USGS-R: A community to support and expand R scientific computing capacity

E. Read, A. Appling, L. Carr, L. DeCicco, J. Read, J. Walker, and L. Winslow
U.S. Geological Survey, Office of Water Information
New regional lake study published

Waste water effluent loads estimated for two streams

National flood forecasting network now active

Streamflow reanalysis dataset released

Local to regional-scale descriptive science

Broad-scale prediction

Data/information/tool divide
Challenges

• Collaborative interdisciplinary work
  – Exchange between hydrologists and limnologists
  – Engage software developers as equals

• Science in the era of web-enabled research
  – Web collaborations
  – Web tools
  – Data web services

• Shared tools improve efficiency
  – Open data -> open tools
  – Reusable solutions
Research workflows

- The Status Quo:
  - Confusing code
  - Poor software development practices
  - Not repeatable
  - "One-offs"
Research workflows

• Analytical building blocks:
  – Tested/versioned
  – Documented
  – Modular
  – Dependable
### rOpenSci packages

#### Data

<table>
<thead>
<tr>
<th>Package</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>spocc</td>
<td>Collect and map species occurrence data from GBIF, iNaturalist, EoEcoengt, AusWeb, elided, and USGS's BISON.</td>
</tr>
<tr>
<td>AntWeb</td>
<td>Access data from the world's largest ant database. Maintained and developed by the California Academy of Science.</td>
</tr>
<tr>
<td>ecoengine</td>
<td>Retrieve 3+ MM records, checklists and photos from Berkeley's ecoengine.</td>
</tr>
<tr>
<td>rebird</td>
<td>Search and acquire occurrence records of bird sightings collated in the ebird database.</td>
</tr>
<tr>
<td>rAvis</td>
<td>Download occurrence data for bird species, and associated metadata from ProyectoAvis, a citizen science bird project in Spain.</td>
</tr>
<tr>
<td>REForData</td>
<td>Connects to instances of web portals for collaborative data management (e.g. REForChina and RUFor).</td>
</tr>
<tr>
<td>rfisheries</td>
<td>Search and retrieve data from the OpenOceana.org, currently providing access to global capture fishing landings from the Food and Agriculture Organization (FAO) of the United Nations.</td>
</tr>
</tbody>
</table>

This package is now deprecated. Programmatic interface to the Metadata portal now available. [contribute](https://github.com/OpenScicore/package).
How do we make the leap?

- Promote and **teach** best practices
- Catalog of open-source tools
- Develop a **community** of learners and teachers
Promote and teach best practices

<table>
<thead>
<tr>
<th>Outreach</th>
<th>Instruction</th>
<th>Advising</th>
<th>Collaboration</th>
</tr>
</thead>
</table>

[Image of USGS logo, R logo, Twitter logo, images of people, and a graph.]

[USGS logo at the bottom left.]
Promote and teach best practices

Outreach
Promote and teach best practices

Intro R Workshop Outline

22 January, 2016

Workshop Schedule

Day 1
- 00 - Welcome (8:30 am - 9:00 am)
- 01 - Introduction (9:00 am - 10:00 am)
- 02 - Get (10:15 am - 12:00 pm)
- 03 - Clean (1:00 pm - 2:45 pm)
- 04 - Explore (3:00 pm - 4:30 pm)

Day 2
- 05 - Analyze: Base (8:30 am - 9:30 am)
- 06 - Analyze: Using Packages (9:30 am - 10:45 am)
- 07 - Visualize: Base (11:00 am - 12:00 pm)
- 08 - Visualize: ggplot2 (1:00 pm - 2:30 pm)
- 09 - Repeat (2:45 pm - 4:30 pm)

Day 3
- 10 - Practice (8:30 am - 10:30 am)

This workshop is designed to provide training and tools to perform common data analysis workflow steps:
get -> clean -> explore -> analyze -> visualize -> repeat

The workshop and materials are hands-on and include examples and exercises. After completing the workshop, you won’t be an R expert, but should have the foundation for getting started on your own data analysis work in R and will know where and how to get help.

We have borrowed from many sources for this material, most significantly from Jeffrey W. Hallister’s IntroR course. Thank you, Jeff, for openly sharing your materials. In addition, material has been drawn from Software Carpentry, Data Carpentry, and from R seminars given by numerous individuals at the USGS Center for Integrated Data Analytics and the USGS Wisconsin Water Science Center at the 2014-2015 R Lunch Bunch Data Crunch seminar series. We are grateful to all of these sources.

