Appendix DR1. Description of datasets and methods, Table DR1, Figures DR1-DR9

Environmental predictors of deep-sea polymetallic nodule occurrence in the global ocean

Adriana Dutkiewicz, Alexander Judge, R. Dietmar Müller

Correspondence to: adriana.dutkiewicz@sydney.edu.au

DATASETS AND METHODS

Sample localities

We selected a total of 2440 deep-sea surface nodules based on a comprehensive search of open access databases of georeferenced seafloor samples: the Index to Marine and Lacustrine Geological Samples (IMLGS; https://www.ngdc.noaa.gov/mgg/curator/, accessed July 2017 (Curators of Marine and Lacustrine Geological Samples Consortium: Index to Marine and Lacustrine Geological Samples (IMLGS). National Geophysical Data Center) and Pangaea (https://www.pangaea.de, accessed October–December 2017). Our selection is based on strict, but simple criteria. The presence of manganese, ferromanganese or polymetallic nodules had to be clearly stated in the grab/dredge/corer/trawl net sample description in the databases and associated datasets. Micronodules, and nodules located on the continental shelf (water depth < 200 m), in lakes, or in marginal seas (e.g., Baltic Sea, Mediterranean Sea) were not included. The nodules are located at or just below the sediment surface (0–10 cm sediment depth). Because some of the environmental datasets used in our analysis (see Figs DR1–DR7) lack a complete coverage of the open ocean (e.g., figs DR1–DR2, DR5), our nodule sites are restricted to localities for which we were able to extract environmental data from all datasets.

Our 8023 control surface samples comprise sites for which the presence of nodules was not reported. These sites were extracted from the Index to Marine and Lacustrine Geological Samples (IMLGS) and equivalent selection criteria were applied as for the nodule sites.

Environmental dataset

The datasets, sampled at 0.1° grid resolution, and their sources are shown in Figs 1 and DR1–DR7.

Statistical Analysis

For the purposes of constructing a discriminative model of nodule occurrence we applied a number of transformations to the oceanic variable data at the exploration locations. First, in order to approximate a spatially uniform sample, a Fibonacci lattice (González, 2010) of 1000001 points was generated for each variable via nearest-neighbor interpolation on the regular latitude-longitude data grids. Then a quantile transformation was applied to the continuous variables based upon the reference sample corresponding to the Fibonacci lattice, resulting in transformed values uniformly distributed between 0 and 1. Quadratic terms of the form $x_i x_j$ were then generated using the continuous variables, while the categorical lithology variable was converted into 8 binary variables $l_k$. Here the subscripts “i” and “j” enumerate the continuous variables. The lithology can be represented by 8 binary variables, used to
represent all possible combinations between the categorical lithologies and the continuous variables \((x_i l_k)\), as well as the interactions of the lithology with the previously computed interactions, leading to an array (the vector \(\mathbf{z}\)) representing the interaction terms \(x_ix_j, x_ilk\) and \(x_ix_jlk\), which together is the full complement of derived features.

To test for the existence of environmental bias due to the location of the exploration samples (i.e., the nodule and control samples) we calculated the p-value for the Kolmogorov-Smirnov statistic obtained from comparing the variable values at Fibonacci lattice sites with those at the exploration sites. This p-value represents the probability of obtaining a larger discrepancy between the empirical cumulative distribution functions of the two samples if they arose from the same distribution. The results displayed in Table DR1 are all vanishingly small, implying that the likelihood of the two samples arising from the same distribution is essentially zero. Therefore, without an overwhelming prior probability of the contrary, this indicates the existence of strong bias in both geographical and environmental space associated with the exploration samples, which must be addressed. We assumed that this bias could be accounted for (i.e. modelled in principle) by variations in the features only. Denoting the occurrence of exploration by the binary variable \(s\), this assumption implies that the occurrence of nodules, indicated by the binary variable \(y\), is conditionally independent of \(s\) given \(\mathbf{z}\), viz.

