Station Identification Process Summary 2014

Locating and Classifying Water Quality Monitoring Stations for the 2012 Integrated Report Re-analysis

Mike Psaris

# Overview

EPA decided not to approve Oregon’s 2012 Integrated Report due to its exclusion of various data sources from analysis. As a result new stations from USGS, EPA, and the LASAR database had to be added, and a new analysis completed. In order to add these new stations, we first had to geo-reference them using the LLID addressing system. This system works by specifying how far upstream from a river’s mouth a station lies. Once the stations were located, each had to be classified as either freshwater, estuary, or marine. This document describes the methodologies, tools, and workflows we used to accomplish these tasks.

# List of Python and R Scripts Used: By Location

[\\Deqhq1\mpsaris\GitHub\ToxicsRedo\Python\_Scripts](file:///\\Deqhq1\mpsaris\GitHub\ToxicsRedo\Python_Scripts)

1. *Assign\_LLID.py*
2. *Assign\_LLID\_Additional\_LASAR\_Stations.py*
3. *Stations\_Analysis.py*
4. *Estuary\_Analysis\_01.py*
5. *Estuary\_Analysis\_02.py*

[\\Deqhq1\mpsaris\GitHub\ToxicsRedo\Estuary\_Analysis](file:///\\Deqhq1\mpsaris\GitHub\ToxicsRedo\Estuary_Analysis)

1. *ConSal\_Data\_Cleanup.R*
2. *Pull\_2010\_StationUseList.R*

# Workflow

To conduct our analysis, we used a mix of automated and manual tasks. Automated tasks were done mostly in Python so we could take advantage of ESRI’s Arcpy spatial analysis tools. For tasks which required tabular data analysis we used the R statistical software environment. Unless specified R scripts were executed inside Python scripts to simplify the process.

Four separate comma separated values (csv) files were provided by Peter Bryant, each with a listing of stations from various government agencies. The first three files were merged and addressed together. The fourth list was created later and therefore had to be addressed separately. The two resulting station lists were then merged and classified as either freshwater, estuary, or marine.

## Data Preparation

### Original Datasets

The three original csv files were combined into one MS Excel workbook located here:

[\\Deqhq1\mpsaris\GitHub\ToxicsRedo\StationsToLocate\FinalList\List\_of\_Stations.xlsx](file:///\\Deqhq1\mpsaris\GitHub\ToxicsRedo\StationsToLocate\FinalList\List_of_Stations.xlsx)

This file has a metadata tab which provides the locations of source files and a description of each worksheet. Each station list was modified to fit one unified format in the “MERGED\_Formatted” worksheet, and then saved as a new csv file located here:

[\\Deqhq1\mpsaris\GitHub\ToxicsRedo\StationsToLocate\FinalList\Master\_List\_of\_Stations.csv](file:///\\Deqhq1\mpsaris\GitHub\ToxicsRedo\StationsToLocate\FinalList\Master_List_of_Stations.csv)

Excel does not handle numbers saved as text well, and some USGS stations which have 15 digit station ID’s would become corrupted once saved as a csv and opened in MS Excel. In order to preserve these IDs the STATION\_ID field was saved as type “special”, and once the worksheet was saved as a csv, the csv would not be modified in Excel again. If changes had to be made, they would be made inside the Excel spreadsheet workbook (.xlsx) and then saved as a csv. This csv was then converted to a shapefile using the *CombineMultipleProjectionfromCSV.py* tool provided by Peter Bryant, so it could be used as input for the *Assign\_LLID.py* script. Due to the way ArcMap determines a column’s type, there was an error when the conversion was run, and each datum had to be converted to a shapefile and then merged.

### Additional LASAR Stations

A different data preparation process was used for the additional LASAR stations. In this case, a csv file with station IDs and descriptions was provided by Peter Bryant, and located here:

[\\Deqhq1\mpsaris\GitHub\ToxicsRedo\StationsToLocate\FinalList\Additional\_LASAR\_Stations\_to\_locate\_06112014.csv](file:///\\Deqhq1\mpsaris\GitHub\ToxicsRedo\StationsToLocate\FinalList\Additional_LASAR_Stations_to_locate_06112014.csv)

With this station list we used a python script to subset the 44 additional stations from the master LASAR station shapefile located here:

\\Deqlead03\gis\_wa\Project\_Working\_Folders\LASAR\_Stations\LASAR\_Stations\LASAR\_Stations\_26sept13.shp

This python script can be found at the beginning of Assign\_LLID\_Additional\_LASAR\_Stations.py

## Assign\_LLID.py (and Assign\_LLID\_Additional\_LASAR\_Stations.py)

These scripts automate the assignment of LLID and RM measures to the stations and perform a basic level of quality control that determines which stations need to be manually inspected.

