Scientific Computing & Analysis

Scientific Computing Topics

- Software Carpentry
- The Joel Test
- Backlash on scientific computing
- Floating point numbers
    - Definition
    - What Every Computer Scientist Should Know About Floating-Point Arithmetic
    - Floating point number overview
- Data access tools
    - Environmental Data Connector (ASA, Inc.)
    - OpenDAP

Scientific Computing Environments

Statistical environments

S-Plus

USGS holds a license for a version of the S-Plus statistical package, and the USGS internal distribution includes USGS-developed statistical and graphics tools.

- TIBCO Spotfire S+ 8.1 for Windows with the USGS Library 4.0 Public Release
- USGS support email: GS-W Help SPLUS
- USGS support searchable archive: http://smtp.water.usgs.gov/splus_help-archive/2011/

R

R is an open-source statistical analysis system built to be functionally equivalent to S. It is gaining in popularity, and has a body of GUI's (such as RCommander) and interfaces available for it. Since R has a command-line interface, it is fairly easy to connect with other software, for example, ArcGIS. and Java Python ("Jython"). Some have described R as a "statistical scripting language."

Although the user community is very good at answering questions, the volume of questions and answers may be overwhelming.

- The R Wiki http://rwiki.sciviews.org/doku.php
- There are VERY active mailing lists http://www.r-project.org/

Python-Pandas

Python-Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. While it is not quite as a rich as R in terms of statistics, it's really getting there. One major benefit is that its data structures are python-native. This means that if you exploit a lot of the python specific functionality to clean up your data, you will not have to transform it for use in R. In some cases, this ca be a substantial bottleneck. There's a pretty decent tutorial video (~30 minutes) (sound cuts out 20-23mins), a class video (~3 hours), and a "tour" video (~10 minutes), among many others.

Python

Python, an interpreted, interactive, object-oriented, extensible programming language, supported on numerous computing platforms. In addition to being relatively easy to use and good for numerical analysis and web programming, it is the de facto scripting language for our corporate GIS platform, ArcGIS.

- The starting point for all things Python is: http://python.org
- There are a large number of instructional videos available, follow this link for an example. Just google "python tutorial" for many great options.
- Python is widely used outside the USGS, so there are lots of great resources hosted in the wide world. We should really participate there, instead of building our own microverse. See the Python Forum, and http://stackoverflow.com/tags (type in "python").

Packages for Python

Generally useful stuff

- NumPy is a *really important* add-on for Python. It provide low-level functions for handling arrays. It comes with ArcGIS
- SciPy is another *really important*add-on for Python. It's built on top of NumPy and provides higher-level (i.e., more user-friendly) functionality. It also comes with ArcGIS.
    - SciPy has a handy Cookbook of worked examples of common tasks
- Python is distributed with a large standard library of modules that support various tasks, but many more are available online. An extensive collection of pre-compiled libraries are available in this collection posted by Christoph Gohlke. Key libraries of interest to scientific computing include NumPy, SciPy, matplotlib, and netCDF4.
- Versions of the GDAL and OGR libraries are now available in Python, in a package called pypi.
- Using Python with Fortran or C sub-page

In addition to a truly dizzying number of individual add-on libraries for Python, there are a few distributions of sets of python libraries that try to eliminate the hassle of pulling together lots of libraries. We should investigate these!

- Enthought Python Distribution (EPD): These guys are really interesting. In addition to grabbing and bundling together a whole bunch of python add-ons, they are apparently responsible for maintaining both the numpy and scipy packages.
    - Rich Signell has been trying to persuade ESRI to not only take a snapshot of EPD with each ArcGIS release, but to make the ArcGIS platform more flexible about using other versions of EPD. One idea is to encourage ESRI to work directly with EPD. Sounds like ESRI is open to this, but that a stronger/clearer business case needs to be made. See this sub-page for a discussion of the potential benefits of EPD with ArcGIS.
- UC-Irvine (unofficial) Windows-compiled libraries for Python - lots of hard-core computing options as individual libraries. Also offers a version of EPD.

