NETLAKE toolbox for the analysis of high-frequency data from lakes

Factsheet #1

Data handling: cleaning and quality control

Elvira de Eyto and Don Pierson

Objective

The objective here is to describe some of the procedures that can be used to process high frequency monitoring (HFM) data to ensure that obvious errors have been removed and that data can be considered quality controlled. Some examples from two long running monitoring stations are discussed.

Specific application

HFM brings immediate gratification in the form of megabytes of data but without quality assurance /quality control (QA/QC) procedures, the confidence in these data will be reduced. Some variables require less care than others, but all variables need to be checked and verified, particularly if the data are being used externally and/or shared openly. Quality indicators are a useful way of informing users to what level QA/QC has taken place. Here is one example, which has been developed for the Lake Erken monitoring station:

Level 0: data straight off the logger as ASCII text files. It is critical that level 0 data is always archived for future reprocessing use.

Level 1: Checked to ensure that all expected time steps or file rows are in the file, even if they only contain missing values. Obvious outliers have been marked or removed and some maintenance log comments added. This is usually the minimum processing before sharing of data.

Level 2: Data are corrected for drift, sensor calibration, compared with neighbouring sensors and corrected accordingly. This level of analysis can be done on a by-needs basis by researchers working with the data, or can be part of a more regular QA/QC program. Typically level 2 processing requires supplementary information such as sensor calibration data.

In some instances, getting data from level 0 to level 1 can be done manually, with checks and comment additions done by experienced personnel. However, some of these processes can be automated. A great advantage to automated processing is that as QA/QC algorithms are improved the level 0 data can easily be reprocessed to an improved level 1 state.

Background

This process is quite specific to each monitoring station, and much will depend on how and where data are stored, whether there are in-built quality checks, and what the intended use of the data is. The specific steps described below are what is carried out at the Irish Marine Institute's research facility in Burrishoole, where five automatic monitoring stations are maintained. Raw data are transmitted from the stations' data loggers via GPRS every couple of minutes to a computer in the research lab. Owing to security issues with institutional firewalls (which is a common problem when transmitting data), this computer is isolated from the main institutional servers. Approximately once a month, these data are copied across to an intermediate storage home on a server which is backed up regularly.

A crucial decision to make at this stage is how you want to store the HFM data long term. In Burrishoole we have decided to store the HFM data at level 1, after a couple of fairly basic checks (described below) and additions. We also store all the level 0 files. Thus, the permanently stored files (which are saved on a SQL server) contain data which have had no data deletion or manipulation. If a data request is lodged with us, the requestor receives these data, with the caveat that the data are at level 1 (see above). We decided that this was the way for us to store and share data after observing how subjective data cleaning and manipulation was. What one person considers to be "bad" data (and may delete from the file), another person, in hindsight, may think is retrievable. This is an important consideration when the intention is to run the station for years or decades, and where it is likely that the staff in charge of the data are likely to change. Of course, where significant data cleaning has been carried out on a particular variable, these data (level 2) will also be stored (and shared if applicable), but in a separate location to the long term level 1 data storage.

Type of data and requirements

Any remotely collected data needs some QA/QC before use. Software that is useful includes MS excel, R, Python, B3 (https://www.lernz.co.nz/tools-and-resources/b3) or Hydras (http://www.ott.com/products/software-solutions/ott-hydras-3-basic/).

Basic procedures

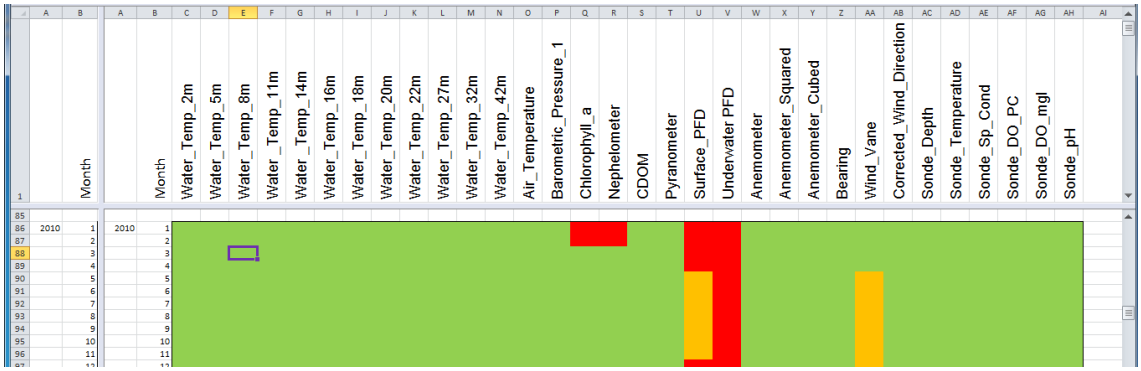
Here we describe how we bring the Lough Feeagh AWQMS (automatic water quality monitoring station) data from **level 0 to level 1**. This is provided only as an example, and should be modified according to your requirements.

