Random Forest (Machine Learning) Interpolation of Seabed Point Data to Grid

dbSEABED Project
4 March 2018

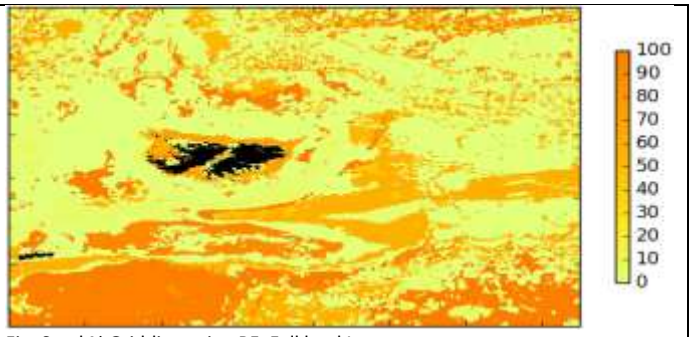

Fig: Sand % Gridding using RF, Falkland Is

Introduction.

Recently, Machine Learning (ML) methods have revolutionized map interpolation/extrapolation and dbSEABED now embraces those methods. However, the implementation is still in a stage of learning and adjustment. The particular ML method used at present is Random Forests (RF), one of the tree-based ensemble methods (Scikit-Learn Developers 2018; Pedregosa et al. 2011). Briefly, a training set of the parameter values and terrain variables ('features') is made and then analysed for structure. Tree collections are tested for natural divisions and variances on the splits are assessed using those collections. Using the trained structure, a larger test dataset – the entire map area in terms of the terrain variables – is submitted and processed. A 'fit' is calculated of the 'test' data to the 'train' data and the parameter values are distributed on those patterns. This is a very high-level explanation of the RF process.

Training Data.

For this study the guiding parameters included: regional bathymetry, seafloor slope, seafloor roughness (BPI & TRI; see Wilson et al. 2007), seafloor curvature (divergence – Figure 3 - and large eigenvalue of Hessian Matrix), bottom-water temperature (Figure 4), and surface-water temperature. Other parameters might be included, but only these had been prepared by the time of the study. There is an obvious rationale for including the seafloor parameters in terms of seafloor smoothing by sediment, rock roughness, sediment slope-stabilities, and the transport distances for sediment grades. The temperature layers can also be rationalized in terms of effects on biological colonizations (e.g., shell, coral), current winnowing of sediments, and on iceberg-drift limits for Ice-Rafted Debris (IRD). Interestingly, the temperature layers achieved very high importance scores in the RF, for all mapped parameters.

Adjustments.

Deciding the best settings for the ML operations required some care. As per normal practice, the number of parameters (features) was made as large as possible, and parameter values were normalized to between 0 and 1 – keeping a modest total number of numerical levels (8). The 'train' dataset also had to be conditioned, since in the input data individual types of seabed could be overloaded with multiple sediment type recordings which led to a winner-takes-all problem. The other controls were set to conventional values (see Koehrsen, W. 2018). Overfitting was avoided, since training accuracy values were ~0.7. At the end of the fitting to full data, the OOB (out-of-bag) accuracies of 0.6 to 0.8 were obtained. The final results were scrutinized for plausibility relative to conventional sedimentologic thinking and were passed. It is clear though, that even though the results are already far better than with linear methods, improvements are still possible.

Reservations.

The sampled training data do not cover the full parameter space of the map. The adjustments that are recommended are not easily settled on, and some users regard the methods as 'black box' and irreproducible – a criticism which is often levelled at ML applications. The environmental layers of the training data are usually not germane (at least not simply) to the problem of sediment types.