUT… What happens if we are not dealing with single spikes? But perhaps spikes lasting across several steps?

If we use the approach of (1) + (2), rate of change may be low in erroneous data after the initial shift

In these cases, we can use the z-score to assess the significance of changes and detect spurious values

Whilst the μ and σ could be calculated for the entire dataset with values exceeding a Z score being removed, this is not very useful for geoscience data (e.g. periodicity, non-linear dynamics)

We want to calculate this dynamically across the data and not use the global mean and standard deviation

We can therefore use additional terms:
A **lag** or moving window which will be used to smooth the data
A **threshold Z** value at which error is reported
An **influence** indicating the effect of new signals on μ and σ

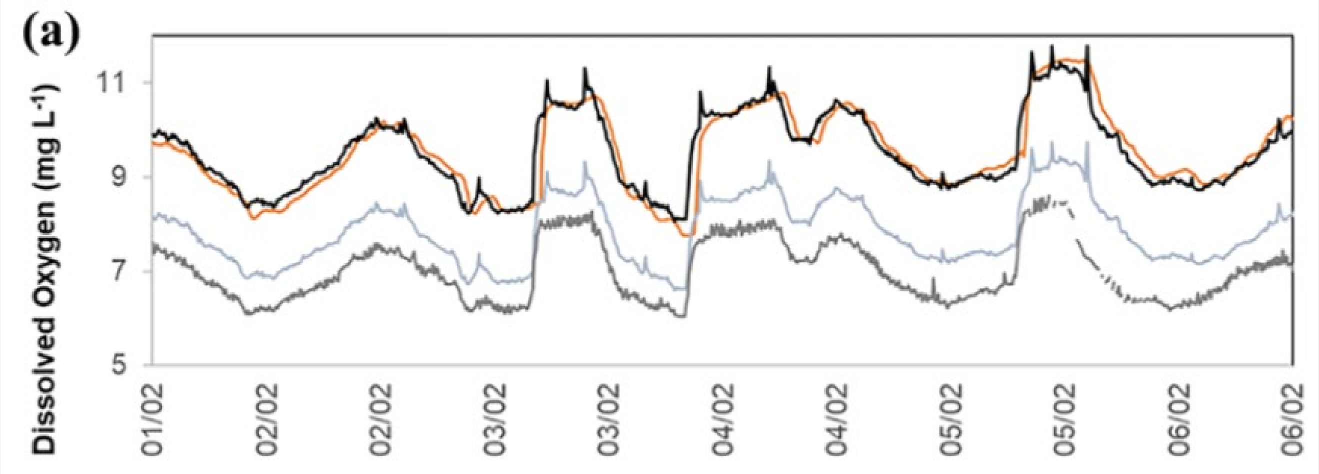
Running quality control

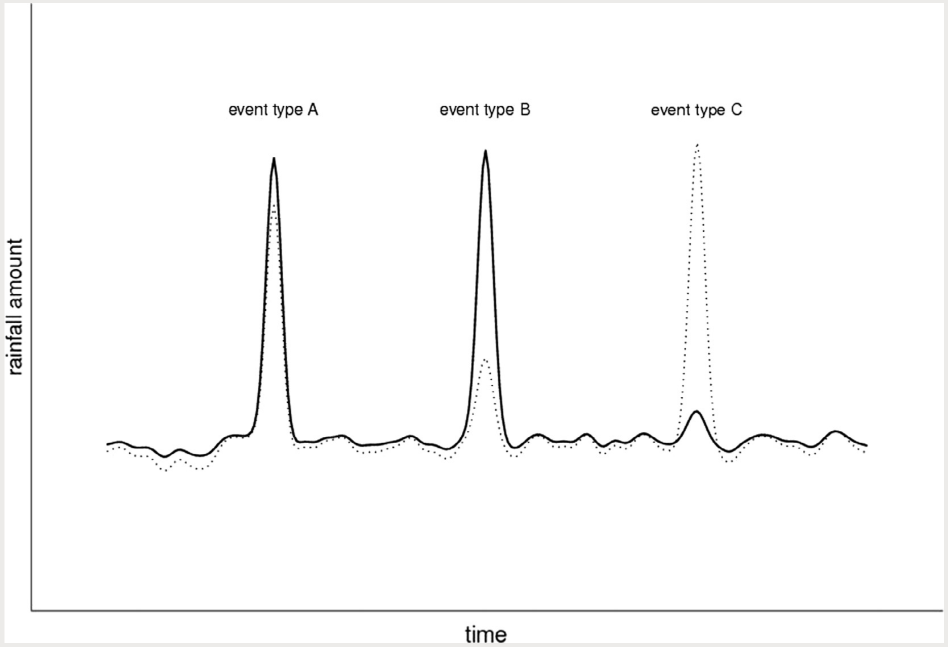
Neighbour test

In some instances, it may be possible to deploy multiple sensors at the same location to provide an additional check on the sensor performance.