1. Join text files containing chunks of data together into an annual file. When data are stored every 2 minutes, an annual file is still manageable with any spreadsheet like Excel for example (262,000 rows). Once you go past this resolution (e.g. minute files, or multiple years), Excel is no longer useable, and manipulation and data viewing probably needs to take some other route.
2. Check for missing time steps and fill in where needed. This is only essential if you expect data to have a regular time step. Having a regular time step makes it easier to aggregate

data from different stations and sensors later, but is not essential. Some useful ways of doing this include:

- Use a pivot table in Excel, with day or date being the aggregating variable (720 measurements per day, 262,800 per year). A day with less than 720 values is easy to spot.
- The Zoo package in R has a function for merging a pre-described time step with a dataset where there are gaps.

3. Fill in blanks with NA. This is for later use in R. Other programs may require blanks to be coded differently (e.g. -999, or simply blank).
    - This can be done in Excel – fill blanks or replace, but can be slow if there are a lot of them.
4. Check for outliers
    - We use the filter in Excel. If an outlier is spotted, we will add a note to that row rather than deleting the value.
5. We fill in comments retrospectively. An extra column is added to the data sheet, and comments are transferred from our field book to the relevant row or rows. The comments might include:
    - Sensors cleaned.
    - Sonde out for calibration.
    - Fluorometer removed for service.
    - Mooring rope replaced.
    - Anemometer looks very high. Check against the manual weather station before use.
    - Batteries flat.
6. We normally extract a subset of data for each day (e.g. 06:00, 12:00, 18:00, 00:00)
    - Do some quick graphs of what things look like. For this purpose simple Excel templates can be prepared into which data is pasted and graphs are automatically generated.
    - Do an informal report on what the year's data look like.
    - Fill in the sensor information plot – this is a simple excel log giving a quick overview of which sensors were working at any time period.



7. Based on the summary plots and the informal report, make some additional comments to the level 1 data.

8. Upload this level 1 data to our long term storage SQL server.

The next step is to get data from **level 1** to **level 2**, which is not done routinely, but on a project basis as data are requested.

1. Make a copy of the level 1 data which can then be changed or manipulated.
2. There are a couple of programs which can be used at this stage; in Burrishoole, we use Hydras or R. Another option is B3. Excel is useful also for some things. At this stage, we would do things like:
    - fill in sensor gaps with interpolated data from another source if applicable
    - Apply temperature corrections to CDOM sensor data
    - Correct data points for sensor drift (e.g. CDOM, chl fluorometer, DO)
    - Do a more detailed analysis for outliers
    - Aggregate the data where required. A useful tool for this is the R package hydroSTM
    - Compare one sensor against another to check for drift or odd values (e.g. multiple thermistors)
    - Remove "bad" data and replace with NA

Pitfalls and tips

- Losing data. This is very common, as files multiple up very quickly. Have a structured folder system, perhaps ordered by site and year, or by sensor. File names should have some logical meaning and be consistent so that they can be sorted as required.
- Overwriting raw data. You might feel the need to correct values as you are sure they are outliers. Then you discover that actually the data were fine, and were, for example, recording an episodic event. Always keep an original version of the level 0 data. Any data manipulations should be made on a copy of these data, and only change data in the copy. The logger text files (level 0) are usually quite small, and you can always store these logger files, along with level 1 and level 2 data.
- Overzealous data cleaning….leaving you with no data!
- Thinking a sensor is working (because the values are changing), but subsequently realising that the logger is just recording some residual current.
- Mixing up your data files from different sites, or different times of data collection. We recommend creating a variable in your data logger program to identify the data logger location and the data logger program version. These variables can be outputted on a daily basis along with other diagnostic information such as logger battery voltage. Having the site and the program that created a file documented in the file itself will prevent location mix-up and will also be of use in linking data changes to program changes.

Further reading

1. http://www.ott.com/products/software-solutions/ott-hydras-3-basic/
2. https://www.lernz.co.nz/tools-and-resources/b3
3. https://cran.r-project.org/web/packages/zoo/zoo.pdf
4. https://cran.r-project.org/web/packages/hydroTSM/hydroTSM.pdf
5. http://www.gleon.org/data/best-practices

Contact details

Elvira de Eyto. Burrishoole research station, Marine Institute, Ireland.
elvira.deeyto@marine.ie

Don Pierson. Lake Erken field station. Uppsala University, Sweden.
don.pierson@ebc.uu.se

Suggested citation

de Eyto, E. and Pierson, D. 2016. Data handling: cleaning and quality control. In Obrador, B., Jones, I.D. and Jennings, E. (Eds.) *NETLAKE toolbox for the analysis of high-frequency data from lakes* (Factsheet 1). Technical report. NETLAKE COST Action ES1201. pp. 2-6. http://eprints.dkit.ie/id/eprint/532.

Acknowledgement

This factsheet is based upon work from the NETLAKE COST Action ES1201, supported by COST (European Cooperation in Science and Technology).