Using big data methods in cartography and modeling

Edward Flathers (flathers@uidaho.edu) UI, Paul Gessler UI, Erich Seamon UI, and Rick Rupp WSU

The rise of "big data" science in recent years has been of great commercial importance to major businesses such as Facebook, which applies powerful data analysis techniques to choose which advertisements to show to each user, and Netflix, which uses data about its customers' rental history to recommend movies that they might enjoy. In the REACCH project, we can use some of the same big data methods to develop a deeper understanding of our environment and agricultural systems in the Pacific Northwest. Here we describe a project in which we are using cloud services and supercomputers in combination with data collected by many different researchers to develop a system that enables us to map out the organic carbon content of the soil across our region, giving us some indication of soil health (Figure 1).

IMPACT

Developing "big data" methods for data analysis helps us to answer current questions about the REACCH study area and serves as a platform for continued study in the future.

Big data is an emerging field that describes new kinds of scientific analysis that have been enabled by recent advances in technology. Inexpensive data storage, broadband Internet connectivity, and computer processor capability come together to allow us to build larger collections of data and to transmit those collections to high-performance computer centers for analysis. This improved technology comes into play, for example, in weather forecasting—short-term weather forecasts today are much better than in the past, in part because of the powerful supercomputers that are used to model weather systems.

In the REACCH project, we are developing methods for applying these big data technologies and techniques to environmental data in ways that are easy to repeat, reuse, and repurpose for use with data that we will collect in the future. As a pilot study to drive the development of our big data tools, we are using data that describe the soils and topography of the REACCH study area to build a statistical model that produces a map of soil organic carbon in our agricultural areas. The organic carbon content of soil can give us some idea of the health of the soil and help guide decisions about the agricultural management practices that we employ in an area (Figure 2). If modeling efforts can produce data at a high enough resolution, the results could even be used to support activities like precision agriculture. Our big data process has four main components: data collection, processing, visualization, and storage.

First, we collect the data. One of the hallmarks of big data science is bringing together data from a variety of sources and assembling them into a single, large collection. In our case, REACCH researchers have collected soil from various locations in the field, taken the samples back to the lab, and measured the soil's organic carbon content. Researchers at the U.S. Department of Agriculture National Cooperative Soil Survey have made similar observations across the country and have made their data available to the public on their website. We combine these two data sources to build a dataset that has more complete coverage than either one of the original sources has on its own. We also include topographic data from the U.S. Geological Service National Elevation Dataset.

Next, we process the data. Another common practice in big data science is the use of "cloud" processing: offloading complex computations to a massive supercomputer that is shared by many clients. At the REACCH project, we can choose to process our data using the powerful supercomputer located at the Idaho National Laboratory, or we can use the Amazon Elastic Compute Cloud, among others. Our choice of a processor is influenced by how complex our model is, how busy each cloud processor is,

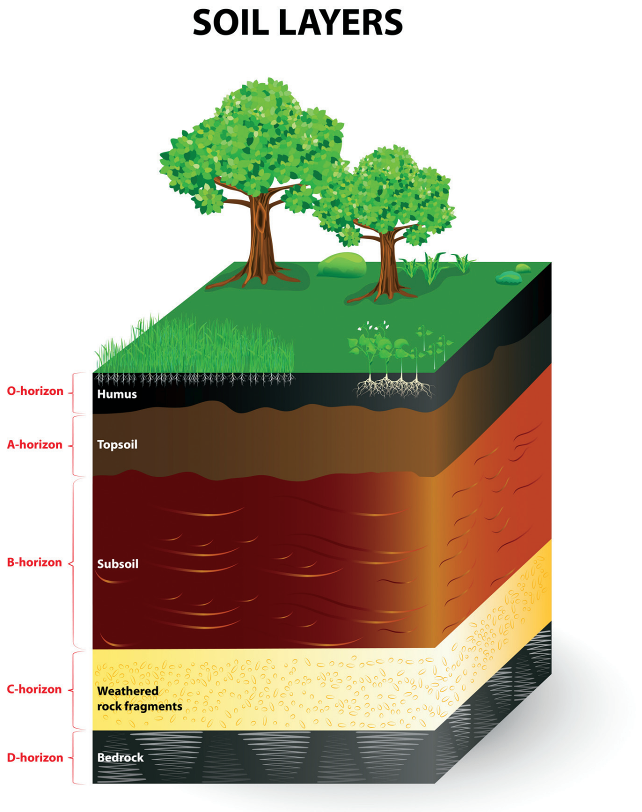

Figure 1. *Soil organic carbon is usually concentrated near the surface, where it accumulates as a product of the decay of plant matter. Image © Designua | Dreamstime.com*

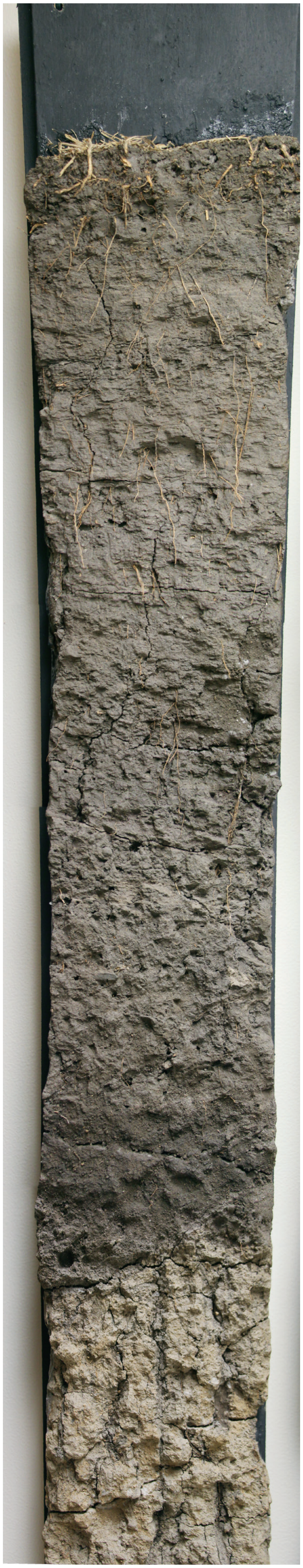

Figure 2. *Soil profile from the wheat-growing area of the Pacific Northwest. Photo courtesy of UI PSES.*

how quickly we need the results, and the cost of computer time. We upload the data to the supercomputing facility for analysis, and then sit back and wait for the results to come back.

When the supercomputer is done, we get back to work. The results of our statistical model run are a large numerical data table. We import this data table into software programs that allow us to build a map of our area of interest (Figure 3). The map, in combination with the data table, can be used by crop consultants and growers to better understand the way that soil organic carbon content varies over our agricultural area, which can support them in making management decisions.

The last step in our big data process is storage. We take the input dataset that we built, the statistical model that we executed, the tabular results that we received from the supercomputer, and the map that we created, and we package it all up for storage in our long-term data library. By archiving the data and the computer code that we used to produce our results, we can ensure that we can always go back and repeat the process, perhaps using additional soil samples that have been collected, or for a different area of the country. We can also share our process with researchers at other institutions, who can help to refine the methods using their own expertise.

This soil-mapping exercise is just one example of the kinds of big data science that can be done in regional projects like REACCH. As we develop our modeling process, we prioritize the use of free, industry-standard software and methods that help us implement a system that is modular and reusable, and that can provide benefits not only to the stakeholders of REACCH, but also to other projects in the future.

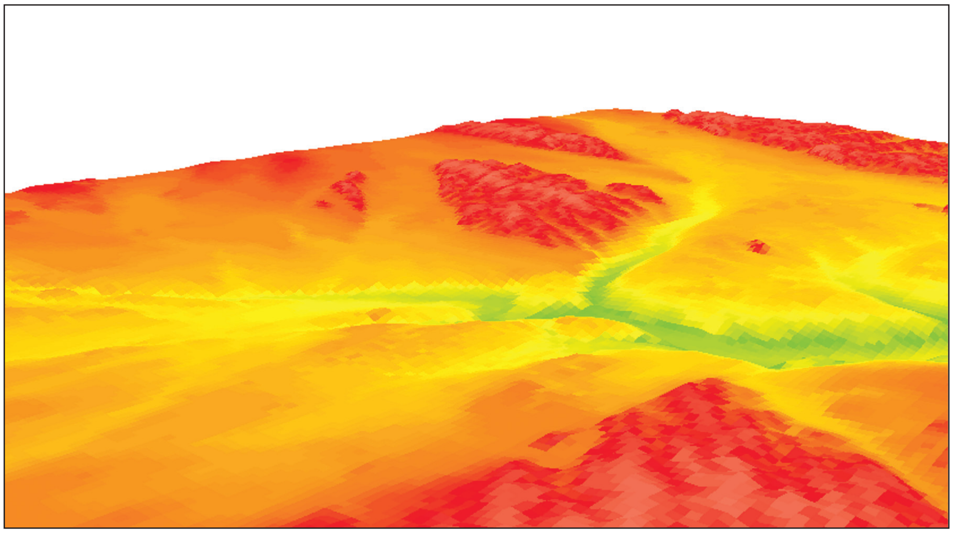

Figure 3. *A 3-D map showing approximate soil carbon concentration in an agricultural area near Umatilla, OR. Carbon migrates from erosional areas (red) to depositional areas (green) due to differences in soil composition and terrain.*