Each step of the workflow has a written component (blog post) describing the topic, a demonstration of example code, and hands-on exercises. The R code found in each blog post is also available as a stand-alone R script of the same name. The blog posts are intended to be used as a reference after the workshop.

Workshop Schedule

Day 1
[00 - Welcome (8:30 am - 9:00 am)]
Promote and teach best practices
Promote and teach best practices

Water Use in the United States

The U.S. Geological Survey’s National Water-Use Information Program is responsible for compiling and disseminating the nation’s water-use data. The USGS works in cooperation with local, State, and Federal environmental agencies to collect water-use information. USGS compiles these data to produce water-use information aggregated at the county, state, and national levels. Every five years, data at the county level are compiled into a national water-use data system and state-level data are published in a national circular. Over the history of these circulars, the water-use categories have had some changes. The USGS Water-Use Data and Research program seeks to develop improved water-use data through agreements with State water-resources agencies.

Estimated Use of Water in the United States in 2010 is available (published November 2014). Report | Data download | Fact sheet


Water Use Overviews

Total Water Use

Surface Water and Groundwater Use

Categories of water use

Public Supply  Domestic  Irrigation  Thermoelectric Power
Promote and teach best practices

Collaboration

John Wesley Powell Center for Analysis and Synthesis

Powell Center Working Group Project Information

Continental-scale overview of stream primary productivity, its links to water quality, and consequences for aquatic carbon biogeochemistry

Principal Investigator(s):
Edward Stets (USGS Branch of Regional Research, Central Region)
Emily Stanley (University of Wisconsin-Madison)
Jordan S Read (Center for Integrated Data Analytics (CIDA))
Robert Hall (University of Wyoming)

Participant(s):
Charles B Yackulic (Grand Canyon Monitoring and Research Field Station, SBSC)
David L Lorenz (Minnesota Water Science Center)
Judson W Harvey (Branch of Regional Research, Eastern Region)
Natalie Griffiths (Oak Ridge National Laboratory)
Alison Apling (University of Wisconsin)
Emily Bernhardt (Duke University)
Jim Heffernan (Duke University)
Maite Arroita (University of the Basque Country)
How do we make the leap?

Get data

Present results
Get data → new → Present results

Catalog of tools

rOpenSci packages

Data packages that interface with data repositories

<table>
<thead>
<tr>
<th>Package</th>
<th>Description</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>spocc</td>
<td>Collect and map species occurrence data from GBIF, iDigBio, iCoEing, AusWeb, dblil, and USGS’s BIOSIS.</td>
<td></td>
</tr>
<tr>
<td>AntWeb</td>
<td>Access data from the world’s largest ant database. Maintained and developed by the California Academy of Science.</td>
<td></td>
</tr>
<tr>
<td>ecoengine</td>
<td>Retrieve 3: MM records, checklists, and photos from Berkeley’s ecoengine.</td>
<td></td>
</tr>
<tr>
<td>rebird</td>
<td>Search and acquire occurrence records of bird sightings collated in the citr database. Download occurrence data for bird species, and associated metadata from.</td>
<td></td>
</tr>
</tbody>
</table>

Catalog of tools

- USGS-R
  - Collection of USGS R codes and packages
  - dataRetrieval
    - This R package is designed to obtain USGS or EPA, as well as user-supplied text. It retrieves and organizes metadata from a variety of sources, including the USGS, EPA, and user-supplied text files.

- sensorQC
  - High-frequency sensor GQG procedures. sensorQC imports data, and runs various statistical outlier detection techniques as specified by the user.

- tom
  - Timeseries Observation Model

- glmtools
  - Tools for Interacting with the Generalized Linear Model (GLM) in R

- smwrBase
  - Base USGS water science R functions.
<table>
<thead>
<tr>
<th>packagename</th>
<th>dl_count</th>
</tr>
</thead>
<tbody>
<tr>
<td>dataRetrieval</td>
<td>1376</td>
</tr>
<tr>
<td>smwrGraphs</td>
<td>1223</td>
</tr>
<tr>
<td>smwrBase</td>
<td>1209</td>
</tr>
<tr>
<td>smwrQW</td>
<td>1054</td>
</tr>
<tr>
<td>smwrStats</td>
<td>1002</td>
</tr>
<tr>
<td>smwrData</td>
<td>934</td>
</tr>
<tr>
<td>WQReview</td>
<td>478</td>
</tr>
<tr>
<td>rloadeast</td>
<td>392</td>
</tr>
<tr>
<td>glmtools</td>
<td>386</td>
</tr>
<tr>
<td>EGRET</td>
<td>345</td>
</tr>
<tr>
<td>sbtools</td>
<td>328</td>
</tr>
<tr>
<td>GLMr</td>
<td>321</td>
</tr>
<tr>
<td>EflowStats</td>
<td>294</td>
</tr>
<tr>
<td>united</td>
<td>249</td>
</tr>
<tr>
<td>geoknife</td>
<td>240</td>
</tr>
</tbody>
</table>

