Machine Learning Applications for International Ocean Discovery Program Geoscience Research

Peter Flaming[1], Brandon De La Houssaye[1], Quinton Nixon[1], and Gary Acton[2]

[1] Master of Science in Data Science, Southern Methodist University, Dallas TX 75275 USA {pflaming,bdelahoussaye,qnixon}@smu.edu
[2] International Ocean Discovery Program, Texas A&M University, College Station, Texas 77845 USA acton@iodp.tamu.edu

Abstract. In this paper, we present the use of a machine learning in order analyze high-resolution still images gathered from decades of scientific ocean drilling data. The International Ocean Discovery Program (IODP)[1] which is part of the National Science Foundation (NSF) has been overseeing the JOIDES Resolution scientific drillship as it travels the oceans and takes samples of the earth and rock beneath the ocean floor. While the images have been in existence for some time, the means to extract the colors in the image into a usable data file for scientists has thus far eluded the IODP and NSF. These sample are what we apply computer vision techniques to in order to classify the colors (based on location in the sample) and then flatten the results into a data format. Or, put another way, in this paper we present a means of using machine learning in order to transform the color from the sample into useful data so that it may be analyzed by scientists in order to study the history of the earth. A computer vision model was utilized in order to first develop a Deep Neural Network (DNN) to mine and extract important features from the sample photos, and a Recursive Neural Network (RNN) was then utilized against the preprocessed DNN data. [NOTE: additional modes & techniques are being explored such as variants of CNN.] The overall goal of the model is to take an image and transform it into a flat data file indicating the shades of color (on a grey scale) against the location in the sample. The conclusion presented is the model built and the indicators of accuracy.

Keywords: IODP · JOIDES Resolution · Geoscience · SST · CCD · Machine Learning · Deep Learning · DNN · RNN · CNN.

1 Introduction

There is significant work undertaken by a broad array of scientists, industrialists, and officials really, all of humanity seeking to understand our Earth. A particular hot topic today is the Earths climate. Geologists have the ability to explain the present and the past of the Earth by analyzing the dirt, rocks, etc.

[1] More information may be found at http://iodp.tamu.edu. Last accessed 14 Jul 2019.

right under foot. However, as a can be imagined, the sheer volume of information that comes from the samples needs to be a format where the geologists can perform their scientific studies.

As it relates to this paper, of particular interest are samples taken the beneath ocean floor. Historically, the data collected from the samples were able to be integrated parametrically in order to describe the present climate. However, with new technologies capturing data in near real-time and in high resolution it presents nonparametric relationships that make it hard for users to model the past and future of the Earths climate with traditional statistical models. Lacking in these traditional methods is the ability to expand these relationships nonparametrically to other locations, such as onshore basins, offshore basins, different wells, and the emergence of fracking. Thus, as the scientific community has expanded their ability to gather information in real time and with better tools (such as high-resolution photography) a new problem has emerged around how to transform this gathered information into more computationally powered modeling techniques.

As more cores are gathered, and the scientific community seeks to gain better insights into the whole of what the core can tell us by using better information gathering tools, there is a real need to transform this additional information into a format where scientific understanding can occur. The above summary is so broad encompassing and the problem statement is so high level, that it may feel of hubris. While the presenters of this paper maintain awareness of such, they also do not limit the potential contributions contained herein towards those broader goals. Nevertheless, in a practical sense, this paper can be described as containing a proof of concept computer vision model that focuses on transforming one element of the information to be gleaned from the sample (the high-resolution photographs) into a data-driven usable format where existing (or soon to be developed) statistical models can be applied in order to better understand the Earths history.

In building the model, the key focus is one of scalability. Scientists can and have gone through the sample photos and translated the colors into a flat (data) file based on color spectrum using spectrophotometry. Basically, this means that for the core sample, at a given location in the photo, on a grey scale from 0 to 255, what is the color shade? The results are then mapped and plotted, and the resulting data can be statistically analyzed. So, the question at hand is why would a team of data scientists be needed? The answer is that it has been done by hand over a painstakingly long and expensive process. What data science can do is develop a model that for a given image, the resulting classification of the image (against a greyscale) becomes automated. The model, once it has been trained to achieve satisfactory accuracy scores, is scalable. That is, the model can be employed in real time to additional core photographs. The end result is a data tidal wave to the population of geologists patiently waiting to their experiments.

In order to develop the computer vision model, we set out to first understand the core sample photographs by deploying DNN techniques along with a RNN technique against the features extracted from the DNN technique. Simply put, we have to first collect the photophraphs and place them in a format whereby the pixels can be analyzed, recorded, and classified accordingly. The team employs a Convolutional Neural Network (CNN) for analyzing the information and building a model. A key point worked through as part of this project, was the notion of time-series. The deeper a core sample goes, the longer that time has passed. i.e., earlier periods in the earths history are represented as lower in the vertical picture. However, there is not a clear indicator from the samples when one time period ends, and another begins. For that reason, RNN would be potentially ideal for developing the model, but CNN is also explored given its reliance on fixed-size inputs and outputs (when analyzing pixelated images and their location within the larger photograph).