\[
P(y = 1|z, s = 1) = P(y = 1|z),
\]

where \(P(y = 1)\) is the probability of nodule occurrence. In this case a Baysian logistic regression model with a likelihood of the form

\[
P(y = 1|\mathbf{z}, \mathbf{\beta}) = \sigma(\mathbf{\beta} \cdot \mathbf{z}) = \frac{1}{1 + \exp(-\mathbf{\beta} \cdot \mathbf{z})},
\]

where \(\mathbf{\beta}\) is a vector of model coefficients, can be shown to be independent of the sampling bias (Zadrozny, 2004) and is therefore, along with the quadratic interaction terms included in \(\mathbf{z}\), a natural choice of discriminative model to capture the basic linear and nonlinear interactions important to the problem at hand. Additionally, in the context of species distribution modelling in ecology, with which the present study holds many parallels, the program of using one set of response and predictor variables and a single model type is considered as standard practice (Araújo et al., 2019). Furthermore, a Bayesian treatment of such a model automatically produces uncertainty estimates associated with predictions. Placing a Gaussian prior with mean zero and precision parameter \(\alpha\) over the coefficients \(\mathbf{\beta}\), the posterior over the model parameters becomes

\[
p(\mathbf{\beta}|\mathbf{D}, \alpha) = \frac{\prod_i P(y_i = 1|z_i, \mathbf{\beta}) p(\mathbf{\beta}|\alpha)}{p(\mathbf{D}|\alpha)},
\]

where \(\mathbf{D}\) represents the \(N\) data pairs \(\mathbf{(y_i, z_i)}\), and \(p(\cdot)\) is a probability density. The probability of nodule occurrence for new data is then obtained as the expected nodule occurrence probability,

\[
P(y = 1|\mathbf{z}, \mathbf{D}, \alpha) = \int d\mathbf{\beta} \; P(y = 1|\mathbf{z}, \mathbf{\beta}) p(\mathbf{\beta}|\mathbf{D}, \alpha).
\]
The logistic regression model was fitted using the maximum marginal likelihood approach employing the Laplace approximation (Bishop, 2006; MacKay, 1992). The posterior for the model parameters then takes the form

\[ p(\beta|D, \alpha) = N(\beta|\hat{\beta}, \Sigma_\beta), \]

where \( N(x|\mu, \Sigma) \) is a normal distribution over \( x \) with mean \( \mu \) and covariance matrix \( \Sigma \), \( \hat{\beta} \) is the value at the posterior maximum, and \( \Sigma_\beta \) is the posterior covariance. This form allows us to write

\[ P(y = 1|z, D, \alpha) = \int da \sigma(a) N(a|\mu_a, s_a), \]

where \( \mu_a = \hat{\beta} \cdot z \) and \( s_a = z^T \cdot \Sigma_\beta \cdot z \). Upper and lower bounds of the 90% credible interval for the probability may then be obtained by evaluating the percent point function of the normal distribution \( N(a|\mu_a, s_a) \) at 0.05 and 0.95, respectively, and inserting the resulting values of \( a \) into the sigmoid function \( \sigma(a) \).

Characterization of the environments where nodules are deemed by the model to likely occur was achieved through the conditional distributions of the oceanic variables viz,

\[ p(x|y = 1, D, \alpha) \propto P(y = 1|z(x), D, \alpha) p(x). \]

We approximated these distributions by using the Fibonacci lattice values as a sample from \( p(x) \) and constructing a histogram with sample weights proportional to the modelled values of \( P(y = 1|z(x), D, \alpha) \). Error bounds on these distributions were then obtained by replacing expected nodule probability with the values representing the upper and lower bounds of the credible interval calculated previously.

