July 2020
I

Joined Newsletter of the *Digital Earth* Project

Contributions from Digital Earth partners presented at the 2nd Annual Meeting

The 2nd Annual Meeting was held from 26-28th May 2020. Due to the COVID-19 pandemic, the meeting could not be hosted at KIT in Karlsruhe as initially planned, but was conducted as a digital format. The first day of the meeting was dedicated to scientific outcome with 16 scientific contributions presented as a public online session. In total, almost 80 people joined this session on this first day. This newsletter shows results presented in the framework of the *Digital Earth* 2nd Annual Meeting.
This newsletter shows a selection of results presented in the framework of the *Digital Earth* 2nd Annual Meeting that took place from 26 to 28 May 2020.

Deep neural networks for total organic carbon prediction and data-driven sampling
Everardo González and Ewa Burwicz
GEOMAR Helmholtz Centre for Ocean Research Kiel

World's global ocean comprises of about 72% of the total Earth's surface. However, due to its size and available technology, direct seafloor samples collected so far are sparse in space. The existing data sets on sediment composition are inadequate to quantify the fluxes of carbon and other seawater constituents across the seabed at global scale. Sediment and ocean models are heavily relying on these fluxes to simulate the uptake of atmospheric CO_2 and the biogeochemical cycles in the ocean. Moreover, the challenging sampling campaigns are often restricted by the amount of ship time, funds, and the lack of consistent methodologies to collect and process the data.

To overcome this problem, machine learning methods were adapted to marine sciences to approximate the seafloor physical and biogeochemical properties without the urge of direct sampling. Some of these methods (e.g. k-Nearest Neighbors) provide a sophisticated averaging tool to estimate the seafloor property (e.g. organic carbon content) based on the data points nearest in space. However, this approach performs better in more homogenous environments, which does not apply to global scale problems.

Over the past decade, deep learning has been used to solve a wide array of regression and classification tasks. Compared to classical machine learning approaches (k-Nearest Neighbors, Random Forests etc.), deep learning algorithms excel at learning complex, non-linear internal representations in part due to the highly over-parameterized nature of their underling models; thus, this advantage often comes at the cost of interpretability. In this work, we used deep neural networks (DNN) to assess global total organic carbon (TOC) seafloor concentration map. Implementing Softmax distributions on implicitly continuous data (regression tasks), we were able to obtain probability distributions and the model's intrinsic information. A variation of Dropout method i.e. the Monte Carlo Dropout is used during the inference step providing a tool to model prediction reliability. Additionally, to global TOC predictions, we used these techniques to create a model information map which is a key element to develop new data-driven sampling strategies for data acquisition. This model information map provides a quantitative analysis of the model information and allows us to define geographical locations that are under-sampled. By acquiring information at these selected coordinates during the research cruises and sampling planning programs, we will be able to quickly improve our overall global predictions.

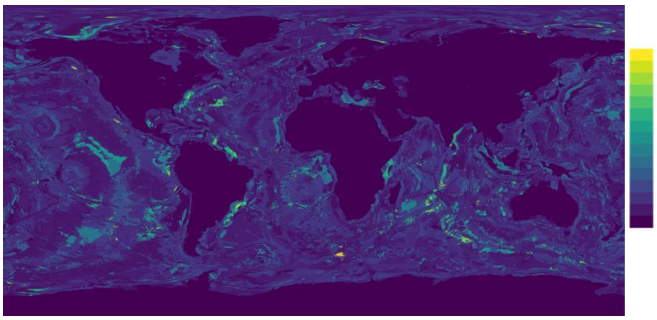

Figure 1: Global seafloor Total Organic Carbon predictions (wt. %) derived from the new Deep Neural Network model.

Estimation of methane and ethane concentrations by means of neural network
Andrey Vlasenko, Volker Matthias, Ulrich Callies
Helmholtz-Zentrum Geesthacht - Centre for Materials and Coastal Research (HZG)

Methane is one of the most important greenhouse gases present in the atmosphere, and therefore its estimate is one of the priority directions in environmental sciences and the *Digital Earth* project in particular. Methane has several natural and anthropogenic sources such as swamps, animal husbandry, growing rice, oil and gas exploration, fuel combustion, etc. At present, the most popular method to estimate methane concentration in air employs discrete chemical transport models (CTMs). Typical CTMs use emissions and meteorological data as inputs from which they calculate concentrations, transport and transformation of chemicals in the atmosphere. Such models have been continuously been improved including more details so that they generally require much computational power. Neural networks (NN) may become a cheaper alternative to CTM in terms of computational resources. We expect that in certain cases fast NNs could substitute CTMs with comparable accuracy. We test this concept considering the example of ethane and methane with two independent NNs.

