minerals

MDPI

Article

Application of General Linear Models (GLM) to Assess Nodule Abundance Based on a Photographic Survey (Case Study from IOM Area, Pacific Ocean)

Monika Wasilewska-Błaszczyk * and Jacek Mucha

Department of Geology of Mineral Deposits and Mining Geology, Faculty of Geology, Geophysics and Environmental Protection, AGH University of Science and Technology, 30-059 Cracow, Poland; jacekm@agh.edu.pl
* Correspondence: wasilews@agh.edu.pl

Abstract: The success of the future exploitation of the Pacific polymetallic nodule deposits depends on an accurate estimation of their resources, especially in small batches, scheduled for extraction in the short term. The estimation based only on the results of direct seafloor sampling using box corers is burdened with a large error due to the long sampling interval and high variability of the nodule abundance. Therefore, estimations should take into account the results of bottom photograph analyses performed systematically and in large numbers along the course of a research vessel. For photographs taken at the direct sampling sites, the relationship linking the nodule abundance with the independent variables (the percentage of seafloor nodule coverage, the genetic types of nodules in the context of their fraction distribution, and the degree of sediment coverage of nodules) was determined using the general linear model (GLM). Compared to the estimates obtained with a simple linear model linking this parameter only with the seafloor nodule coverage, a significant decrease in the standard prediction error, from 4.2 to 2.5 kg/m^2, was found. The use of the GLM for the assessment of nodule abundance in individual sites covered by bottom photographs, outside of direct sampling sites, should contribute to a significant increase in the accuracy of the estimation of nodule resources.

Keywords: polymetallic nodules; nodule abundance; general linear models; linear regression; nodule coverage of seafloor; sediment coverage of nodules; Clarion-Clipperton Zone (CCZ); image analysis

check for updates

Citation: Wasilewska-Błaszczyk, M.; Mucha, J. Application of General Linear Models (GLM) to Assess Nodule Abundance Based on a Photographic Survey (Case Study from IOM Area, Pacific Ocean). *Minerals* **2021**, *11*, 427. https://doi.org/10.3390/min11040427

Academic Editors: Pedro Madureira and Tomasz Abramowski

Received: 12 March 2021
Accepted: 12 April 2021
Published: 17 April 2021

1. Introduction

Deposits of polymetallic (manganese-bearing) nodules occurring at the bottom of all oceans are an attractive alternative to onshore deposits from the point of view of metal resources such as Ni, Co, Cu, Mn, Li, REE (the rare earth elements), and others. The Clarion–Clipperton Fracture Zone (CCZ) in the tropical NE Pacific is the area of greatest economic interest for nodules [1–3] (Figure 1A). It is expected that the demand for some metals will soon exceed their supply due to the depletion of onshore ore deposits resulting from the intensive development of modern branches of the economy (high technology, green technology, emerging industries, and military applications). In the future, the shortage of some metals can be covered by the exploitation of offshore deposits. This issue was widely discussed in many publications, e.g., [1,2,4–14]. The exploitation of these deposits, apart from their proper recognition and estimation of their resource [15], requires solving a number of problems related to the technique of mineral extraction and ore processing [7,16–18].

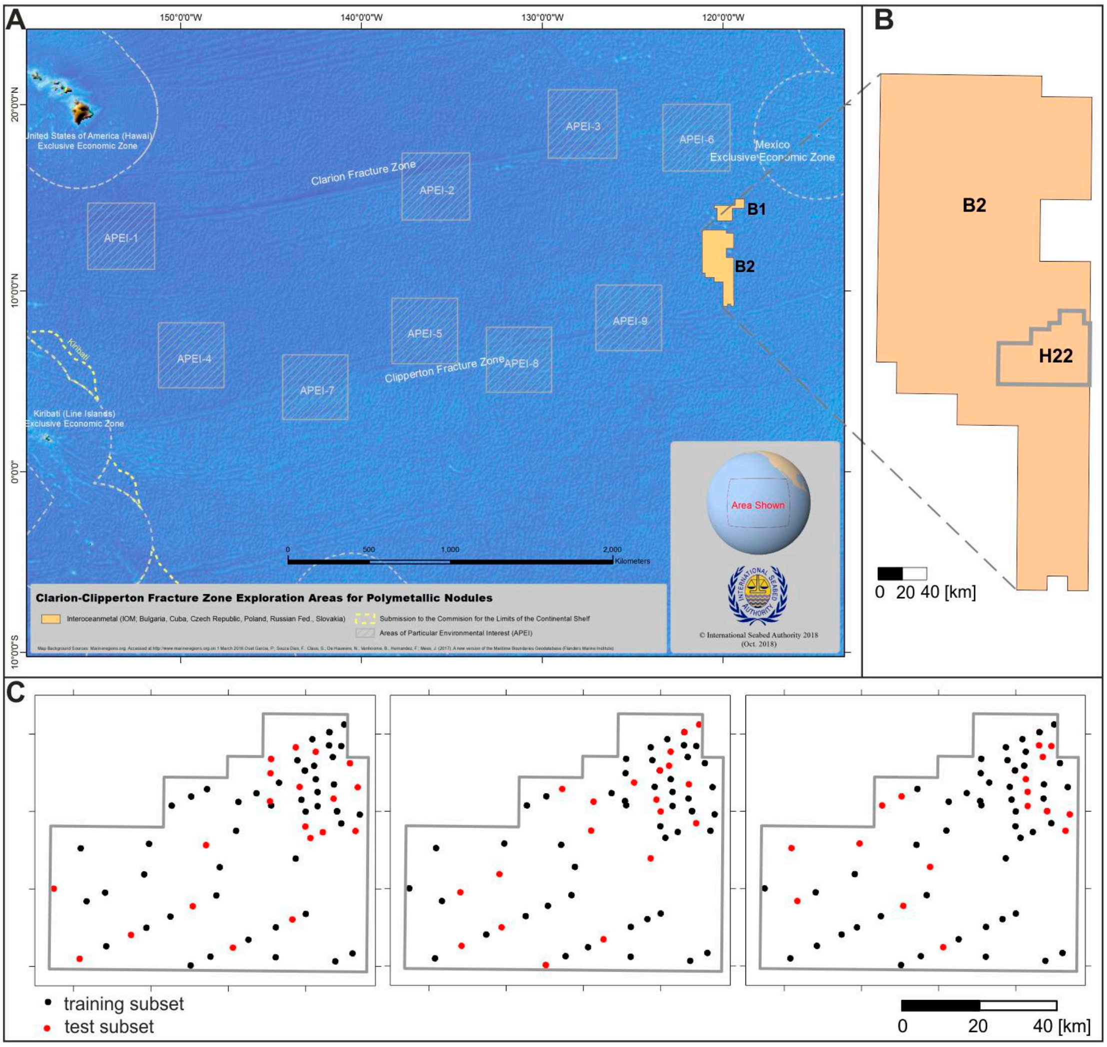

A
150°0'0"W
140°0'0"W
130°0'0"W
120°0'0"W
20°0'0"N
10°0'0"N
0°0'0"
10°0'0"S
United States of America (Hawaii)
Exclusive Economic Zone
Clarion Fracture Zone
Clipperton Fracture Zone
APEI-1
APEI-2
APEI-3
APEI-4
APEI-5
APEI-6
APEI-7
APEI-8
APEI-9
Mexico
Exclusive Economic Zone
Kiribati
Kiribati (Line Islands)
Exclusive Economic Zone
B1
B2
0
500
1,000
2,000
Kilometers
Area Shown
Clarion-Clipperton Fracture Zone Exploration Areas for Polymetallic Nodules
Interoceanmetal (IOM; Bulgaria, Cuba, Czech Republic, Poland, Russian Fed., Slovakia)
Submission to the Commision for the Limits of the Continental Shelf
Areas of Particular Environmental Interest (APEI)
Map Background Sources: Marineregions.org. Accessed at http://www.marineregions.org on 1 March 2018 Oset Garcia, P.; Souza Dias, F.; Claus, S.; De Hauwere, N.; Vanhoorne, B.; Hernandez, F.; Mees, J. (2017). A new version of the Maritime Boundaries Geodatabase (Flanders Marine Institute)
© International Seabed Authority 2018
(Oct. 2018)
B
B2
H22
0 20 40 [km]
C
training subset
test subset
0
20
40 [km]

Figure 1. Location of B2 Interoceanmetal Joint Organization (IOM) exploration area for polymetallic nodules against the background of the Clarion–Clipperton zone [19] (**A**) and H22 exploration block (**B**); the location of the box corer sampling sites and seafloor photographs in the H22 exploration block (three variants of training and test subsets) (**C**).

