Developing a Soil/Terrain Geospatial Database to Support Soil Carbon Modeling

Edward Flathers[1], Paul Gessler[1], Richard Rupp[2], Erich Seamon[1]
[1]University of Idaho College of Natural Resources
[2]Washington State University College of Agricultural, Human, And Natural Resource Sciences
flathers@uidaho.edu

<Open Science>

As open-access data policies become more common among government agencies, publicly-funded research programs, and others, scientists are gaining access to vast collections of potential research inputs. The free availability of science data today enables the creation of science "mashups"—the combination of data from a variety of sources to serve new analyses and generate new data products.

The concurrent rise of the "open science" movement is a natural complement. Advocates of open science argue that scientists should go beyond the publication of traditional papers describing research methods and results; we should also publish collected data, computer algorithms used in analysis, and resulting outputs. By exposing these aspects of the research process, we enable reproducibility, enhance reusability, and expand the opportunities for peer review of our work.

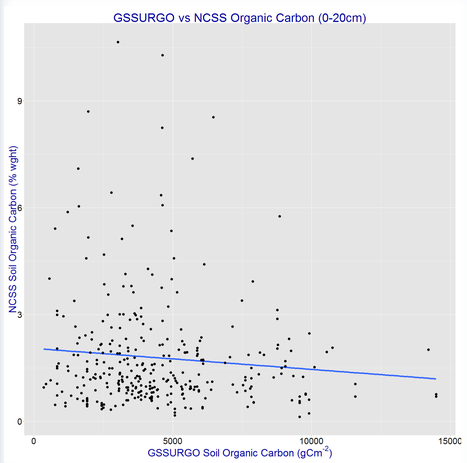
It is our goal in the course of this project to apply the principles of open science. During the project, we will collate data from REACCH researchers, the Natural Resources Conservation Service, National Cooperative Soil Survey, the USGS National Elevation Dataset, and others. Each stage of the project will be documented and all data, computer code, and output products will be made freely available. Our goal is to release the final product as a complete package of data and analytical software that can be downloaded and used to reproduce our results, apply our methods to different input data, or replace our analytical methods with more advanced methods from the soils literature.

Papazoglou, MP, and WJ Van Den Heuvel. 2006. "Service-Oriented Design and Development Methodology." International Journal of Web Engineering and Technology, 1–17.

