Streamline QA/QC for Observational Data

Li Kui[1], Kristin Vanderbilt[2], John H. Porter[3]
1. University of California, Santa Barbara, CA lkui@ucsb.edu
2.Florida International University, Miami, FL krvander@fiu.edu
3.University of Virginia,Charlottesville, VA jhp7e@virginia.edu

ABSTRACT

Background/Question/Methods

Observational data, a form of data observed or measured by humans, has been used widely in a variety of disciplines to gain first-hand knowledge on target objects in their natural setting. The workflow for processing observational data typically involves data collection on a paper survey sheet, transfer to a computer, QA/QC, and production of a data product. Because of data handling by humans and human interactions with software, e.g. Excel, Google Sheet, Pages or a database, there is a significant chance that human errors will be introduced at both observation (in the field) and data transfer stages. This creates needs for an error checklist, containing the types of errors and how to find them, and tools to assure high quality data.

To improve data quality, we describe best practices for QA/QC of observational data based on our experiences with datasets from the Long-Term Ecological Research (LTER) Network. Potential errors and their causes at each step of data processing are summarized. The corresponding recommendations with examples are provided for each of the data processing steps.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation to the Santa Barbara Coastal LTER project (OCE-1831937), Florida Coastal Everglades LTER project (DEB-1832229), and Virginia Coast Reserve LTER project (DEB-1832221).

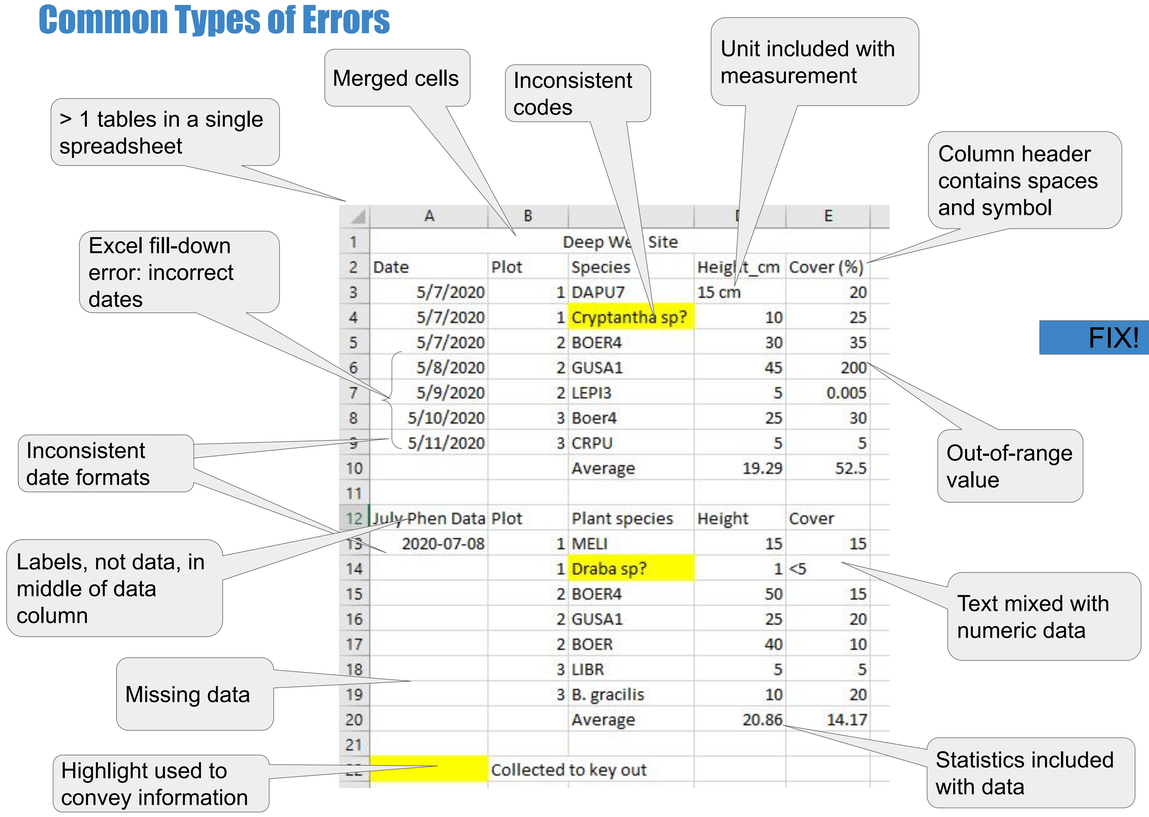
INTRODUCTION

Errors may arise when data are transferred from datasheets into the computer. Also, the way in which data are organized upon entry can greatly affect the ease with which they are analyzed.

Here we illustrate, using MS Excel, some types of errors that can arise when appropriate QA/QC and data organization strategies are not implemented during data digitization. We also show the preferred format for the same data. This format supports further vetting of the data using automated programmatic solutions.

To assure high quality of observational data, we suggest best practices for the QA/QC process.

Common Types of Errors

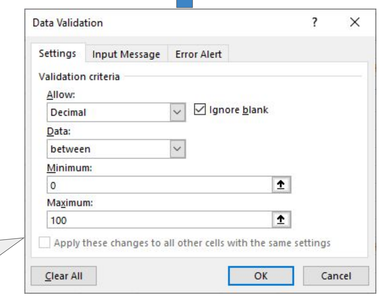


Well-formatted and Clean Data

FIX!

Date	Site	Plot	Species	Height	Cover		
2020-05-07	DeepWell	1	DAPU7	15	20	Notes	
2020-05-07	DeepWell	1	CRYPT	10	25		
2020-05-07	DeepWell	2	BOER4	30	35	collected to key out	
2020-05-07	DeepWell	2	GUSA1	45	20		
2020-05-07	DeepWell	2	LEPI3	5	0.005		
2020-05-07	DeepWell	3	BOER4	25	30		
2020-05-07	DeepWell	3	CRPU	5	5		
2020-07-08	DeepWell	1	MELI	15	15		
2020-07-08	DeepWell	1	DRABA	1	0.005		
2020-07-08	DeepWell	2	BOER4	50	15	collected to key out	
2020-07-08	DeepWell	2	GUSA1	25	20		
2020-07-08	DeepWell	2	BOER	40	10		
2020-07-08	DeepWell	3	LIBR	5	5		
2020-07-08	DeepWell	3	BOGR2	10	20		

Use of “format cells” to organize date and time into consistent format



Use of validation tools in spreadsheets can help assure 1) that codes are consistent, 2) that values fall within valid ranges

BEST PRACTICES

A. Use software and tools that detect errors in the data transfer process from paper to computer
- MS Excel or google sheet with validation rules
- MS Access to create data entry forms
- Create a backend database to allow data input on website with restricted controls across multiple survey tables

B. Implement programs/scripts that streamline and automate the QA/QC process in a reproducible way
- R packages such as Datamaid
- OpenRefine
- Python
- SQL

C. Format and organize data to increase the reusability of the data
- One variable per column, one observation per row
- Long format instead of wide format
- Zero-filled for closed species list data
- Export data in CSV or TXT format with column headers included

CONCLUSION

These QA/QC recommendations should be implemented in all observational data processing workflows. Once the data are entered into a computer, depending on the staff’s technical skills and knowledge about the data, error checking can be conducted within one software (e.g., programming language such as R, Python, or integrated database structure) or through multiple platforms (e.g., from Excel sheets to any programming language). Most importantly, the development of community-wide data processing procedures are essential for QA/QC that would instill confidence in observational data and improve data interoperability within a scientific network.