We design the first NN for searching anomalies in methane and ethane concentrations during cruise measurements near oil fields in the North Sea basin. Note that nineteen percent of atmospheric methane is associated with oil and gas mining production, more than half of it leaks directly from the fields. The NN, installed on a laptop, can use current physical parameters of the surrounding atmosphere from the onboard ship sensors and estimate the methane and ethane concentration in this location. If the estimate does not match

measured concentrations, one can conclude that an anomaly was detected which may be associated with new oil or gas fields or substantial changes in the known ones. In this sense, the NN is a compact smart monitoring tool. The NN doing this job is quite simple: it consists of three dense layers, nine inputs, and one output. We used the available cruise data to test and train it. An example of NN estimates compared to the real measurements is given in Figure 2. Being relatively simple, this NN can nevertheless reliably predict local methane concentrations near the previous cruise measuring routes. It must be noted, however, that oil fields contribute only an estimated 19% to methane concentrations in the atmosphere. Thus, measured anomalies can also originate from other sources.

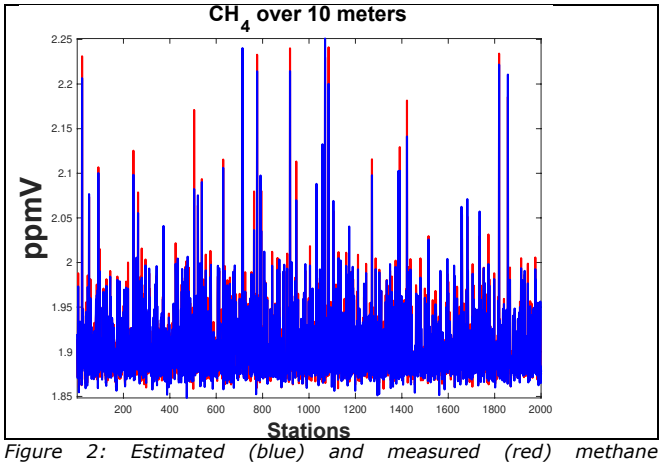

Figure 2: Estimated (blue) and measured (red) methane concentrations from cruise data in the North Sea.

To exclude the impact from other sources, we developed the second NN, which estimates ethane concentrations in the atmosphere. Note that natural gas contains up to several percent of ethane, giving 60% of the atmospheric ethane. The ethane/methane concentration ratio is unique for each oil or gas field, serving as a kind of fingerprint. Methane has a long lifetime (decades), it has several sources, and therefore it is hard to detect whether measured gas is leaking now from an oil field, or if it came from a different source. Ethane is less persistent, but it has a long enough lifetime (several weeks) for detection. We trained and validated the second NN on output from the **C**onsortium **M**ultiscale **A**ir **Q**uality Model (CMAQ) for the European domain. At present, the second NN explains more than 50% off-seasonal ethane variability. An example of ethane estimates by means of CTM CMAQ and NN is shown in Figure 3.

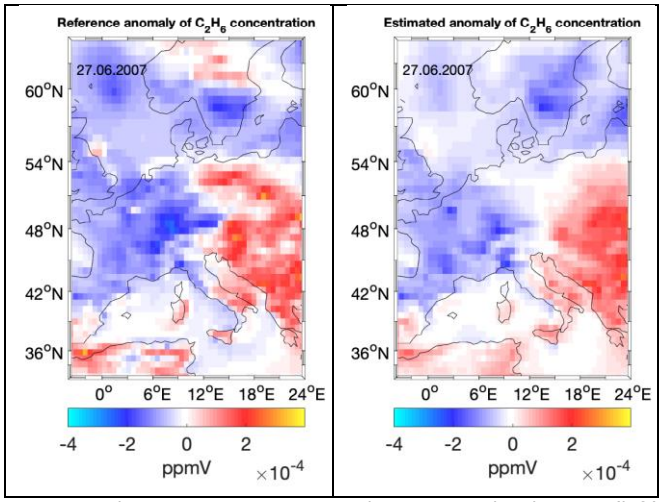

Figure 3: Ethane concentration anomalies estimated with CMAQ (left) and NN (right).

Similar to the CTM, the NN calculates concentrations for all nodes of the discretized European domain. In our case, the spatial resolution is 64x64 km. Such discretization is too coarse to detect anomalies in ethane emission from gas fields as their characteristic scale is only a couple of kilometers. However, the NN can identify ethane stemming from other sources. Calculations with the first NN do not involve any spatial discretization, but applicability of this NN is limited to a very narrow area. The strength of our approach lies in the combination of the two NNs. Improving the method of such combination will be a focus point of our future research within the Digital Earth project.

Using machine learning for automated site detection of seafloor massive sulfides

Amir Haroon
GEOMAR Helmholtz Centre for Ocean Research Kiel

In 2016, geophysical and geological data were acquired during research cruise M127 at the slow-spreading TAG Midocean Ridge Segment to detect known and unknown seafloor massive sulfides (SMS), and asses the resource potential of the ocean floor in this region. The obtained data has been previously evaluated using standard geoscientific methodologies and now serves as a pilot test case for implementing data science methodology in an automatic SMS site detection workflow using multivariate geophysical and geological data.

In the framework of a Digital Earth Bridging Postdoc, we aim to not only implement existing machine learning methodologies to predict the resource potential of the seafloor, but also obtain a robust quantification of the prediction uncertainty. A first milestone needed to achieve this overarching goal deals with geophysical/geological data acquired on various spatial scales. During the 2nd Digital Earth Annual Meeting we presented an initial workflow for integrating sparsely and spatially sampled data onto a continuously sampled grid using a sequential application of fuzzy clustering with random forest regression.
The spatially sampled data sets (e.g. depth, slope, aspect, ruggedness, reduced-to-pole magnetics) were used to create a segmented map of the seafloor combining regions of similar behavior into common clusters. The pixel fuzziness is then used in a random forest regression approach at the defined nodes of the sparsely sampled training data set (apparent conductivity) to derive a model that allows us to extrapolate the sparsely sampled conductivity data onto the local scale of our region of interest. Note that apply this approach to conductivity data as it is a direct indicator for the presence of SMS occurrences on the seafloor.
Figure 4 illustrates how the sparsely sampled apparent conductivity data (left panel) is first applied to derive a model capable of predicting values at the defined points of the validation data set (central panel), and subsequently applied to extrapolate across the entire region of interest (right panel). Areas appearing in purple/pink within the vicinity of the sparsely sampled data are in good agreement with the dimensions of known SMS occurrences. The areas highlighted in purple and pink to the lower left side of the right image are associated with biases or prediction errors of an insufficient regression model and do not coincide with known SMS sites. Here, model improvements and a robust assessment of uncertainty in our prediction are needed.

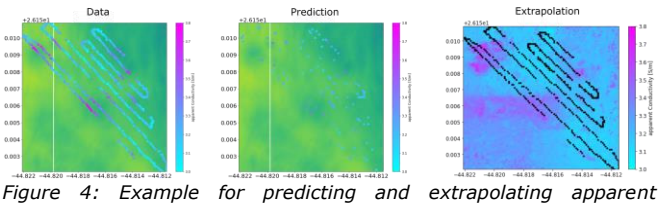

Figure 4: Example for predicting and extrapolating apparent conductivity data using a sequential fuzzy clustering with random forest regressor models.

Seafloor & Terrain Sampling based on auxiliary information: A collection of methods

Iason-Zois Gazis
GEOMAR Helmholtz Centre for Ocean Research Kiel

A common concern in scientific research is the amount and quality of the available data to derive 'data-supported' conclusions. In marine research, and particularly in deep-sea studies, acquiring the needed amount of data at the required position is difficult as it is time consuming (e.g. taking one samples in 4km water depth takes 4h) and positioning the sampling device has typically an error between 100 and 10m.As such, deep sea studies typically suffer from the scarcity of ground truth samples. Physical samples taken by grabs or corers as well as photo/video surveys done by ROV, AUV or towed cameras have several constraints, regarding weather conditions and limited spatial coverage. A box corer taken within 4h only samples 0.25m^2 of seafloor Similar to marine research, terrestrial studies (e.g. soil mapping) also face difficulties in ground truth sampling due to physical (e.g. high terrain roughness and steepness), consensual (e.g. presence of infrastructures and land ownership) and conservation barriers (e.g. protected areas). Since the data availability is limited, the quality of the data has increased weight.

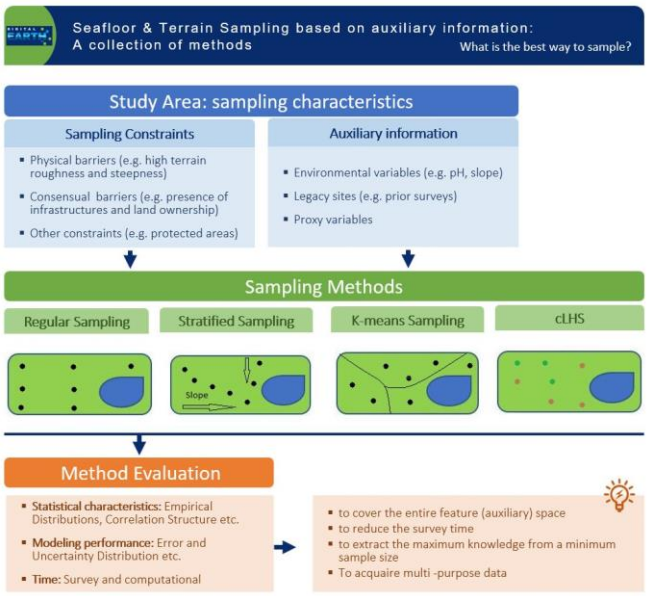

Figure 5: The proposed workflow for the evaluation of different sampling methods in seafloor and terrain research.

One milestone in data quality is the ability of the sample (physical sample or photo) to adequately represent the variable of interest, allowing a rigorous inference about the sample composition. Moreover, the same sample should provide information for more than one examined variable. and it should be able to capture all environmental variables that have a causal effect (direct or indirect) to the examined variables. This highlights the importance of **sample representativeness on a multivariate level**. In this respect, different sampling strategies have been evaluated: random sampling, regular sampling, and feature space clustering based on Clustering Large Application method - CLARA (Kaufman and Rousseeuw, 1990), and conditional Latin Hypercube – cLHS (Minasny & McBratney, 2006). Their performance is evaluated based on their ability to a) represent a uniform distribution of the examined variables (range, mean value etc.), b) to predict the empirical distributions of continuous and categorical auxiliary variables, c) to keep the correlation structure among the continuous auxiliary variables, and d) to consider survey cost and time and keep it to a minimum.

The results show that in small and homogeneous areas all method can provide sufficient information, with the regular sampling having the maximum geographical coverage. As the seafloor/terrain heterogeneity increases, CLARA clustering and cLHS can capture more efficiently the auxiliary space, especially when a decreased amount of sampling points is used. However, CLARA and cLHS are sensitive to the parameters that are used. For CLARA, the number of clusters has to be selected based on internal clustering criteria (cohesion & separation), evaluated by statistical indexes such as Calinski-Harabasz (Calinski & Harabasz, 1974). After an optimal clustering has been achieved, a stratified sampling among the clusters (based on survey needs) is performed. In cLHS the number of iterations and the cooling temperature have to be set appropriately during the simulated annealing in order to achieve the optimum global solution. Nevertheless, the ability of using cost surfaces (e.g. restricted areas, or legacy sites) makes the use of cLHS a useful decision sampling tool in complex environments or/and in repeated monitoring surveys.

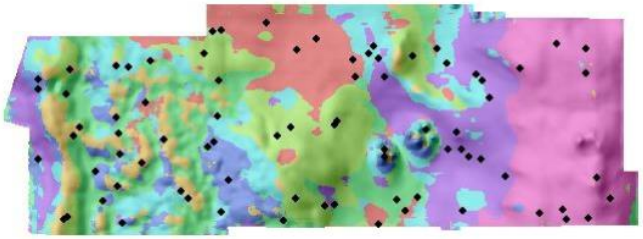

Figure 6: Seafloor future space clustering (CLARA method), based on bathymetry and several bathymetric derivatives (e.g. slope, rugosity, concavity). On top (black dots) are the sampling locations based on stratified sampling among the clusters.