[Image showing a webpage with a table and a screenshot of a package listing with the USGS logo.]
Catalog of tools

*Get data*

*Present results*
Can we better incentivize tool code sharing?
How to ensure maintenance and community support?
How do we combat tool misuse?
How can we know about everything that exists?
How do we make the leap?

Get data → new → Present results
Web-enabled analysis

California drought, visualized with open data

Scroll down to learn more.

The State of California is experiencing one of the most severe droughts on record, which has implications for citizens of California and beyond. Many State, Federal, and Tribal agencies make routine observations of the water cycle. This website graphically visualizes these data to help understand the effect of drought on rivers, streams, lakes, and reservoirs. Below, learn more about water use and changes in the water cycle due to the current drought in California.

The data presented here are drawn from free and publicly accessible sources. In addition, the analytical, graphical, and software tools used are open-source and available for public re-use.
“dataRetrieval is designed to obtain its water quality sample data, streamflow data, and metadata directly from either the USGS or EPA, as well as user-supplied text files.”
Web-enabled analysis

NWIS

sensorQC

XML

new

cida.usgs.gov/ca_drought

U.S. Drought Monitor

December 16, 2014

U.S. Drought Monitor

December 16, 2014
Community-based ‘help’ platform

NWIS → sensorQC → XML → new

U.S. Drought Monitor

cida.usgs.gov/ca_drought
Community-based ‘help’ platform
Community-based ‘help’ platform

GitHub repository: USGS-R / dataRetrieval

Issue: readWQPdata - fewer results returned than expected

jjwill2 opened this issue 18 days ago - 4 comments

jjwill2 commented 18 days ago

I used the code below to try to query all nutrient chemistry data from lakes in OR. There was a warning message: "In importWQP(urlCall, FALSE, tz = t2): 27571 sample results were expected, 1611 were returned". Is there any way to address or troubleshoot these warnings? Doing the same query through the WQ portal web interface yielded 27,374 results. What is happening here?

```r
OR_lake_data <- readWQPdata(
  # specify states
  statecode = "US:41",
  # specifies water samples
  ActivityMediaName = "water",
  # specifies lakes
  siteType = "Lake, Reservoir, Impoundment",
  # specifies parameter category
  characteristicType = "Nutrient")
```

https://github.com/USGS-R/dataRetrieval/issues/203
Community-based ‘help’ platform

Allow more times zones when retrieving data from webservices please. #152

I have an enhancement request, when retrieving data from waterservices using the readNWISuv command. All of the timeseries data collected in MI is stored as EST, so when we retrieve through dataRetrieval and use "America/New_York" (the only eastern timezone allowed) as the timezone all data retrieved during daylight savings time is 1 hour off. This poses a problem when we pull a year of data; some of the data is correct but some of the data is an hour off, so we have to do some data manipulation to adjust all that. If EST were an allowed timezone it would save us some time from having to reformat the times when we retrieve the data. Anyhow thank you for considering this enhancement request.

for example:

test <- readNWISuv('415318085243401', startDate="2015-06-01", endDate="2015-08-31", parameterCd = 'head(test,1)

agency_cd site_no dateTime tz_cd X_72019_00011 X_72019_00011_cd
1 USGS 415318085243401 2015-06-01 01:00:00 America/New_York -0.28 P
2 USGS 415318085243401 2015-06-01 01:15:00 America/New_York -0.27 P
Community of learners and teachers

Outreach  Instruction  Advising  Collaboration

[USGS logo]  [Twitter logo]  [Graph]  [Images of people]
Data/information/tool divide

Local to regional-scale descriptive science

Broad-scale prediction
In summary

• Building blocks for analysis can expand scope and increase efficiency of science efforts
• A learning community and clear access to tools is important

eread@usgs.gov