The end result the output of the model is a flattened data file that shows on a color scale against a location allowing the possible graphing of the data. The conclusions that can be reached is that a proof of concept computer vision model was developed whereby the images used to train and test the model were correctly assigned XX amount of the time.

The remainder of this paper includes an overview of the data gathered and utilized; the methods and experiments performed; the results (of the model developed); an analysis of those results (i.e., accuracy indications); ethical considerations; and overall conclusions. Finally, this paper contains a summary of potential future work as well as the associated references and appendices.

2 Background

The Integrated Ocean Drilling Program Expedition 320/321, ”Pacific Equatorial Age Transect” (Sites U1331U1338), was designed to recover a continuous Cenozoic record of the paleoequatorial Pacific by coring above the paleoposition of the Equator at successive crustal ages on the Pacific plate. These sediments record the evolution of the paleoequatorial climate system throughout the Cenozoic. As we gained more information about the past movement of plates and when in Earth's history ”critical” climate events took place, it became possible to drill an age transect (”flow line”) along the position of the paleoequator in the Pacific, targeting important time slices where the sedimentary archive allows us to reconstruct past climatic and tectonic conditions.

The Pacific Equatorial Age Transect (PEAT) program cored eight sites from the sediment surface to at or near basement, with basalt aged between 53 and 16 Ma, covering the time period following maximum Cenozoic warmth, through initial major glaciations, to today. The PEAT program allows the reconstruction of extreme changes of the calcium carbonate compensation depth (CCD) across major geological boundaries during the last 53 m.y. A very shallow CCD during most of the Paleogene makes it difficult to obtain well-preserved carbonate sediments during these stratigraphic intervals, but we recovered a unique sedimentary biogenic sediment archive for time periods just after the Paleocene/Eocene boundary event, the Eocene cooling, the EoceneOligocene transition, the ”one cold pole" Oligocene, the OligoceneMiocene transition, and the middle Miocene cooling. Together with older Deep Sea Drilling Project and Ocean Drilling Program drilling in the equatorial Pacific, we can also delineate the position of the paleoequator and variations in sediment thickness from 150W to 110W longitude.

The results of our paleoequator reconstruction and drill site locations are shown in Figure (see Fig. 1).

2.1 Expedition 320/321 Site Location Strategy

IODP positioned drilling Sites U1331 through U1338 somewhat south of the estimated paleoequatorial position at their target ages (see Fig. **??**) to maximize the time that the drill sites remained within the equatorial zone (i.e., 2 of the equator), to allow for some error in positions (evidence suggests a southward bias of the equatorial sediment mound relative to the hotspot frame of reference [Knappenberger, 2000]), and to place the interval of maximum interest above the basal hydrothermal sediments.

The Eocene was a time of extremely warm climates that reached a global temperature maximum near 52 Ma, a period around the Early Eocene Climatic Optimum (EECO) (see Fig. **??**) (Zachos et al., 2001a; Shipboard Scientific Party, 2004). From this maximum there was gradual climatic cooling through the Eocene to the Eocene/Oligocene boundary. There appears to have been a slight reversal to this trend in the middle Eocene near 43 Ma and in the late Eocene at 3436 Ma, just prior to the pronounced drop in oxygen isotopes that marks the Eocene/Oligocene boundary and one of the most dramatic changes of the CCD (see Fig. **??**).

Throughout the Eocene, the CCD lay near a depth of 3.23.3 km, albeit with potentially significant short-term fluctuations (Lyle et al., 2005). Thus, recovering well-preserved carbonate sediments from the equatorial region is a substantial challenge but not impossible if the depth of the East Pacific Rise lay near the global average of 2.7 km. We presently lack calcareous sediments from the region of the equatorial circulation system during this time of maximum Cenozoic warmth (Zachos et al., 2001a), elevated atmospheric pCO2 concentrations (Lowenstein and Demicco, 2007), and a shallow early Eocene CCD estimated between 3200 and 3300 m water depth (Lyle, Wilson, Janecek, et al., 2002; Lyle et al., 2005; Rea and Lyle, 2005). The Eocene equatorial upwelling system appears to differ from the modern equatorial upwelling regime by having strong secondary upwelling lobes 10 in latitude away from the primary equatorial region (see Figs. **??**). These lobes produced a much broader region of (relatively) high productivity than is present today.

Subsequent paragraphs, however, are indented.

Table 1 gives a summary of all site locations and the geologic ages of the samples recovered.

Table 1. Expedition 320 Coring Summary.