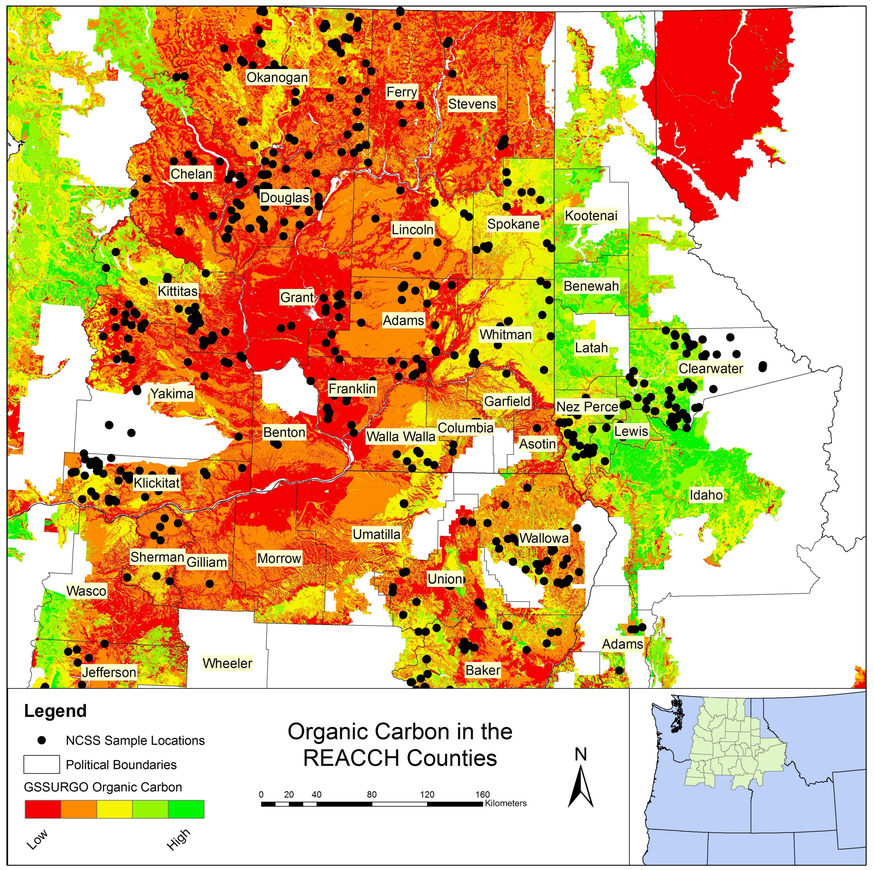
<Challenges of Open Science>

One challenge of participating in open science is in consuming data that are produced by other researchers. In order to be able to apply data in our own studies, we must know something about how the data were collected and recorded, among other details. We rely on accurate metadata records to understand the content and context of data that we are repurposing from another source.

As an example, consider the graph of GSSURGO vs NCSS Organic Carbon (right). This is a scatterplot of soil organic carbon content at roughly 600 sample locations across the REACCH region, according to the GSSURGO database (x-axis) and the NCSS soil samples (y-axis). Since both datasets represent the same physical attribute, we might expect to see a strong correlation between them. In this case, we see a weak, negative correlation (-0.09). The problem is a units mismatch, and is an extremely common obstacle encountered while attempting to collate heterogeneous data.



<GSSURGO Soil Organic Carbon>



GSSURGO data from Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Available online at http://websoilsurvey.nrcs.usda.gov/. Accessed February 2015. NCSS Samples from National Cooperative Soil Characterization Database Available online at http://ncsslabdatamart.sc.egov.usda.gov/. Accessed February 2015. Boundaries from US Census Bureau.

<GSSURGO at Large Scale>

The US Department of Agriculture Natural Resources Conservation Service maintains a soils information geodatabase called SSURGO—Soil Survey Geographic Database. This database geographically represents the United States as a collection of vector-polygon map unit areas defined by their soil characteristics. In 2011, the NRCS began publishing the GSSURGO database, which is a gridded (raster) dataset that is derived from the vector SSURGO data. The raster is composed of 10-meter cells that are snapped to the same grid as the United States Geological Survey's land cover products, which makes GSSURGO convenient to overlay with USGS products and with others that use the same grid system.

GSSURGO is a rasterization of the original SSURGO polygons, which means that each polygonal map unit from SSURGO is represented as a collection of 10x10 meter cells in GSSURGO. The attribute information for each grid cell is taken from the attribute of its parent map unit, which means that all cells within the map unit carry identical attribute values. This means that while the 10-meter grid appears to afford a higher spatial resolution than the original polygons, the effective scale for analysis remains at the map unit level. This scale is appropriate for many regional- and smaller-scale analyses, but may not be adequate for localized studies.

Value can be added to the GSSURGO database by producing accompanying data products that represent attributes at the 10-meter grid cell scale. Clearly, it is impractical to produce an empirical soil database built upon sampling each grid cell on the ground. We propose using spatial statistical models to build interpolated 10-meter layers that are compatible with the GSSURGO product.

In keeping with the goals of the REACCH project, we will focus on soil organic carbon (SOC) as our soil characteristic of interest. Soil carbon is primarily associated with soil organic matter and is a proxy for many soil properties related to resiliency and soil health for agriculture. We will gather terrain and soils information and apply analytical methods following those previously developed (Gessler et al. 1995; 2000) to develop raster maps of SOC across the REACCH region.

Gridded Soil Survey Geographic (GSSURGO) Database User Guide, United Stated Department of Agriculture. Available online at http://www.nrcs.usda.gov/wps/PA_NRCSConsumption/download?cid=nrcs142p2_051847&ext=pdf. Accessed February 2015.

<Methods>

The goal of this project is to demonstrate open science methods by building a software framework that allows the integration of heterogeneous data from a variety sources. The project framework includes three main elements:

- Extract, Transform, Load (ETL)
  - A series of small programs that download data from remote sources and homogenizes and organizes them into compatible formats to be combined and analyzed
- Analytical Code
  - A computer program that processes the homogenized data using a spatial statistical model to produce the output products
- Documentation
  - Standards-based metadata that describes data and analytical methods to aid in review and reuse

<Acknowledgments>

Funded in part through Award #2011-68002-30191 from the USDA National Institute for Food and Agriculture, and with support from these organizations: