COAL: Coal and Open-pit surface mining impacts on American Lands

Taylor Alexander Brown, Heidi Ann Clayton, and Xiaomei Wang

Group 18

CS 461: Senior Capstone Fall 2016

Oregon State University

Abstract

Mining is known to cause environmental degradation, but software tools to identify mining impacts are lacking. Researchers studying this problem possess large imaging spectroscopy and environmental quality data sets as well as high-performance cloud-computing resources. This project provides a suite of algorithms using these data and resources to identify signatures of mining and correlate them with environmental impacts over time.

CONTENTS

1 Problem Definition

The problem that COAL seeks to solve is to develop original software for identifying mining impacts using imaging spectroscopy data and geographic information. Large data sets of imagery and geographic information are available to data scientists and high-performance cloud-computing resources exist to process this data. Currently there is an abundance of literature describing projects with specific use cases for processing these resources, however the software these projects have developed is generally specialized and not available to the general public. COAL will use insights gained from prior research to provide a reusable framework for processing the data that can be generalized from our case study of environmental impacts of mining to a broad range of applications.

2 Proposed Solution

2.1 Capabilities

Our solution is to develop a Free and Open Source suite of imagery processing algorithms. They will be implemented as library functions which can be incorporated into application programs with a variety of goals. The algorithms we are specifically interested in implementing will identify minerals based on their spectroscopic signatures, use mineral distributions to locate mining operations, correlate geographic information about environmental conditions with mining operations, and derive conclusions about mining and its effects on the environment over time. The algorithms will be verified by automated tests which compare expected output to actual values. Because source code rarely exists in a vacuum, our library will also require a detailed API reference for end users and a website for the general public.

2.2 Requirements

These features provide solutions to each part of the problem described. The software we develop will fill a niche application to environmental data science that is as yet unfilled. Because the library will be licensed under a permissive Free and Open Source Software license it will be available to the widest possible audience, including but not limited to federal agencies, non-governmental organizations, and private enterprises. Because it will be developed in a popular programming language, it will be accessible to developers with a broad range of backgrounds.

The methods we develop to process spectroscopic imagery and geographic information will solve the problem of processing these datasets. They will allow users to derive meaningful conclusions from the data by generating representations and visualizations to illustrate relationships that might not otherwise be inferred. Users with access to high-performance computing systems will be able to take advantage of efficient implementations of scientific and numerical operations to process large data sets.

The documentation accompanying our project will provide both a detailed and a high-level view of our software and its usage. The API reference will allow end users to understand the structure and function of the library in order to apply it to their own application. The website and wiki will provide a more general overview to communicate with the public and share the results of our efforts.

2.3 Presentation

We expect to present a number of facets of our software and development team. We will first introduce ourselves, our client, and our client's affiliation with the Jet Propulsion Laboratory. Describing the stake which each team member holds

in the project and the importance of the project to the scientific community will provide context about the motivations of the software.

The technical basis of the project will be described to provide background about the sources of the data. A brief description or depiction of imaging spectroscopy will illustrate the sensor which records the imagery. The data structures used to store and manipulate the imagery must then be described in order to provide a sense of scale about the amount of data to be analyzed.

The algorithms and techniques used to implement the library will then be presented in a simplified fashion. We will discuss how minerals were identified based on their spectrographic signatures, how mines were located based on mineral distributions, how environmental conditions were associated with mining operations using geographical information, and how temporal analyses were performed over historical datasets.

Examples of input and output data will demonstrate the use of our library as well as the value it provides to the scientific community. The case study we examine as we implement the project will provide a source of specific imagery and geographic information to depict. The results of analyzing this data will feature prominently to show the kinds of results we were able to derive. Bitmap images, hyperspectral cubes, GIS layers, tables, graphs, and maps are all candidates for inclusion into the presentation.

As we develop the project we will be noting our progress and our results in order to present a research paper. The paper may not be finalized by the time of our presentation, but we should be able to share the highlights of our draft which will provide a scientific foundation to our work.

3 Performance Metrics

The most basic metric for our software will be that each component satisfies its function. The mineral identification module should, when given a data set and a signature to search for, return a distribution that corresponds to results that are known from observation. Likewise, the mining identification step should be able to locate known mines based on their surface features. The correlations we derive between mining and environmental conditions may not have been previously published, so the basic metric is that the predictions should be testable by observation in order to confirm or disconfirm the correctness of our algorithm. Likewise the historical analysis may be novel, but its results should be able to be analyzed by scientists to provide new insights and directions for study.

More concrete metrics are imposed by performance constraints. Our algorithms should be efficient enough to process realistic data sets within practical time and space constraints of the time sharing systems we use to execute them. Specifically, we should be able to complete our analyses without exceeding the allotment of supercomputer time available to us. These requirements will require careful development to develop sufficiently efficient algorithms.

Automated tests will provide an additional metric for measuring the outcome of our project. Tests should exist to exercise multiple features of program behavior. Because tests serve as an informal specification and verification of correctness, there should be no failing tests in the finished application.

Because we seek to publish our results, our implementation and its results should be conducted in a reproducible manner. The source code will exist in a version control system that will maintain a history of the project development. The results we derive should be accompanied by detailed information on how they were produced. This way, not only will we be ready to present a research paper, but our results will allow others to learn from our successes and mistakes and advance scientific understanding.