Geologic Epoch	Site Location; Age (Ma)	Core Recovered
Middle Miocene	U1338; 18 Ma crust	0 meters
Miocene	U1337; 24 Ma crust	0 meters
Latest OligoceneEarliest Miocene	U1335; 26 Ma crust	850.78 meters
Oligocene	U1336; 32 Ma crust	438.71 meters
Eocene/Oligocene Boundary	U1334; 38 Ma crust	869.26 meters
Middle and Late Eocene	U1333; 46 Ma crust	531.65 meters
Early and Middle Eocene	U1331-U1332; 53 and 50 Ma crust	427.65, 433.98 meters

3 Dataset and Data Exploration

The team has received copies of core images taken aboard the JOIDES Resolution for geoscience research and paleoclimate modeling. The core images consist of color reflectance line scan images from cores drilled during Expedition 320. The images are medium to ultra high-resolution (centimeter to sub-millimeter scale) consisting of a color space described as L*, a*, b* = lightness reflectance values of sediment as defined in the LAB color model. Upon request, the team will receive additional expedition core images to contribute towards the geocsience research conducted at IODP Headquarters. The data exploration consists of extracting the dominant color reflectance for each core image, which is needed to build the paleoenvironmental models that are researched by the IODP (Texas A&M) scientists and published for world-wide open source use.

Once the images are denoised and processed they represent numerous geologic cross sections of time when the Earth's climate raised to a global maximum and fell to the global minimum during glaciations. These core samples are taken from deep beneath the ocean and below the oceans subsurface floor. As the exploration drilling occurs, the sample is pulled up and analyzed as well as cut in half and then photographed giving valuable data that is undisturbed. Each image is stored and tagged with the sample location as well as the geologic time it represents on Earth.

4 Methodology and Experiments

Note: This section is left intentionally blank as no methods have been finalized and experiments have yet to be conducted beyond the exploration phase of the project.

Displayed and numbered equations such as Equation 1 are centered and set on a separate line.

$$x + y = z \tag{1}$$

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Proof. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

5 Results

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

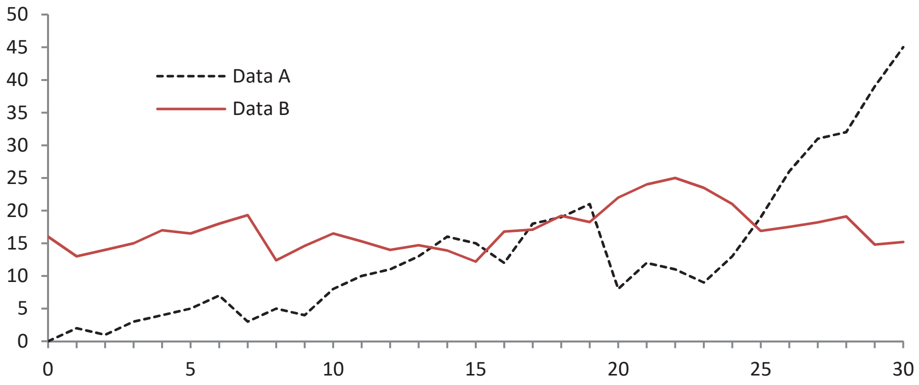

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

6 Results

This section should show the test results when taking the model through the selected test image(s). As no methods or experiments are ready for provision, no results are available as well.

7 Analysis

Note: The use of Mean Absolute Error is proposed for comparison of model accuracy and the use of McNemar's hypothesis test is being researched for use in a statistical test measuring the difference in the paired core image color reflectance data measured.

8 Ethics

This work can be seen as contributive to larger scientific endeavors focused on better understanding Earth in terms of a variety of considerations including climate. The information, data, and assistance required to perform this work cannot be seen as infringing upon another entitys labors or proprietary, valuable information. In a very general sense, this work has extraordinarily limited ethical concerns.

9 Conclusions

The overall conclusions reached are that...

10 Future Work

The model contained in this paper was built and trained using a limited number of sample photographs. As additional core samples are gathered and analyzed, the model should continue to be trained for increased performance in terms of accuracy.

Additionally, as this paper sets out, the model being built is one geared towards computer vision with a focus on transforming image information into data files. Once this is accomplished, and a model is put into production, the possibility for future work is so broad that this paper cannot possibly list in completeness. However, some immediate steps would be possibly deploying the computer vision model to other images sets (for retraining). Additionally, there will be multiple opportunities to continue deploying machine learning methods for other issues facing the geologists (now that they have this data) including models to infer the contents of the Earth between the location of the samples; predicting characteristics of the ocean the Earth from the color shades of the core; etc.

For citations of references, we prefer the use of square brackets and consecutive numbers where the references are numbered according to the ascending alphabet (i.e., a to z) of the first author's last name. Citations using labels or the author/year convention are not acceptable. The following reference provides a sample reference list with entries for journal articles [2] and [1].

Note that URLs should be placed in Footnotes[2] and not in the references. Only documents that will not change over time should be placed in the references and cited.

References

1. Barker, S., Diz, P.: Timing of the descent into the last ice age deter- mined by the bipolar seesaw. Paleoceanography and Paleoclimatology (29(6)), 489–507 (2014)
2. Lamy, F., W.G., Alvarez Zarikian, C.: Expedition 383 scientific prospectus: Dynamics of the pacific antarctic circumpolar current (dynapacc). International Ocean Discovery Program (2018)

[2] More information may be found at http://www.smu.edu. Last accessed 31 Dec 2018.