AUTOMATED ASSESSMENT: MANAGING LARGE DATA SETS WITH R

Jack Pflaumer, Biswarup Guha, Dave Rosenman

Division of Water Monitoring & Standards
New Jersey Department of Environmental Protection (NJDEP)

May 29, 2018
Overview

- Why Automation Needed?
- Manage Data
  - Data for Integrated Report
  - National Water Quality Portal (WQP)
  - Continuous Data Sets
- Why Use R?
  - New Jersey’s Evolution to R
  - Benefits of R
  - Challenges and Lessons Learned
Why Automation Needed - Data & Assessment

- Difficult to obtain data from multiple sources and formats
- Issues with managing large data sets
- How to QA large number of stations / parameters
- Need to eliminate many repetitive data analyses steps
- Focus on water quality evaluations by scientific staff
- Improve accuracy, more complete assessments
- Improve transparency and reporting tools
Integrated Report Process

- Chemistry Data Crunching
  - 958 assessment units (HUC14)
  - 5 - >10 years of data
  - > 10,000 discrete stations
  - > 3.2 million discrete data
  - > 300 continuous monitoring stations
  - > 90 parameters
  - Biological data

- Assessments
  - Station level
  - HUC level chemistry and biological assessment rollup
  - Designated Use assessment
2016 NJ Integrated Report Data Sources

- **State**: 41%
- **Utility**: 28%
- **Federal**: 26%
- **Academia**: 1%
- **Interstate**: 2%
- **Volunteer**: 1%
- **County**: <1%
- **Watershed**: <1%

* Federal data includes:
  Cooperative NJDEP/USGS
  River and Stream Monitoring Network
Data Repositories

- NWIS
- STORET (pre–WQX)
  - Clearinghouse for nationwide water quality data
  - Difficult to import data
  - Difficult to retrieve data
  - Uneven participation and data availability
- Myriad of other databases (excel, access, text)
Water Quality Portal

Benefits

- Since 2012: EPA (STORET), USGS (NWIS), USDA (STEWARDS)
- One Stop Shop for data (except for continuous)
- Standardized data protocol
- Easy to use interface with help desk
- Closer collaboration with stakeholders with WQX
- Publicly available and easily accessible
- NJ requires approved QAPP
Data Access & Assessment BEFORE the Water Quality Portal

MUCH EFFORT

USEPA STORET

Retrieve

Reformat & Normalize

Merge

USGS NWIS

Retrieve

Reformat & Normalize

Merge

Local Database

Assessment

Integrated Report & ATTAINS
Data Access & Assessment in AFTER the Water Quality Portal

- USEPA STORET
- USGS NWIS
- NWQMC WQP
- Local Database
- MOST EFFORT
- Integrated Report & ATTAINS

Export
Retrieve

NJDEP Water Monitoring and Standards
Continuous Data Solution

- National Issues with Continuous Monitoring Data
  - STORET not designed for this
  - No data standard
  - Difficulty sharing data
- NJ Solution – web accessible database (Rutgers U)
  - Interactive
  - Easy to use
  - Expandability
    - Other data generators
    - EPA is working with Rutgers
Continuous Data Solution
History of Automated Assessment Process

- Incremental improvements to automation process
- Excel (pre-2006)
- R (2016)
Why Manage Data with R

- Free software
- Easily handles large datasets, easy to use
- No special data management system needed
- Store data in ordinary ascii or binary files
  - Large datasets – binary format improved efficiency
- Automated downloads (WQP, USGS)
- Vast community and resources available
- Can do everything with R: data management, assessment, reporting
- Integrates with other languages and platforms
NJ IR Automated Process

Step 1. Data Integration
- Continuous
- Discrete
- Biological, Shellfish, Fish Tissue

Step 2. QA
- QAPPs
- Duplicates
- Data / Laboratory / Field Comments
- Validity

Step 3. Analysis
- Data Level
- Station Level
- HUC14 Level (Biological, Chemistry)
- Designated Use

Step 4. Integrated Report
- Content and Text
- Figures
- Tables

Step 5. ATTAIANS batch upload

NJDEP Water Monitoring and Standards
Reporting Tools

- Maps
- Tables
- Piecharts,
- Bar Charts
- Scatter Plots
- Box Plots
- GIS
- Etc....
R Automation Benefits

- Improved Efficiency
  - ~ 70% time savings for NJ
- More Time for Comprehensive Assessments
- Improved Accuracy
- Improved Validation
- More Information for Technical Decisions
- User Friendly Visualization Tools
- Overall Improved Transparency of Process

Challenges with automation:
- Programming Skills & Time
- Validation & QA: requires scientific staff verification
Lessons Learned

- Use the resources on the internet
- New features / packages / tools make a difference
- Train more than one person, get a computer programmer (non-environmentalist) or contract out
- It takes much longer time to program if doing it in-house
- Not a one-time deal: continuous updates needed
- IT Department approvals can be a long process
- Lookup tables – accuracy is critical
- Inconsistent data standards – continued improvement required
Questions

Point of Contacts:
Jack Pflaumer: (609) 633-0499  jack.pflaumer@dep.nj.gov
Roop Guha: (609) 292-1592  biswarup.guha@dep.nj.gov

New Jersey Department of Environmental Protection (NJDEP)
Division of Water Monitoring and Standards (WM&S)
Quality Assurance Tools

- Data without approved QAPPs
- Duplicate records
- Result Comments
- Quality Control Data
- Data Errors
- Site Conditions
- Lab Comments
- Parameters reported at same location and same time with different units
- Cont. Data including Max / Min
- Cont. Data in WQP download
- Check Entire Record

QA Tools
- Substitute Record
- Substitute Censored Data
- Normalize reporting unit and parameter names
- Flag preliminary and estimated Data

Data Errors

- slope
- intercept
- calibration
- matrix interferences
- lab comments
- site conditions

Result Comments

- precision
- accuracy
- sample collection
- sample preservation
- sample transport

Quality Control Data

- control limits
- quality control
- data reduction
- data validation

Cont. Data in WQP download

- data integration
- data analysis
- data interpretation

Parameters reported at same location and same time with different units

- unit conversion
- data reconciliation
- data comparison

Cont. Data including Max / Min

- maximum values
- minimum values
- range
- dispersion

Duplicate records

- data redundancy
- data overlap
- data verification

Check Entire Record

- data integrity
- data completeness
- data consistency
- data accuracy

Data without approved QAPPs

- data gap
- data inconsistency
- data error
- data deletion

Remove Data

- data deletion
- data correction
- data adjustment
- data exclusion

Substitute Record

- data substitution
- data replacement
- data interpolation
- data extrapolation

Substitute Censored Data

- censored data
- truncation
- censoring
- data suppression

Normalize reporting unit and parameter names

- unit conversion
- unit standardization
- unit normalization
- unit transformation

Flag preliminary and estimated Data

- preliminary data
- estimated data
- provisional data
- cautionary data

NJDEP Water Monitoring and Standards