Hands-on Activity 5: Data Quality Control & Assurance

Associated DataONE Lecture: Lesson 5: *Data Quality Control and Assurance* (Note that instructors and students should have reviewed Lesson 4: *Data Entry and Manipulation* beforehand.)

Objectives: Students recognize the importance of organizing data so that others can understand them, by experiencing the challenges of reviewing data sheets that are organized in a common but inappropriate way, and they get hands-on experience in improving them.

Outcomes: Students strategize about ways to improve data organization.

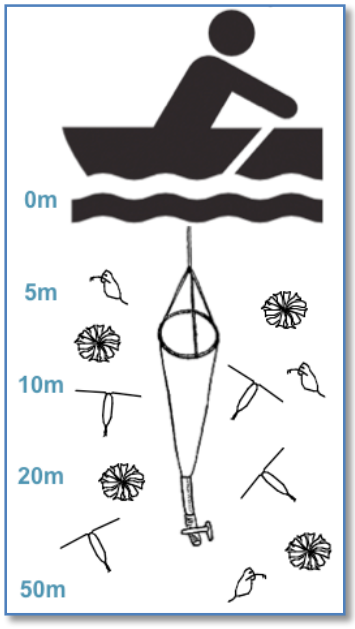
Time Needed: 45 minutes in class.

URLs: None.

Additional Files Needed: pond2010.xlsx, zoop- temp-main.xlsx; zoop-temp.xlsx

Key Readings:
Borer, E.T., Seabloom, E.W., Jones, M.B., Schildhauer, M., 2009. Some Simple Guidelines for Effective Data Management. Bulletin of the Ecological Society of America 90, 205–214.

Campbell, JL, Rustad LE, Porter JH, Taylor JR, Dereszynski EW, Shanley JB, Gries C, Henshaw DL, Martin ME, Sheldon WM et al.. 2013. Quantity is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data. Bioscience. 63:574-585



Notes and Instructions for Instructors:

Data Quality Control and Assurance

Background: Plankton are microscopic organisms that form the base of many aquatic food webs – fueling the growth of fish and other larger organisms. It's common to sample them using a net or another container that can be controlled to collect water just from certain depths; so you can see how plankton collected at the surface (0 meters) might be different from plankton at another depth (e.g. 10 meters below the surface).

(For more information: http://en.wikipedia.org/wiki/Phytoplankton and http://en.wikipedia.org/wiki/Zooplankton.)

They are identified and counted under a microscope, and

usually their numbers are reported as individuals per liter or milliliter.

Frequently, aquatic scientists collect plankton samples during both day (e.g. noon) and night (e.g. 2am) because plankton change their distributions from day to night, and not all species alter their distributions in the same way. (For more information, search "diel vertical migration" on the web.)

You should have 3 (fictional) data files: pond2010.xlsx, zoop-temp-main.xlsx; zoop-temp.xlsx.

These 3 files were all intended to be part of the same study – the investigators wanted to examine the day-night distribution of 2 species of zooplankton across multiple years. The type of zooplankton they studied is called rotifers generally, and specifically the genus *Conochilus,* in which groups of individual rotifers stick together in colonies (see http://eol.org/pages/43393/overview). The investigators plan to repeat this study for several more years.

The files have some problems in how they are organized, which you have already discussed in exercise 4. Now let's think about assuring and controlling the data quality.

Activity 1

In small groups, open the 3 files and inspect them. It may be easiest to split this work up among 3 people. Plot the data to look for anomalous values, e.g. a scatter plot (x-y), or identify the maximum and minimum values in each column. These are easy ways to get a sense for where problems in the data may have been created, for example, where there have been mistakes in data entry or migrations, or problems with equipment.

Some example answers:

- Chippo #/L in zoop-tmp.xlsx has a few negative values that are impossible. They should be flagged.
- Temperature in zoop-temp-main.xlsx has one large outlier – it should be flagged and the scientists who collected the data should evaluate it, if possible.

Activity 2

Suggest a system for flagging data as anomalous.

Example answer:

- A separate column can be created, e.g. "temp_flag", in which codes are entered that correspond to the reasons that a value may be flagged.