The script takes the following as input:

1. A routed streams layer
2. A station feature class
3. The desired search radius in feet
4. The desired location and name of the output geodatabase

The script outputs a geodatabase with the following feature classes:

1. “qc\_success” – stations in this feature class snapped to a stream, were assigned an LLID and RM measure, and the station description matched the stream name. These stations are properly addressed but still need to be check for lake LLIDs.
2. “qc\_needs\_review” – stations in this feature class snapped to a stream, were assigned an LLID and RM measure, but the station description did not match the stream name. These stations require manual inspection, and possibly require manual assignment of a stream and/or lake LLID and RM measure.
3. “outside\_threshold” – stations in this feature class did not have a stream within the specified search radius, and therefore possibly require manual assignment of a stream and/or lake LLID and RM measure.

The scripts also add the following new fields

1. LLID – The LLID of the stream the station was snapped to
2. Measure – How far upstream the station lies in feet (ft).
3. Distance – How far away from the stream centerline the station is located. Stations to the left of the upstream direction are positive.
4. QAQC1 – Indicates the outcome of the first round of qaqc.

## Manual LLID and RM assignment

Once the two Assign\_LLID scripts were run, manual adjustments had to be made. To do this, the attribute tables of all output feature classes were exported to a .dbf file. The files were then combined into two excel workbooks – one for the initial station list, and one for the additional LASAR station list.

Three columns were added to each list so the qaqc process could be tracked:

1. Lake LLID – Streams which were located in a waterbody were assigned a lake LLID.
2. QAQC2 – This field describes what needed to be done on the second round of qaqc.
3. Comments – Any important comments were included here.

Once the manual review process was complete, QAQC1 and QAQC2 fields were modified slightly to synchronize all output. The following is the list of possible entries for these two fields:

1. QAQC1
   1. ‘Correct’
   2. ‘Needs Secondary Review’
   3. ‘Outside Threshold’
2. QAQC2
   1. 'Added Lake LLID and removed stream LLID' – for stations located in a lake without an outlet.
   2. ‘Correct’ – Needed no manual adjustment
   3. ‘Further Review Needed’ – Unclear how to proceed. Look at comments for more details.
   4. ‘Likely Correct’ – Some uncertainty, but still highly confident in the assignment.
   5. ‘Manually adjusted with IRL tool’/’Manually Revised with IRL tool’ – LLID and Measure recalculated using the ‘Identify Route Location’ tool.
   6. ‘Manually Revised with LFAR Tool’ – LLID, Measure and distance recalculated with the Locate Feature Along Route tool.
   7. ‘Not Required’ – First round of QAQC was successful
   8. ‘Potential Digitization’ – No stream or lake present in DEQ streams or lakes dataset. These waterbodies may be added to the dataset at a later time.

Once the manual review was complete, a new worksheet for each shapefile was created where field names were adjusted to match those of the DEQ station list for the 2010 Integrated Report. These worksheets were then exported to .csv files. Since the additional lasar stations list was small, the stations were merged together before exporting to .csv. Therefore there were three csv files for the original station list, and one csv file for the additional lasar stations list. The stations were then ready to be merged together with the *Stations\_Analysis.py* script.

## Stations\_Analysis.py

This script takes the various shapefiles and csv files that have been created up to this point and merges them into one master shapefile with field names matching those from the IR 2010 station use list. It also checks to see whether there are any duplicate stations and removes them. The result is located here:

\\Deqhq1\mpsaris\GitHub\ToxicsRedo\StationsToLocate\FinalList\All\_Final.gdb\All\_stations\_final

And has the following fields:

1. LLID
2. LAKE\_LLID
3. AGENCY
4. AGENCY\_ID
5. STATION
6. DEC\_LAT
7. DEC\_LONG
8. DESCRIPTION
9. QAQC1
10. QAQC2
11. Comments
12. RIVER\_MILE
13. HUC\_6
14. HU\_6\_Name
15. HUC\_8
16. HU\_8\_Name
17. GIS\_STREAMNAME
18. LAKE\_NAME
19. GIS\_Source
20. GIS\_Source\_LAKE