The development of an appropriate scenario and schedule for the future short and medium-term exploitation of nodules, after meeting the environmental protection requirements and solving technical problems of mining, depends on a detailed recognition of the distribution of nodule abundance and resources and an analysis of the ocean floor topography based on reliable contour maps.

An accurate assessment of the abundance of polymetallic nodules at seafloor sites located far away from direct sampling stations causes many problems. They result mainly from the large distances between sampling stations (which vary depending on the stage of recognition of the nodule-bearing areas), high variability of nodule abundance, and to a lesser extent, from the inevitable errors during the sampling process [20]. Therefore, the attempts to combine the estimation of nodule abundance based on the classical direct sampling (e.g., using box corers) [15,21] with indirect methods, such as photographic surveys [22–25] or widely understood hydro-acoustic methods [26,27], seem natural and rational.

The routine and continuous video and photo-profiling (photographic survey) of the ocean floor along the course of a research vessel from which direct sampling is carried out provides a huge number of photographs. Their analysis provides indirect, approximate information on the percentage of seafloor nodule coverage (i.e., the percentage of seafloor covered by the nodules, hereinafter abbreviated NC-S), the degree of sediment coverage of nodules (SC), and the dominant genetic type of nodules (GT) between sampling stations [28]. The data obtained from the photographs in sampling sites are correlated with various strengths with the nodule abundance based on box core samples. The previous attempts to develop a simple linear regression between the nodule abundance and the percentage of seafloor nodule coverage did not yield unequivocal results. In some parts of the studied area of the Interoceanmetal Joint Organization (IOM), a statistically significant and strong linear correlation between these parameters was found (with linear correlation coefficients of 0.6–0.7), while in other parts there was no statistically significant correlation and the coefficients of linear correlation were close to zero [20,28,29]. The main reasons for the weaker correlation of both parameters can be seen in the varying the degree of sediment coverage of nodules [28–30], diversity of the nodule genotypes [31,32], small scale variability of nodule abundance, different geometrical basis of measurements (area of the bottom covered by the photograph several times larger as compared to the area of the horizontal section of the box corer), and the variation in the quality of seafloor photographs. For these reasons, some researchers introduced various coefficients correcting the relationship between the nodule abundance and the percentage of seafloor nodule coverage determined on the basis of photographs [30,33].

Improvements in the accuracy of the estimates of nodule abundance can be expected when additional independent qualitative variables (ordinal variable), determined based on the photographs, such as the distribution of nodule fractions associated with the genetic type of nodules and the degree of sediment coverage of nodules, are introduced to the relationship model. For the data from the H22 exploration block in the IOM area (Figure 1B), a statistically significant and relatively strong correlation was found between the nodule abundance and the genetic type of nodules, while the correlation between the nodule abundance and the degree of covering the nodules with bottom sediments was weak [28].

The coexistence of independent variables of different types (continuous and ordinal) requires the use of an appropriate mathematical model linking them with the nodule abundance used as a dependent variable. This can be achieved using general linear models (GLM). The results of their application to a dataset from a part of the area administered by the IOM [34] in the Clarion–Clipperton Zone in the Pacific are the subject of this article.

2. Research Objective

The main aim of the research was to assess the accuracy of the prediction of nodule abundance (APN) at ocean floor points outside the sampling stations based on the multivariate regression technique called General Linear Models (GLM) (Figure 2). In the present case study, GLM was used to determine the form of the relationship linking the nodule abundance (continuous dependent variable) with the percentage of seafloor nodule coverage NC-S (independent continuous variable) and two qualitative variables (independent ordinal variables)—the genetic type of nodules in the context of their fraction distribution (hereinafter referred to in the text abbreviated as the genetic type of nodules and marked as GT) and the degree of sediment coverage of SC nodules. The values of all independent variables were determined based on photographs of the ocean floor at box corer stations. The values of the NC-S variable were determined automatically with the use of computer software, while the values of the GT and SC variables were determined visually based on expert evaluation.

