VolGrabData.RProj

# Summary

The VolGrabData.RProj creates a version controllable set of R scripts for processing grab water quality data submissions from ODEQ monitoring partners to the volunteer monitoring program. The processing includes cleaning up submitted Excel workbooks with instantaneous water quality results so the file can be loaded into R scripts that will eventually load the data into the database used for volunteer data storage by the volunteer program

# Processes

## Project Information

The VolWQdb must have project, station and data submission information in the respective tables in order for database dependent portions of the code to work. This includes making sure that valid values are available for formatting the Excel file.

## Excel File Formatting-

Required fields and format for metadata need to be applied in Excel. Both a project info and data tab are required.

## Rcode: Loading Data-

### User inputs:

Requires the input by the user of details about the data submission ID, the organization conducting the activities, the type of batching used for QC, the directory for the input and saving of files, and the file path. The user provides specific information about which worksheets to include and the starting and ending rows of data in each.

### File loading

The XLConnect package is used to load the workbook and then read the specified worksheets into the ‘gd’ dataframe for grab data and the ‘prj’ dataframe for project data. A number of steps are then taken to add fields and make sure the existing fields are of the correct data type (numeric, POSIXct, character, factor).

### Duplicate Data Procedures

A series of processes to remove any duplicated data based on (1) all columns for two rows match; (2) result and DateTime + Station match. Activity root field is qualified with a ‘b’ pasted to the end of the activity root if more than one set of rows are found with matching DateTime/Station combos are present. Each step exports a csv file with the redundant rows of data.

### Reshape messy flat file to Tidy Columnar Format

Take the wide format from standard Excel data submission file and stack each characteristic on top of each other with all the corresponding metadata for each result. Duplicate pairs still in the same row.

### Create QC Data Constants, Criteria, etc.

Import the QCcrit.csv file from the file directory and calculate the correct type of QC calculation for the duplicate pairs. Report precision value and QCcalculation type and ODEQ\_DQL in fields for each duplicate.

### Complete Tidy Process

Assign duplicates activity type of FD and move to gdl format, then rbind samples, field primaries and field duplicates together.

‘gdtidy’ dataframe is then ready to be used for summaries and plots. Possible break in the processing where a file subID-gdtidy.RData could be saved to the folder.

## Rcode: Data Summaries

Create tables of data summaries for each characteristic for all stations combined and for each individual station. Export a data summary text file.

Create QC summaries about the distribution of duplicates and % of results at different DQL’s. Export QC Summary text file

## Rcode: Create Plots

### Prep

Load CharNames.RData with unis and descriptors for each of the characteristics to be plotted.

Identify which characterstics should NOT be plotted on a log scale

### Characterstic Box Plot v. Station

For each parameter a box plot is displayed for each station.

### Duplicate Differences v. Time

Point charts for absolute difference v Time or RPD v Time (1 or 2 charts per parameter)

DQL criteria lines plotted based on characteristic and QCcalc method,

### Duplicate Differences v Concentration

Field primary plotted against field duplicate. DQL criteria lines are created across the range of concentrations

Another logical break in script includes Data summaries and plots together. No changes to the gdtidy. Need gdtidy, QCcrit and CharNames as inputs.

## Rcode: VolWQdb Prep and Loading

### Generate Activity and Result ID’s

Pulls required fields from gdtidy. Determines if lab or field characteristic from ‘prj’ dataframe. Converts activity types to storet codes and creates activity ID and then result ID.

### Create Result Table upload file

Pull required fields from prj dataframe and then merge with dataframe containing result ID’s (act). Final formatting of column names and then save a text file of the result table uploads.

### Create Activity Table upload file

Check to make sure there is only one result for each characteristic per activity ID. Then summarize the result table by the activity ID for the activity table fields. Then add up to 15 fields based on matching up actroot field in gd dataframe and actinfo dataframe. Clean up the formatting. Save text file of activity table.

### Rcode ODBC: Load Activity and Result tables.

Open the channel to VolWQdb. Run sqlSave to append to VolWQdb.t\_Activity table.

Save t.results to a temporary new table in VolWQdb then run SQL query to append the temporary table to the VolWQdb.t\_Result table. Then delete the temporary table.

Generate Activity Group and Junction table to Activity Table

Activity Group ID’s are differently depending on the ‘dbatch’ method specified. Series of if statements used take unique values from one of the three methods below. Day batch actgrp is submission ID and date. Day+Crew is day batch plus the DupBatchKey. For Sampler batch the actgrp is the dupbatch key. Save text file of output and save to VolWQdb.t\_ActGrp. For each batch type the actgrp ID is partially matched with the fields used to generate the group ID. The t.ActGrp dataframe is saved as a text file and loaded into VolWQdb.tjct\_ActGrp2Act.