# Estuary Analysis

We used the following set of criteria to determine whether a station was located in freshwater, estuary, or marine waters:

1. If the station lies west of the borders of a bay, estuary, or the Pacific coastline, the station is classified as marine.
2. If the maximum salinity concentration for a station is greater than 0.09717282 ppth (200uS/cm), the waterbody is tentatively labeled estuary.
3. Final classifications are based on a station’s geography and proximity to surrounding stations
   1. If an inland station is considered estuary based on Salinity concentration, but has stations labeled freshwater both up and downstream from the station, it’s classified freshwater.
   2. If a station near the coast is considered freshwater based on salinity concentration, but had stations labeled estuary up and downstream from it, it is classified estuary.
   3. If stations do not meet either of these criteria, a mixture of elevation and heads of tide was used to make an informed decision.

## Estuary\_Analysis\_01.py

Since many stations lie inland and are unquestionably freshwater, this script only operates on a subset of new stations. The subset consists of all new stations that lie within HUC8 watersheds that either intersect the pacific coast, or contain 2010 stations that are classified estuary. This subset is located here:

\\Deqhq1\mpsaris\GitHub \ToxicsRedo\Estuary\_Analysis\Estuaries.gdb\stations\_subset

The data table for the 2010 stations was accessed using the *Pull\_2010\_StationUseList* R script located here:

[\\Deqhq1\mpsaris\GitHub\ToxicsRedo\Estuary\_Analysis](file:///\\Deqhq1\mpsaris\GitHub\ToxicsRedo\Estuary_Analysis)

The resulting csv file is joined up to the 2010 station use list shapefile for analysis.

*NOTE: In order for this join to be successful, there must be a properly formatted Schema.ini file. Refer to the comments starting on line 28 of the* ***Estuary\_Analysis\_01.p****y script for more details.*

If a new station is located downstream of any 2010 stations labeled estuary, the new station is classified as estuary. If there are no upstream stations classified as estuary, the station is marked for further review. The resulting shapefile is located here:

\\Deqhq1\mpsaris\GitHub \ToxicsRedo\Estuary\_Analysis\Estuaries.gdb\stations\_subset\_est2010

This script adds one new field called “Estuary\_2010”, and has the following values:

1. ‘Needs Further Review’
2. ‘Estuary’
3. ‘No stream LLID’

## Estuary\_Analysis\_02.py

This script computes maximum salinity concentrations for all new stations, and classifies stations accordingly. This results in the following two fields:

1. Sal\_ppth – Maximum salinity computed for this station
2. Est\_Sal – Classification based only on the maximum salinity concentration registered for each station. Possible values include:
   1. Freshwater – salinity below threshold.
   2. Estuary – Salinity above threshold
   3. Needs Further Review – No data available.

Once this data has been included, a new shapefile is created, and is located here:

\\Deqhq1\mpsaris\GitHub\ToxicsRedo\Estuary\_Analysis\Estuaries.gdb\stations\_subset\_est2010\_consal

### Manual Review

At this point, there is a break in the script, and a manual review is conducted. Final classifications are made in an edit session in ArcMap, based on the criteria laid out at the top of this section. This results in two final columns:

1. Est\_Final – the final classification for each station. This has the following possible values:
   1. Estuary
   2. Freshwater
   3. Marine
   4. NULL – Stations which were not finalized in the original Stations Analysis process were not included in the estuary classification, and therefore have NULL values. These stations fulfill the following criteria:
      1. QAQC2 = ‘Potential Digitization’ (None of the subseted stations were classified as ‘Remove’)
2. Est\_Comments
   1. This includes any necessary comments describing the rational for classifying a station a certain way.

### Final Product

Once the above manual review is conducted, the remainder of the script can be run. A copy of the master shapefile (*All\_stations\_final*) is made and the new fields are joined to it. All finalized stations which are not located in the selected HUC8 boundaries are then labeled “freshwater”. Stations which have not yet been finalized, ie “QAQC2” IN (‘Remove’, ‘Potential Digitization’), remain NULL. This final product is located here:

\\Deqhq1\mpsaris \GitHub\ToxicsRedo\Estuary\_Analysis\Estuaries.gdb\All\_stations\_final\_est

In addition to the twenty fields contained in *All\_stations\_final*, the final shapefile has the following fields:

1. Estuary\_2010
2. Sal\_ppth
3. Est\_Sal
4. Est\_Final
5. Est\_Comments