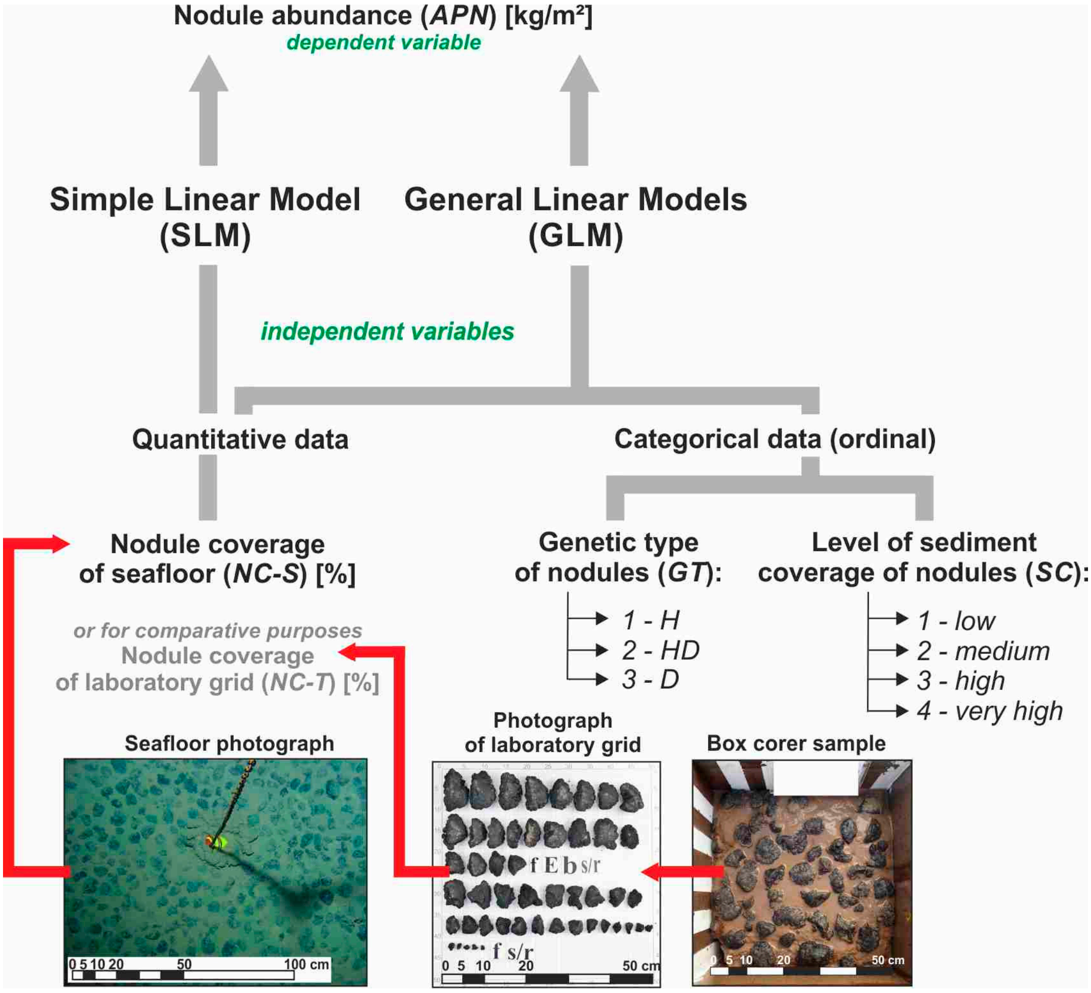

Nodule abundance (*APN*) [kg/m²]
dependent variable
Simple Linear Model (SLM)
General Linear Models (GLM)
independent variables
Quantitative data
Categorical data (ordinal)
Nodule coverage of seafloor (*NC-S*) [%]
or for comparative purposes
Nodule coverage of laboratory grid (*NC-T*) [%]
Genetic type of nodules (*GT*):
1 - H
2 - HD
3 - D
Level of sediment coverage of nodules (*SC*):
1 - low
2 - medium
3 - high
4 - very high
Seafloor photograph
Photograph of laboratory grid
Box corer sample
0 5 10 20 50 100 cm
f E b s/r
f s/r
0 5 10 20 50 cm
0 5 10 20 50 cm

Figure 2. Diagram of the types of data (variables) obtained from photographs of the seafloor and laboratory grid used in the general linear model (GLM) and the simple linear model (SLM) to predict the nodule abundance. Explanations: H, HD, D—genetic types of nodules in the context of their fraction distribution (H—hydrogenetic, HD—hydrogenetic-diagenetic, D—diagenetic).

The fraction distribution of genetic type of nodules (GT symbol, ordinal variable) acts as an identifier and expresses differences in the probability distributions of nodule sizes characteristic for the distinguished genetic types of nodules. The exact characteristics of both ordinal variables and other continuous variables describing the nodule-bearing areas in the H22 exploration block were presented by Wasilewska-Błaszczyk and Mucha [28]. The analysis of variables included in the cited article was the basis for the selection of independent variables in GLM.

To evaluate the effectiveness of GLM as a method for predicting nodule abundance, the obtained results were compared with the results of the prediction for the simple linear model (SLM) linking the nodule abundance APN and the percentage coverage of ocean floor NC-S. For comparative purposes, analogous analyses were also performed for nodules from the box corer samples after their removal and washing and arranging on a laboratory grid, i.e., for the percentage of grid coverage with nodules NC-T (Figure 2).

3. Materials

The usefulness of GLM was assessed based on 68 measurements of the nodule abundance in samples collected from the ocean floor using box corers and the results of the analysis of 68 photographs taken at the direct sampling sites. The datasets are derived from the H22 exploration block (4151 km^2) (located in the central-eastern part of the B2 sector), best recognized in the area administered by the IOM (Figure 1).

The data were obtained during two cruises of a research vessel in 2014 and 2019. Both cruises used the same methods of sampling and photographing of the ocean floor and for computer determination of the nodule coverage of the seafloor based on bottom photographs. Therefore, from the point of view of the accuracy of determining the values of the variables, the homogeneity of the initial dataset can be assumed. The box corer sampling covered a 0.25 m^2 (0.5 m × 0.5 m) square section of the seafloor, while 62 photographs covered a bottom section of approximately 1.6 m^2. In 6 out of 68 sampling sites, no bottom photographs were taken before the box corer sample was collected; therefore, the seafloor photographs obtained from photo-profiling (the device Neptun C-M1, Russia [35], covering an area of about 5 m^2, taken approximately 5–50 m from the box corer sampling sites, were used instead.

The statistics of continuous variables, i.e., the nodule abundance based on wet nodule weight (APN symbol, dependent variable) and the percentage of seafloor nodule coverage (NC-S symbol, independent variable), are presented in Table 1, and their empirical distributions in graphical form are presented in Figure 3.

Table 1. Statistics of the nodule abundance (APN) in the box core and the percentage of seafloor nodule coverage (NC-S) determined from the photographs.

Statistics	APN (kg/m^2)	NC-S (%)
Count	68	68
Average	13.47	39.51
Median	13.55	42.0
20% Trimmed mean	13.81	41.16
Standard deviation	4.64	12.65
Coeff. of variation	34.4%	32.0%
Minimum	1.5	7.0
Maximum	23.1	72.0
Range	21.6	65.0
Stnd. skewness	−1.30	−1.89
Stnd. kurtosis	−0.10	0.77
p-value (Shapiro–Wilk test)	0.428	0.029

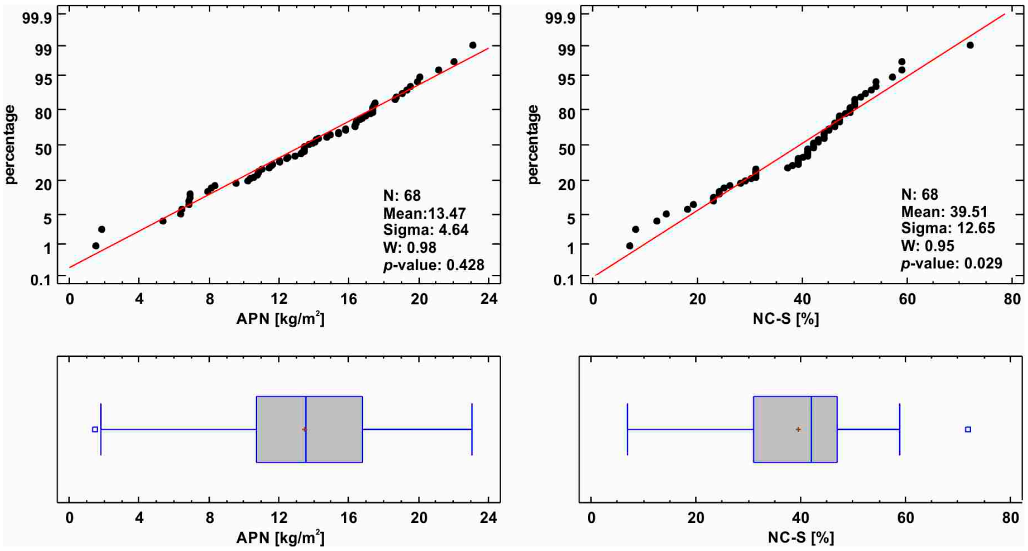


Figure 3. Normal probability plots (above) and box and whisker plots (below) for nodule abundance (APN) and nodule coverage of the bottom (NC-S).

The values of all statistical measures of central tendency (average, median, 20% trimmed mean) both within the APN and NC-S sets differ only slightly (Table 1). The variability of both parameters (APN and NC-S) measured by the coefficient of variation is similar (32–35%) and can be described as moderate. Standardized skewness and kurtosis are within the range expected for the data from a normal distribution (the range is from −2 to 2) [36]. The more precise normality test (Shapiro–Wilk test) did not provide grounds for rejecting the hypothesis of the normality of the distribution of the nodule abundance at the significance level of 0.05. An examination of the empirical distributions of APN and NC-S using the box and whisker method did not show the presence of outliers, which allowed us to consider both datasets as homogeneous (Figure 3).

The other two independent variables used in GLM were categorical (ordinal) and defined as factors with some number of assigned levels. In the case of the sediment coverage of nodules (SC), four levels of this factor were distinguished based on a visual assessment of the seafloor photographs (Figure 4), with numbers from 1 to 4 in the order corresponding to the increasing degree of coverage (low coverage (1), medium coverage (2), high coverage (3), and very high coverage (4)) assigned as identifiers.

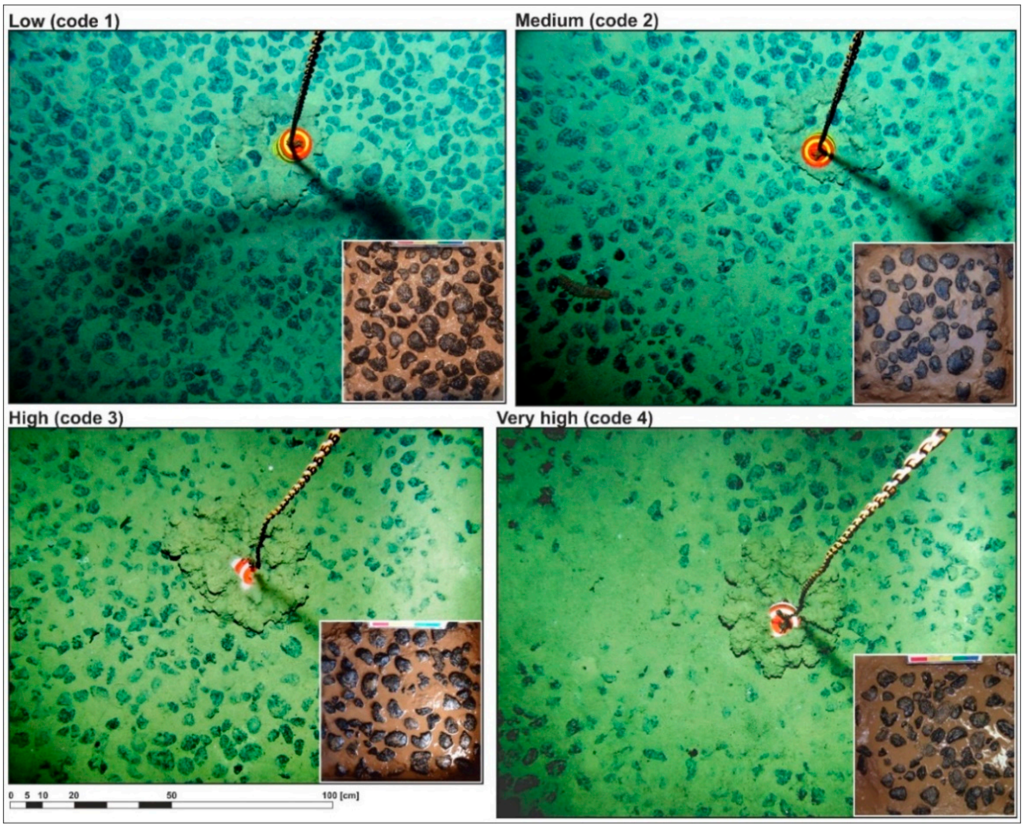


Figure 4. Examples of photographs (seafloor and box corer) from the H22 exploration block showing different levels of sediment coverage of the nodules; the values (codes) of the ordinal variables are given in parentheses.

The nodules occurring in the CCZ are most commonly classified into the three genetic types [37,38]:

- H (hydrogenetic)—small nodules up to 3 cm [37] or 4 cm [38] in diameter, most frequently spheroidal and with smooth surfaces;
- HD (hydrogenetic-diagenetic)—nodules intermediate in size (by convention, from 3 to 6 cm in diameter) with a smooth upper surface and a rough lower surface, predominantly ellipsoidal, flattened, and plate-shaped;

- D (diagenetic)—large nodules, 6–12 cm in diameter, predominantly discoidal and ellipsoidal in shape and with rough surfaces.

Based on the scaled seafloor photographs, it is possible to determine the dominant fractions of nodules, and thus, with high probability, the genetic type of the nodules [28]. In the IOM area, the correctness of the determination of the genetic type of the nodules dominant in the photograph or its part (in the case of only locally increased sediment coverage of nodules) is usually not questionable. With regard to genotypes (GT), three levels of the factor were distinguished based on the nodule fractions dominating in the bottom photographs (Figure 5A): 1-H (hydrogenetic), 2-HD (hydrogenetic-diagenetic), and 3-D (diagenetic). Figure 5B presents the nodule fraction distributions for the different dominant genetic types of nodules with the example of the three box core samples. To illustrate the specificity of the fraction distributions of a given genetic type, mean fraction distributions averaged based on 8 (H), 17 (HD), and 43 (D) of 68 all box core samples in the H22 exploration block were used (Figure 5B). The fraction distributions of nodules for different dominant genetic types usually have characteristic shapes (Figure 5B): strongly skewed to the right (H), moderately skewed to the right (HD), and close to symmetric (D). This factor (GT) related to the differentiation in nodule sizes directly translates into the value of nodule abundance (APN), which is confirmed by a strong positive nonlinear correlation between weight (mass) of the nodules and their surface area [28,29].

Their statistical description was limited to providing the number of observations for individual levels of factors due to the specificity of both ordinal variables (Table 2, Figure 6).

The SC factor levels are dominated by moderate sediment coverage (level 2), which constitutes 50% of all observations, while the GT factor levels are dominated by the diagenetic type D (level 3), slightly exceeding 63% (Figure 6, Table 2).

Table 2. Frequency and relative frequency for different levels of factors (ordinal variables): genetic type of nodules (GT), sediment coverage (SC).

Ordinal Variable (Factor)	Level of Factor	Code	Frequency	Relative Frequency
Sediment coverage (SC)	Low	1	24	35.3%
	Medium	2	34	50.0%
	High	3	6	8.8%
	Very high	4	4	5.9%
Genetic type of nodules (GT)	H	1	8	11.8%
	HD	2	17	25.0%
	D	3	43	63.2%