The marginal importance of each variable was modelled by an estimate of the mutual information between the spatial distribution of the variable and that of the probability of nodule occurrence. A set of estimates corresponding to 100 samples from the posterior distribution over \( \beta \) was calculated for each variable empirically (Kraskov et al., 2004) using the variable and probability values at the Fibonacci lattice points. The final estimates and their errors were obtained as the mean and standard deviation, respectively, of the 100 samples, this number being chosen on the basis that it provided stable statistical estimates within a practical computational timeframe.
Figure DR1. Nodule occurrence (A) and control sample occurrence (B) relative to long-term sedimentation rate. Nodules preferentially occur in regions dominated by low decompacted long-term sedimentation rates (Dutkiewicz et al., 2017) that are < 0.5 cm/ky. Relative variable importance is 1±0.05. White dots denote polymetallic nodule locations. Dark gray denotes regions of no data. White lines denote mid-ocean ridges. Mollweide projection.
Figure DR2. Nodule occurrence (A) and control sample occurrence (B) relative to bottom water oxygen concentration. Nodules preferentially occur in regions dominated by moderately high bottom water oxygen concentration (Seiter et al., 2005) between 150 and 200 mmol/m³. Relative variable importance is 0.49±0.03. Black dots denote polymetallic nodule locations. Dark gray denotes regions of no data. Black lines denote mid-ocean ridges. Mollweide projection.
Figure DR3. Nodule occurrence (A) and control sample occurrence (B) relative to average summer sea-surface productivity. Nodules preferentially occur on the seafloor where the average summer surface productivity is < 300 mgC/m²/day. The surface productivity dataset based on Sea-viewing Wide Field-of-view Sensor (SeaWiFS) r2010 reprocessing results was obtained from http://orca.science.oregonstate.edu/1080.by.2160.monthly.hdf.vgpm.s.chl.a.sst.php, with summer productivity calculated as outlined in Dutkiewicz et al. (2015). Relative variable importance is 0.31±0.02. Black dots denote polymetallic nodule locations. Black lines denote mid-ocean ridges. Mollweide projection.
Figure DR4. Nodule occurrence (A) and control sample occurrence (B) relative to seafloor megafaunal biomass (epibenthic invertebrates and demersal fishes). Nodules preferentially occur in regions where the seafloor megafaunal biomass (Wei et al., 2010) is < 1 log mgC/m². Relative variable importance is 0.28±0.02. Black dots denote polymetallic nodule locations. Black lines denote mid-ocean ridges. Mollweide projection.
Figure DR5. Nodule occurrence (A) and control sample occurrence (B) relative to bathymetry. Nodules preferentially occur in regions where the bathymetry (Amante and Eakins, 2009) is > 4500 m. Relative variable importance is 0.21±0.02. Black dots denote polymetallic nodule locations. Black lines denote mid-ocean ridges. Mollweide projection.
Figure DR6. Nodule occurrence (A) and control sample occurrence (B) relative to total organic carbon in seafloor sediment. Nodules preferentially occur in regions where the total organic carbon (Lee et al., 2019) content in seafloor sediments (< 5 cm sediment depth) is 0.25 wt% and 0.5 wt%. Relative variable importance is 0.25±0.02. Black dots denote polymetallic nodule locations. Dark gray denotes regions of no data. Black lines denote mid-ocean ridges. Mollweide projection.
Figure DR7. Nodule occurrence (A) and control sample occurrence (B) relative to bottom water current speeds. Nodules preferentially occur in regions where the bottom current speed fluctuations (Thran et al., 2018) (expressed as standard deviation) is < 1 cm/s. Relative variable importance is 0.10±0.02. Black dots denote polymetallic nodule locations. Black lines denote mid-ocean ridges. Mollweide projection.
Figure DR8. Upper bound on the probability of nodule occurrence. The upper bound of the 90% credible probability interval of nodule occurrence is shown. Dark gray denotes regions of no data. Black outline indicates known nodule fields and poorly-explored areas that are deemed to be permissive for the economic development of nodules (Hein et al., 2013). Black dots denote nodules. White lines denote mid-ocean ridges. Mollweide projection.

Figure DR9. Lower bound on the probability of nodule occurrence. The lower bound of the 90% credible probability interval of nodule occurrence is shown. Dark gray denotes regions of no data. Black outline indicates known nodule fields and poorly-explored areas that are deemed to be permissive for the economic development of nodules (Hein et al., 2013). Black dots denote nodules. White lines denote mid-ocean ridges. Mollweide projection.
Table DR1. **Kolmogorov-Smirnov statistic p-values to test for spatial sample bias.** The p-values are obtained from comparing variable samples at the spherical Fibonacci lattice locations to those from the (irregularly distributed) exploration locations. The p-values indicate significant sample bias.

<table>
<thead>
<tr>
<th>Variable</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sedimentation rate</td>
<td>$7 \times 10^{-134}$</td>
</tr>
<tr>
<td>Bottom water oxygen concentration</td>
<td>$6 \times 10^{-196}$</td>
</tr>
<tr>
<td>Average surface productivity</td>
<td>$6 \times 10^{-14}$</td>
</tr>
<tr>
<td>Seafloor megafaunal biomass</td>
<td>$1 \times 10^{-23}$</td>
</tr>
<tr>
<td>Seafloor organic carbon content</td>
<td>$5 \times 10^{-228}$</td>
</tr>
<tr>
<td>Bathymetry</td>
<td>$2 \times 10^{-143}$</td>
</tr>
<tr>
<td>Bottom current speed standard deviation</td>
<td>$1 \times 10^{-19}$</td>
</tr>
</tbody>
</table>

**References and Notes**

Amante, C., and Eakins, B. W., 2009, ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis, National Geophysical Data Center, NOAA, NOAA Technical Memorandum NESDIS NGDC-24, NOAA Technical Memorandum NESDIS NGDC-24, p, 10.7289/V5C8276M.


Dutkiewicz, A., Müller, R. D., O’Callaghan, S., and Jónasson, H., 2015, Census of seafloor sediments in the world’s ocean: Geology, v. 43, p. 795-798, [https://doi.org/10.1130/G36883.1](https://doi.org/10.1130/G36883.1).