Are duplicate sensors offering data within acceptable uncertainty limits?

Flags	**Condition**
Fail = 4	Significant deviation between sensors.
Pass = 1	Within measurement range





Running quality control

Multivariate test

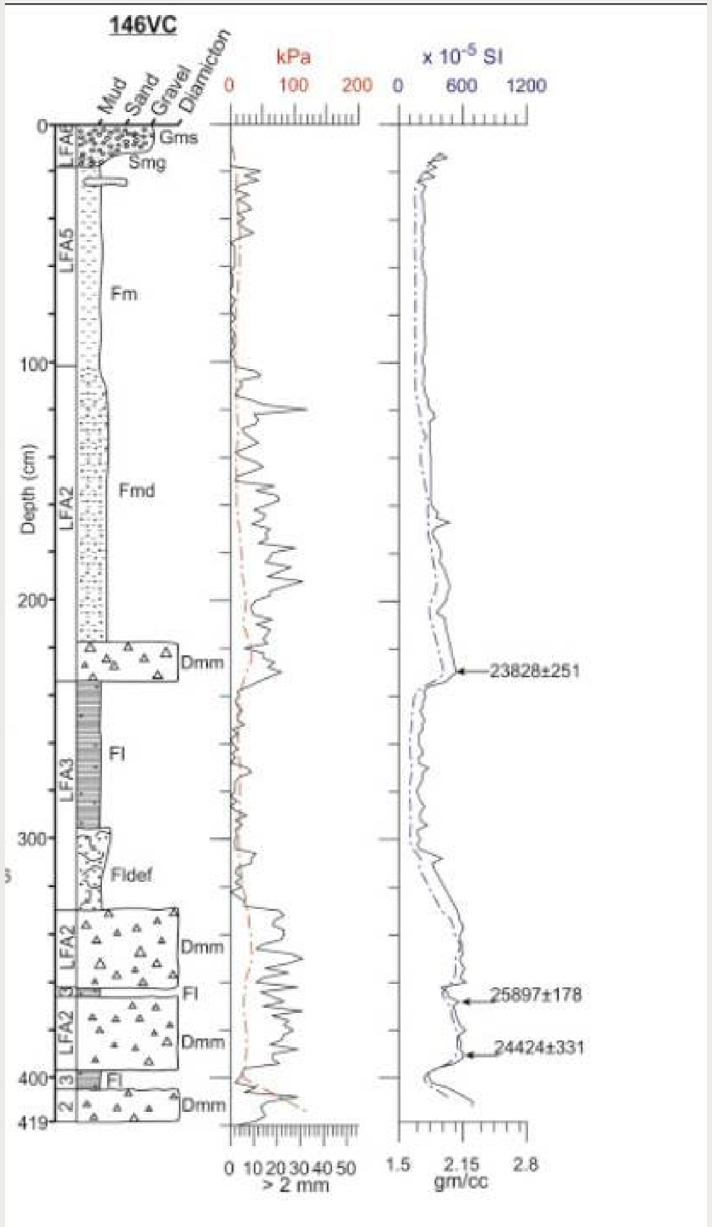
Where multiple measurements are made at the same point in space/time it may be possible to develop relationships between two or more variables. For example, relationships between total dissolved solids and conductivity, or bulk density and particle size.

By analysing the covariance, it may be possible to identify times where the relationship between these variables weakens.

If a deviation in the expected differences exceeds a threshold (percent or absolute), we may flag this data as being suspect.

Can only be reliably applied when the covariance between variables is high and physically stable (a high correlation coefficient exists).

Flags	Condition
Fail = 4	Measured value >> greater than predicted
Suspect = 3	Measured value > greater than predicted.
Pass = 1	Within expected range



Data export

Once we have the data readily processed we can consider how to export the flagged data

For ascii

```
writecell(C,'C_tab.txt','Delimiter','tab')
```

For ascii, xslx

```
writetable(C, filename)
```

For netCDF

```
ncid = netcdf.create(filename,cmode)
varid = netcdf.inqVarID(ncid,'temperature');
data = [100:109];
netcdf.putVar(ncid,varid,0,10,data);
netcdf.close(ncid);
```