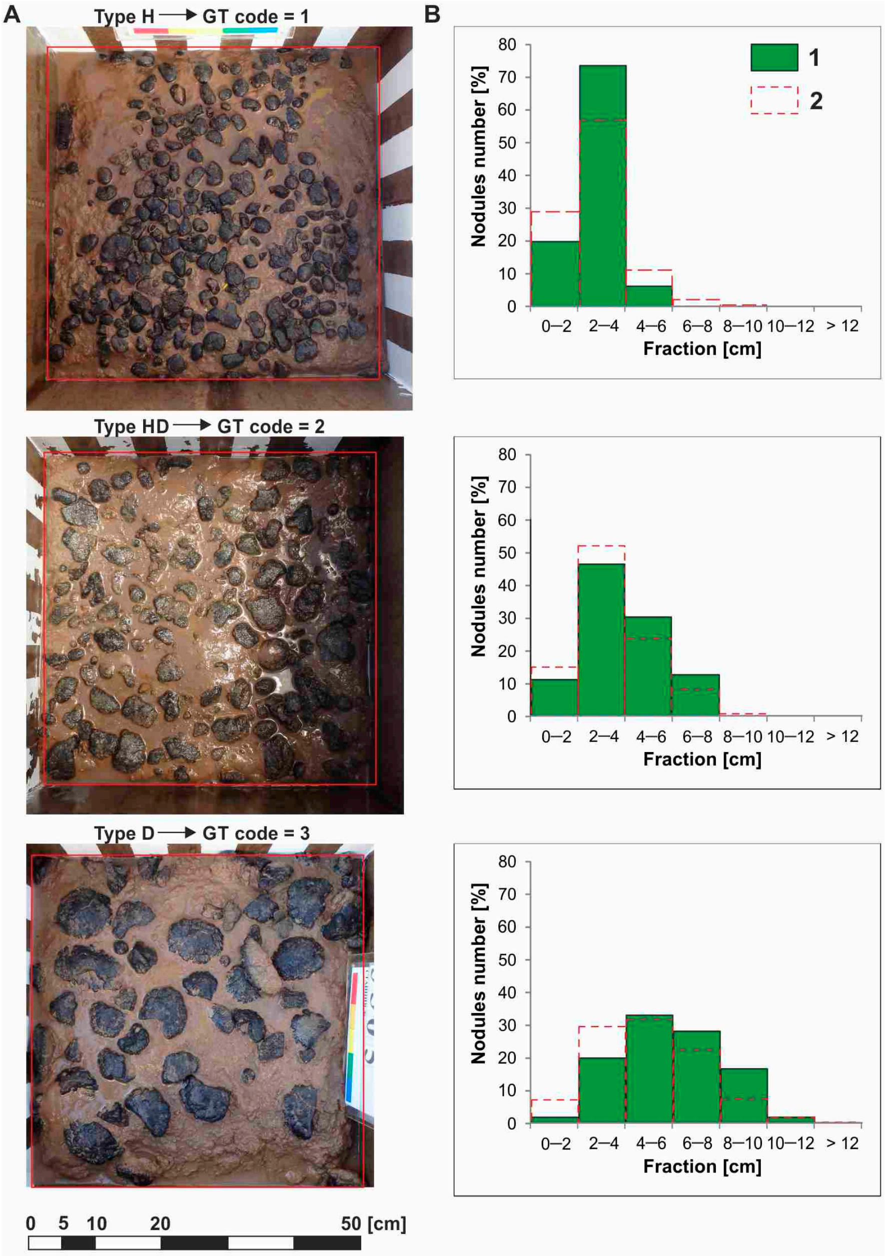


Figure 5. Examples of box corer photographs for three genetic types of nodules (**A**); fraction distributions of the nodules defined based on the presented (on the left) box core samples (1—green bars) and the mean number of nodules in the fractions based on the box core samples from H22 exploration block (2—dashed red line) (**B**).

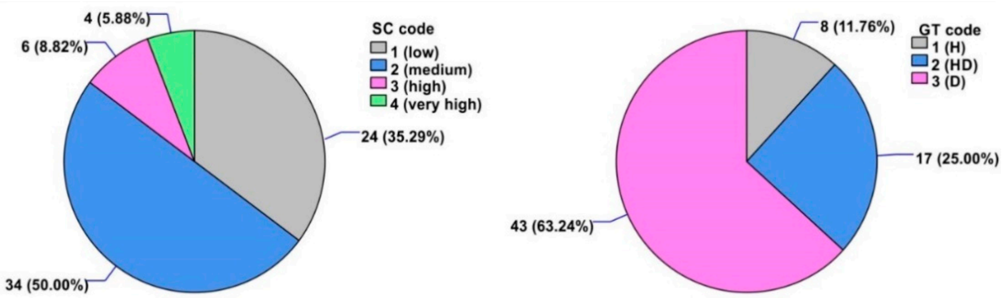


Figure 6. Pie charts for sediment coverage (SC) and genetic type of nodules (GT).

4. Methods

The general linear models (GLM) procedure is used to construct statistical model describing the impact of any set of explanatory variables (X), quantitative (continuous) or qualitative (categorical), on one or more dependent variables (Y). An independent categorical variable (nominal or ordinal) is called a factor, and its categories are called the levels of the factor [39].

In its simplest form, GLM determines the (linear) relationship between the one dependent (response) variable Y and the set of predictors (explanatory variables) X_i and is expressed by the general formula:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k \tag{1}$$

where b_0—intercept, b_{1-k}—partial regression coefficients.

This particular case of a general linear model limited (restricted) to one dependent variable was used in the research. In terms of the obtained results, the method is equivalent to the multiple regression model.

For use in regression, a categorical variable with k categories (levels) must be transformed (coded) into a set of (k–1) indicator variables also known as a dummy variable [40]. An example of such a transformation is given in the caption of the Table 3.

Table 3. Regression models between nodule abundance (APN) and seafloor nodule coverage (NC-S) or grid nodule coverage (NC-T) supported by the categorical variables: level of sediment coverage of nodules (SC) and genetic type of nodules (GT) based on photographs (count of data = 68).

Data Set (Count of Data = 68)	Regression Method	Independent Variables	Equation of Estimated Model	R^2_{adj} *p*-Value	SEE	MAE	MPE	MAPE
Grid photographs	SLM	NC-T	APN[kg/m^2] = 1.33 + 0.27 NC-T[%]	59.2 (0.0130)	2.96	2.23	−7.3	20.9
		ln(NC-T)	APN[kg/m^2] = −14.17 + 7.38 ln(NC-T[%])	52.6 (0.0000)	3.19	2.58	−2.0	28.0
	GLM	NC-T, GT	APN[kg/m^2] = 0.57 − 2.70I1(1) − 0.42I1(2) + 0.25NC-T[%]	80.7 (0.0000)	2.04	1.53	−6.8	17.3
		ln(NC-T), GT	APN[kg/m^2] = −15.82 − 3.20I1(1) − 0.40I1(2) + 7.34ln(NC-T[%])	81.7 (0.0000)	1.98	1.56	0.1	14.8
Seafloor photographs	SLM	NC-S	APN[kg/m^2] = 7.55 + 0.15 NC-S	15.4 (0.0000)	4.27	3.56	−21.4	41.7
		ln(NC-S)	APN[kg/m^2] = −5.46 + 5.25ln(NC-S[%])	23.3 (0.0000)	4.06	3.39	−15.5	34.9
	GLM	NC-S, GT	APN[kg/m^2] = 2.65 − 4.16I2(1) − 0.48I2(2) + 0.22NC-S[%]	60.2 (0.0000)	2.92	2.26	−14.6	29.8
		NC-S, GT, SC	APN[kg/m^2] = 0.95 − 1.73I1(1) − 0.01I1(2) + 0.02I1(3) − 4.30I2(1) − 0.46I2(2) + 0.27NC-S[%]	61.0 (0.0000)	2.90	2.14	−13.6	28.0
		ln(NC-S), GT	APN[kg/m^2] = −13.25 − 3.82I2(1) − 0.73I2(2) + 6.79ln(NC-S[%])	67.4 (0.0000)	2.65	2.05	−8.4	22.3
		ln(NC-S), GT, SC	APN[kg/m^2] = −20.02 − 2.10I1(1) − 0.60I1(2) − 0.83I1(3) − 4.10I2(1) − 0.61I2(2) + 8.90ln(NC-S[%])	70.4 (0.0000)	2.52	1.88	−6.2	18.8

Explanations: SLM—simple linear regression model, GLM—general linear models, R^2_{adj}—adjusted coefficient of determination, SEE—standard error of estimation, MAE—mean absolute error, MPE—mean percentage error, MAPE—mean absolute percentage error. The values of indicator variables in GLM regression were determined based on the values of categorical (ordinal) variables according to the following scheme: I1(1) = 1 if SC = 1, −1 if SC = 4, 0 otherwise, I1(2) = 1 if SC = 2, −1 if SC = 4, 0 otherwise, I1(3) = 1 if SC = 3, −1 if SC = 4, 0 otherwise, I2(1) = 1 if GT = 1, −1 if GT = 3, 0 otherwise, I2(2) = 1 if GT = 2, −1 if GT = 3, 0 otherwise.