Activity 3

Suggest a system for flagging data as missing.

Example answer:

There are several ways to do this. If missing values are left blank in the data column, a flag column can indicate that they are missing.

Some researchers choose to fill in missing values in the data columns with various standard values, such as "NA" or "-999999". Discuss the merits and potential challenges of approaches that are suggested. The convention chosen depends on the data – you wouldn't want "-9999999" if the actual values were in the same range of numbers! And similar would happen with "NA" if you had a list in which North America were abbreviated as "NA".

Date	Time	NO3_N_Conc	NO3_N_Conc_Flag
20081011	1300	0.013	
20081011	1330	0.016	
20081011	1400		M1
20081011	1430	0.018	
20081011	1500	0.001	E1

M1 = missing; no sample collected

E1 = estimated from grab sample

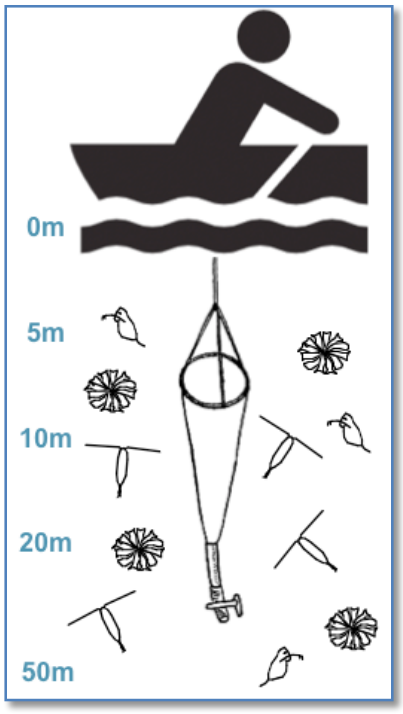
Activity 4

Suggest a system for data entry that can be used in the future to prevent data entry errors.

Example answer:

Trainees should have lots of good ideas by now. They may suggest things such as: constraining the data entry choices to avoid misspellings (such as having pull-down lists); having two people independently enter data; using text-to-voice to have computer read back the value that is typed; limiting the entry field to a specific data type (date, text, etc.).

Student Instructions:



Data Quality Control and Assurance

Background: Plankton are microscopic organisms that form the base of many aquatic food webs – fueling the growth of fish and other larger organisms. It's common to sample them using a net or another container that can be controlled to collect water just from certain depths; so you can see how plankton collected at the surface (0 meters) might be different from plankton at another depth (e.g. 10 meters below the surface).

(For more information: http://en.wikipedia.org/wiki/Phytoplankton and http://en.wikipedia.org/wiki/Zooplankton.)

They are identified and counted under a microscope, and usually their numbers are reported as individuals per liter or milliliter.

Frequently, aquatic scientists collect plankton samples during both day (e.g. noon) and night (e.g. 2 am) because plankton change their distributions from day to night, and not all species alter their distributions in the same way. (For more information, search "diel vertical migration" on the web.)

You should have 3 (fictional) data files: pond2010.xlsx, zoop-temp-main.xlsx; zoop-temp.xlsx.

These 3 files were all intended to be part of the same study – the investigators wanted to examine the day-night distribution of 2 species of zooplankton across multiple years. The type of zooplankton they studied is called rotifers generally, and specifically the genus *Conochilus,* in which groups of individual rotifers stick together in colonies (see http://eol.org/pages/43393/overview). The investigators plan to repeat this study for several more years.

The files have some problems in how they are organized, which you have already discussed in exercise 4. Now let's think about assuring and controlling the data quality.

Activity 1

In small groups, open the 3 files and inspect them. (It may be easiest to split this work up among 3 people.) Plot the data to look for anomalous values, e.g. a scatter plot (x-y), or identify the maximum and minimum values in each column. These are easy ways to get a sense for where problems in the data may have been created, for example, where there have been mistakes in data entry or migrations, or problems with equipment.

Activity 2

Suggest a system for flagging data as anomalous.

Activity 3

Suggest a system for flagging data as missing.

Activity 4

Suggest a system for data entry that can be used in the future to prevent data entry errors.