More advanced stuff

- **Integrated Development Environments (IDEs)** - There are lots of choices, many of which allow a user to write in more than one language. See the sub-page on this.
    - iPython Notebook- kind of an IDE, but a great web-based user-interface for mixing markup, code snippets, and results in a unified presentation. This is a fantastic way to teach! Especially because the code snippets are live, meaning users can adjust and rerun them to update the results.
        - Rich Signell's colleague Massimo (last name?) has been working with this, configuring server images that he can virtualize. He's got one that has python, R statistics, and GRASS. Very cool!
        - Would be interested to see if this could be made to work in conjunction with ArcGIS (probably the Server product) so folks can explore with the ESRI arcpy object for python.
        - Would also be very interested to see, if this server image matures, we can encourage folks developing ArcGIS Server images for cloud deployment to include these add-ons/work with us.
        - This also makes one wonder about web-accessible/executable python-based functionality. iPython Notebook uses 0MQ for messaging, but this almost doesn't matter. Might be best to focus on using OGC WPS instead.
- NetworkX - a pretty hard-core library for making complex networks and graphs. Could also be useful for TINs.

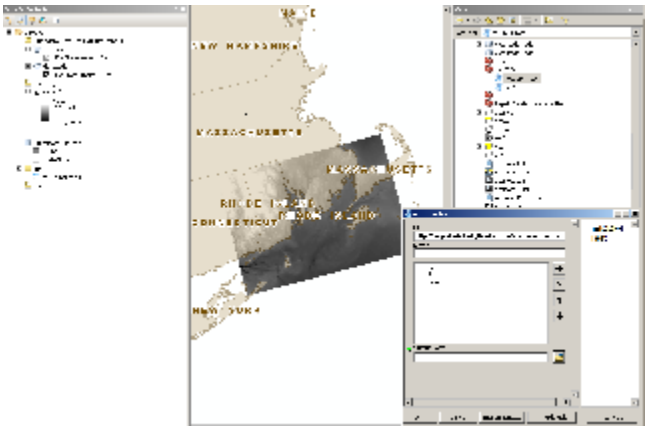
Python and ArcGIS

- Python, aside from being a standalone general scripting language, has become the main scripting language for the ArcGIS platform. Versions of Python 2.x and Python libraries that are included in different versions of ArcGIS are as follows:

ArcGIS	*9.3*	*10.0*	*10.1*
Python	2.5.1	2.6.2	2.7.2
NumPy	1.0.3	1.3	1.6.1
matplotlib	–	–	1.1.0

- The Enthought Python Distribution (EPD) version 7.3 is fully compatible with ArcGIS 10.1. This means you can install the EPD, then install ArcGIS and Arc will see the EPD distribution and simply add the arcpy modules without installing another python instance. Since the EPD 7.3 comes with an OPeNDAP-enabled NetCDF4 Python, there is no need to run Gohlke's installer (described below).
- The netCDF4 module compiled for ArcGIS 10.0, 10.1 allows fairly straightforward access of netCDF and OPeNDAP data from ArcGIS Python scripts. Thanks to Rich Signell and, most of all, Christoph Gohlke (who compiled the module so it will work with Arc). Rich and Curtis Price provided this python script and script tools (zipfile).

- The image below shows an example of a raster that has been loaded into ArcMap from a remote dataset using a Python script tool that accesses data using the netCDF4 library and the OPeNDAP access protocol. (Click it for a full-resolution view.)



Discussion topics

- A discussion of ways to write data objects to files

MATLAB

MATLAB is commonly used for data and compute-intensive scientific analysis.

Known USGS MATLAB users: Rich Signell, Ashley Van Beusekom

Microsoft Office

Although Microsoft Office is very useful for general-purpose computing widely used in science, it has also been also widely criticized by the scientific community (especially by statisticians). The largest problem by far is data import/export, and the misuse of the tools, for example the (far too common) use of Excel as a database, and errors in worksheet cell references.

USGS holds a site license for MS Office, through the Bureau Windows Technical Support Team (BWTST).

Geographic Information Systems (GIS)

Since much of USGS scientific computing involves spatial data, it is no suprise that more than half of the attendees of the 2011 CDI meeting were polled identified themselves as users of Esri's ArcGIS product.

USGS Core Science Systems supports the Enterprise GIS (EGIS) team, who supports GIS activities in the Bureau. EGIS supports USGS-wide site licenses for Esri's ArcGIS suite, and Global Mapper.

- ArcGIS Toolbox for NAWQA - presentation by Curtis Price

Contributors

User	Edits	Comments	Labels
Viger, Roland	49	0	5
Signell, Richard P.	6	0	0
Benson, Abigail L.	1	0	0
Blodgett, David L.	1	0	0