Each of the determined regression models was tested for its statistical significance by calculating the so-called *p*-value. When the *p*-value $\leq$0.05, the model can be considered statistically significant with a risk of error not greater than 5%.

In addition, five measurements were determined to indicate the model's goodness of fit (the strength of relationships):

- The adjusted coefficient of determination $\mathrm{R}^2_{\mathrm{adj}}$ expresses the percentage of the variability in the dependent variable, which has been explained by the fitted model, ranging from 0% (lack of the dependency) to 100% (ideal, full relationship), adjusted for the number of coefficients in the model:

$$\mathrm{R}^2_{\mathrm{adj}} = \left[1 - \left(\frac{\mathrm{n}-1}{\mathrm{n}-\mathrm{p}}\right) \times \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}\right] \times 100\% \tag{2}$$

where n—count of data, p—the number of estimated model coefficients, $\hat{y}_i$—theoretical value of the dependent variable Y determined from the model equation for the observation "*i*", y_i—empirical value of the dependent variable Y for the observations "*i*", $\overline{y}$—arithmetic mean of the empirical values of the dependent variable Y.

- The standard (prediction) error of estimation (SEE) characterizing the average scatter of the measured values of the dependent variable in the regression model:

$$\mathrm{SEE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - \mathrm{y_i})^2}{n}} \tag{3}$$

- The mean absolute error (MAE) characterizing the mean absolute deviation of the measured Y values from the values indicated by the model:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{4}$$

- Mean percentage error (MPE):

$$\mathrm{MPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{\hat{y}_i - y_i}{y_i} \tag{5}$$

- Mean absolute percentage error (MAPE):

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{6}$$

For comparison purposes, simple linear models (SLM) of the general form:

$$\mathrm{Y} = \mathrm{b}_0 + \mathrm{b}_1 \cdot \mathrm{X}_1, \tag{7}$$

linking the nodule abundance (Y) to the percentage of seafloor nodule coverage (NC-S) and box corer (NC-T) and their natural logarithms, were also analyzed.

Logarithmically transforming variables in a regression model is a very common way of handling situations where a non-linear relationship exists between the independent and dependent variables [41]. The use of the natural logarithm in the analyzed cases resulted from preliminary studies which showed a slightly stronger correlation of APN with ln(NC-S) than with (NC-S) (Figure 7).

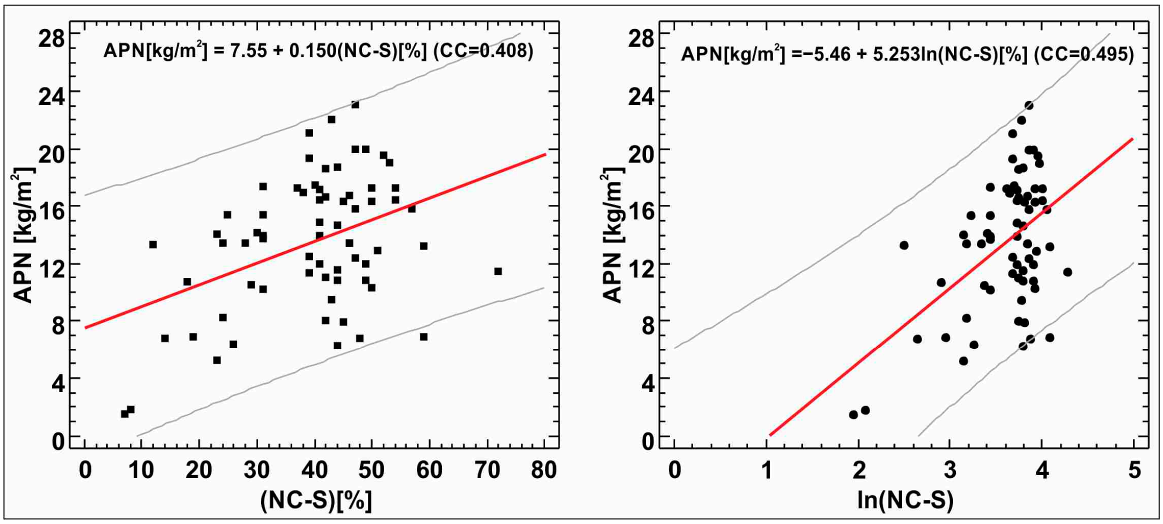


Figure 7. Simple linear relationship between APN and (NC-S) (**left**) and APN and ln (NC-S) (**right**); CC—correlation coefficient.

The determination of simple linear models allowed a preliminary assessment of the improvement in the accuracy of APN prediction using GLM. The quality of all models was verified using the cross-validation method for a randomly selected training and test data subsets.

5. Results and Discussion

In accordance with the adopted methodology for APN prediction, the first step was to develop GLM equations, which, for the variables selected for the study, have the following form:

$$\text{APN} = b_0 + b_1 \times (\text{NC-S}) + b_2 \cdot (\text{SC}) + b_3 \cdot (\text{GT}) \tag{8}$$

and

$$\text{APN} = b_0 + b_1 \cdot \ln(\text{NC-S}) + b_2 \cdot (\text{SC}) + b_3 \cdot (\text{GT}) \tag{9}$$

where b_0—intercept, $b_{1\text{–}3}$—partial regression coefficients, NC—seafloor nodule coverage (continuous variable) (%), SC—level of sediment coverage of nodules (ordinal variable), GT—genetic type of nodules (ordinal variable); NC-S, SC, and GT values were determined based on photographs of the bottom taken in the place (or near) of the box core sample sites.

The continuous and ordinal variables used in the GLM analysis along with the adopted levels are shown schematically in Figure 2.

For the analyzed variables, the equations of simple linear models (SLM) are as follows:

$$\text{APN} = b_0 + b_1 \cdot (\text{NC-S}) \tag{10}$$

and

$$\text{APN} = b_0 + b_1 \cdot \ln\,(\text{NC-S}) \tag{11}$$

where b_0—intercept, b_1—slope, NC-S—seafloor nodule coverage (continuous variable) [%].

For comparative purposes, identical variants of regression models were determined for the data obtained from the grid photographs (e.g., Figure 5):

- GLM:

$$\text{APN} = b_0 + b_1 \cdot (\text{NC-T}) + b_2 \cdot (\text{GT}) \tag{12}$$

and

$$\text{APN} = b_0 + b_1 \cdot \ln(\text{NC-T}) + b_2 \cdot (\text{GT}) \tag{13}$$

- SLM:

$$APN = b_0 + b_1 \cdot (NC\text{-}T) \tag{14}$$

and

$$APN = b_0 + b_1 \cdot \ln(NC\text{-}T) \tag{15}$$

where NC-T—nodule coverage of the grid.

The analysis of the results contained in Table 3 allows us to make a number of observations about the H22 IOM exploration block, which are presented below.

The use of GLM in place of the simple linear regression model (SLM), both for grid and seafloor photographic data, in all cases leads to a significant increase in the accuracy of the APN prediction, as evidenced by the increased values of R^2_{adj} for the grid and seafloor photographs by over 20% and approx. 50%, respectively, and reduced SEE, MAE, MPE, and MAPE values (e.g., SEE by about 1.0 kg/m^2 for the grid photographs and about 1.7 kg/m^2 for the seafloor photographs). The improvement in modeling quality is also visually confirmed by the plots of empirical and theoretical relationships shown in Figure 8.

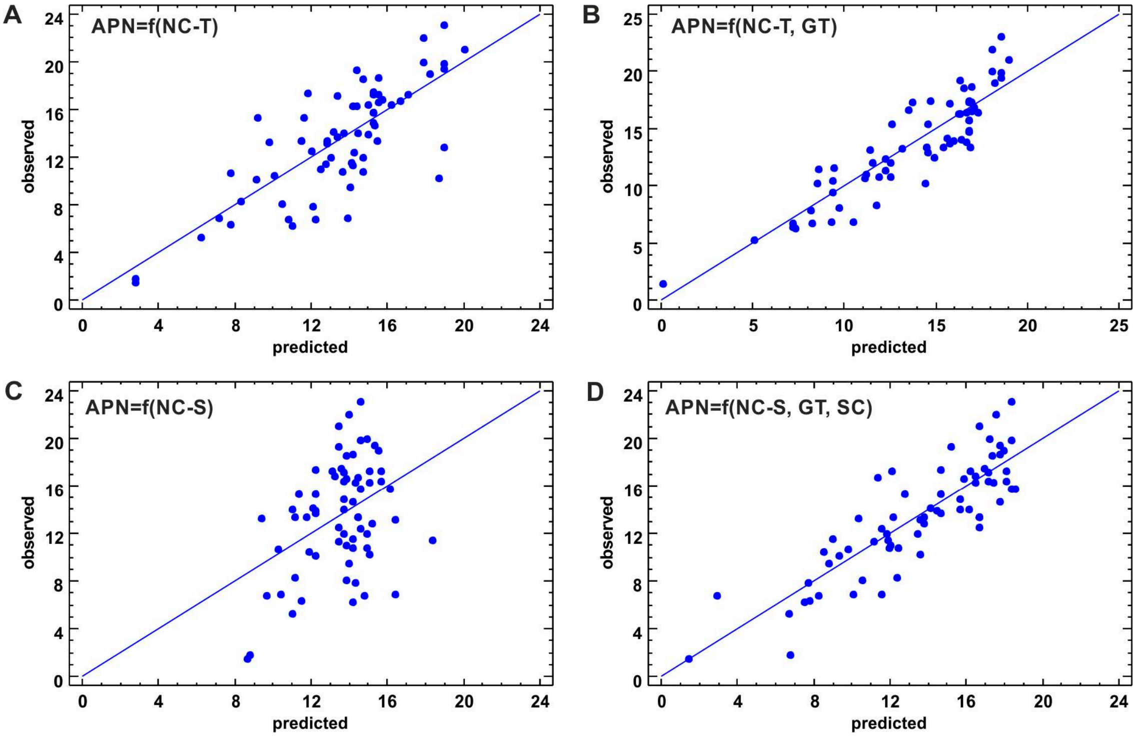


Figure 8. Scatter plots between predicted and observed APN; predicted APN was calculated based on grid photographs from the SLM (**A**) and GLM regression model (**B**) and based on seafloor photographs from the SLM (**C**) and GLM regression model (**D**).

The linear relationship between APN and NC-T (or ln(NC-T)) is much stronger (with R^2_{adj} of 50–60%) than between APN and NC-S (or ln(NC-S)) (with R^2_{adj} of 15–25%). This can be explained, first of all, by the lack of sediment covering the nodules in the box corer after drainage and the identical surface area for which the abundance (APN) and nodules coverage (NC-T) are determined (Figure 5).

The prediction of nodule abundance (APN) based on GLM with a high value of R^2_{adj} = 61% (and 70% for ln(NC-S)) is associated with SEE values of 2.7 and 2.5 kg/m^2, respectively, and MAE values of 2.1 and 1.9 kg/m^2. These values, related to the average nodule abundance (13.5 kg/m^2), represent 20.0% (and 18.5% for ln(NC-S)) for SEE and 15.6% (and 14% for ln(NC-S)) for MAE.

Replacing NC-S with its natural logarithm clearly improves the quality of approximation of the empirical relationship with regression models, characterized by an increase in

R^2_{adj} by about 8–10% and accompanied by a corresponding decrease in SEE and MAE. The opposite effect is observed for NC-T since the use of the natural logarithm NC-T in simple linear model results in a decrease in R^2_{adj} by about 7% and a corresponding increase in SEE and MAE.

Surprisingly, the addition of the ordinal variable SC to the GLM does not result in a significant increase in the accuracy of the APN prediction, which is confirmed by overlapping confidence intervals for APN determined for particular levels of this factor (Figure 9). It proves that this factor has little influence on the accuracy of determining the APN from the model. It is supposed that this due to a strong negative statistically significant correlation between the sediment coverage of nodules (SC) and the seafloor nodule coverage (NC-S) with Kendall's rank correlation coefficient of −0.57 (Figure 10).

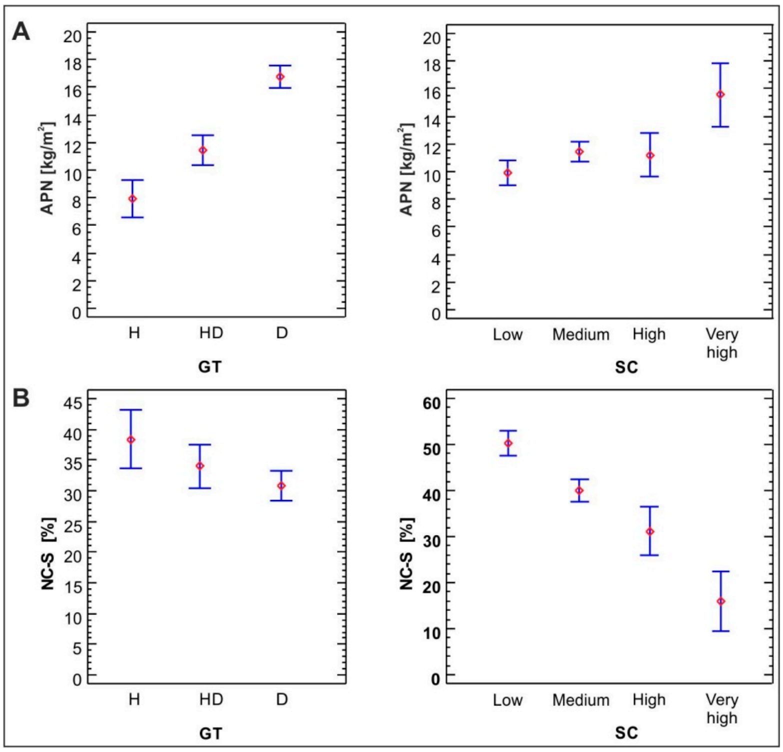
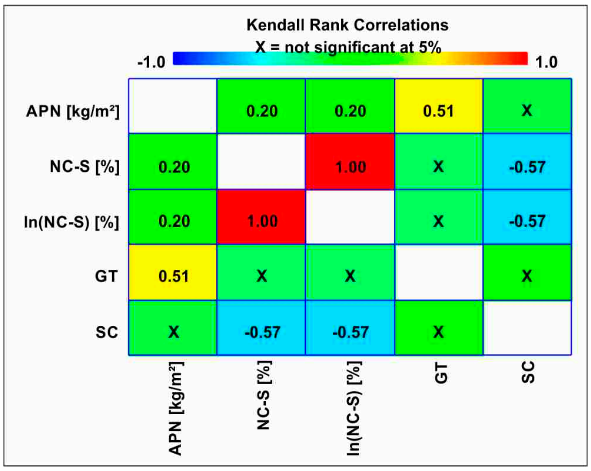


Figure 9. LSD (least significant difference) type confidence intervals of the nodule abundance (APN) (**A**) and seafloor nodule coverage (**B**) for different levels of factors: GT (**left**) and SC (**right**).


Kendall Rank Correlations
X = not significant at 5%
-1.0
1.0

	APN [kg/m²]	NC-S [%]	ln(NC-S) [%]	GT	SC
APN [kg/m²]		0.20	0.20	0.51	X
NC-S [%]	0.20		1.00	X	-0.57
ln(NC-S) [%]	0.20	1.00		X	-0.57
GT	0.51	X	X		X
SC	X	-0.57	-0.57	X	

Figure 10. Kendall's rank correlation coefficients for pairs of analyzed variables.

Theoretical measures of the accuracy of the approximation of empirical dependence by the applied models (GLM, SLM) are not fully sufficient to definitively confirm or reject their usefulness for APN prediction.

To strengthen the conclusions about the effectiveness of the obtained regression models, the cross-validation procedure was applied to independent data subsets. Although the dataset (of 68 observations) is large enough to determine a reliable form of the APN dependency model on the analyzed factors, it is not large enough to reliably verify the quality of models for independent subsets of data distinguished within it. Nevertheless, such verification was performed by randomly selecting three subsets of 48 observations from the basic set, which were treated as training sets and for which the forms of three regression models (SLM, GLM, and GLM with ln(NC-S)) were determined. These models were used to estimate the nodule abundance in the remaining data subsets, consisting of 20 observations, treated as test sets (Figure 1C).

The verification of the quality of the models for the training sets and their usefulness for predicting the abundance of nodules in the test sets consisted of the following:

- Determining the statistical significance of the linear relationship between the nodule abundance predicted from the models (for the training data) with the real nodule abundance in the test sets (using *p*-value) and the strength of this relationship using the adjusted coefficient of determination R^2_{adj});
- Determination of the arithmetic mean (MD) and mean absolute difference (MAD) between the nodule abundance predicted from the model and found in the test sets.

The results of the validation of the training models presented in Table 4 fully confirm the previous observations made for the complete initial data set (Table 3).

Table 4. Estimation errors of the nodule abundance in three test subsets (of 20 observations each) based on three regression models determined from the training data subsets of 48 observations each); the location of training and test subsets is shown in Figure 1C.

Model	Parameter	Test Subset 1 ($\overline{APN}$=13.63 [kg/m^2])	Test Subset 2 ($\overline{APN}$=13.14 [kg/m^2])	Test Subset 3 ($\overline{APN}$=14.65 [kg/m^2])
SLM *Simple linear model* APN = f(NC-S)	*p*-value	0.0146	0.2182	0.0497
	R^2_{adj}	24.9%	3.2%	15.3%
	MD	−0.23 (−1.7%)	0.49 (3.8%)	−1.50 (−10.2%)
	MAD	3.49 (25.6%)	3.57 (27.2%)	3.67 (25.1%)
GLM *General linear model* APN = f(NCS, GT, SC)	*p*-value	0.0000	0.0000	0.0009
	R^2_{adj}	68.0%	65.7%	43.7%
	MD	0.73 (5.4%)	0.39 (3.0%)	−1.01 (−6.9%)
	MAD	2.22 (16.3%)	2.03 (15.5%)	2.71 (18.5%)

GLM(ln(NC-S))
General linear model with ln(NC-S)
APN = f(ln(NC-S), GT, SC)
	p-value	0.0000	0.0000	0.000

	R^2_{adj}	77.9%	65.9%	59.6%
	MD	0.11 (0.8%)	0.37 (2.8%)	−1.05 (−7.2%)
	MAD	1.86 (13.6%)	2.06 (15.7%)	2.40 (16.4%)

Explanations: $\overline{APN}$—arithmetic mean of the nodule abundance in the test subset, R^2_{adj}—the adjusted coefficient of determination, MD—mean difference, MAD—mean absolute difference, NC-S—seafloor nodule coverage, GT—genetic type of nodule, SC—sediment coverage.

The linear relationships of the nodule abundance (estimated from the training models and found in the test data subsets) are highly statistically significant for GLM with a *p*-value of <0.001, while for SLM they are statistically significant only in two cases but at a significance level of $\propto$ = 0.05 (test subsets 1 and 3), and in one case there is no basis to reject the hypothesis that there is no such relationship (test subset 2).

The coefficients of the determination R^2_{adj} of linear relationships of the nodule abundance (both found and estimated using SLM) are many times lower than those determined using GLM.

The use of a more advanced general linear model (GLM) instead of a simple linear model (SLM) to predict the nodule abundance in the test subsets leads to a significant reduction in the random prediction error represented by the MAD value. These values, expressed as a percentage of the average nodule abundance in the test subsets, range from 25% to 28% for SLM, through 15% to 19% for GLM and 13% to 17% for GLM (with ln(NC-S)). With one exception, the MD values, used as a measure of systematic prediction error for GLM models, are also lower than for SLM.

Despite the limitations of the validation method used, related to the small number of test sets and partial overlapping of data in three variants of both training sets and test sets (Figure 1C), the obtained results are unequivocal and confirm the usefulness of using GLM to predict nodule abundance with the use of ordinal variables and in particular GT, indirectly characterizing the nodule fraction distribution.

6. Conclusions

The results of using general linear models (GLM) to predict the nodule abundance based on seafloor photographs of the H22 exploration block (IOM) can be considered promising in terms of increasing the accuracy of prediction. The advantage of GLM is the possibility of including both quantitative continuous variables (seafloor nodule coverage) in addition to ordinal variables (the dominant size of nodules related to their genetic type, the degree of sediment coverage of nodule) in the regression model. All these variables can be quantified, albeit with a different accuracy, from seafloor photographs. The nodule coverage of the ocean floor (visible in photographs with areas ranging from 1.5 m^2 to 5 m^2, depending on the technique used and the conditions of photographic recording) estimated with the use of computer programs is subject to error resulting mainly from at least partial nodule coverage with sediments, with which it is negatively and strongly correlated. Determining the values of the two qualitative variables on an ordinal scale (GT and SC) requires some experience with the visual assessment and the analysis of photographs. This approximate visual assessment, however, seems to be sufficient to significantly increase the reliability of the prediction of the nodule abundance. Compared to the simple linear model linking the nodule abundance found in the box corer with the seafloor nodule coverage estimated based on the photographs, the use of GLM leads to a significant increase in the accuracy of the nodule abundance estimates. For the analyzed data set, it is expressed by a significant increase in the adjusted coefficient of determination (from 15.4% to 70.4%) and by a significant reduction in the dispersion measures around both models: for SEE from 4.2 kg/m^2 to 2.5 kg/m^2 and for MAE from 3.6 kg/m^2 to 1.9 kg/m^2.

The accuracies obtained using GLM may seem not entirely satisfactory, but one should take into account factors affecting modeling quality resulting from differences in the horizontal surface of the box corer and the ocean floor covered by the photograph, the local variability of nodule abundance, and their sediment coverage and uneven quality of photographs resulting from a variable distance from the seafloor and its uneven illumination. The obtained results require verification on a larger dataset due to the limited dataset from the fragment of the nodule-bearing area in the Pacific administered by the IOM.

The use of a different number of levels of sediment coverage of nodules and photographic patterns facilitating the appropriate categorization of this factor should also be considered.

The final confirmation of the usefulness of GLM for the prediction of nodule abundance on the basis of data obtained from the photographic survey of the seafloor will enable the geostatistical estimation of nodule resources, e.g., with the use of kriging with the variance of measurement errors [42] integrating the measurements made with different accuracies: higher in the case of the box corer data and lower in the case of the data from photographs.

Author Contributions: Conceptualization, M.W.-B. and J.M.; methodology, M.W.-B. and J.M.; software, M.W.-B.; writing—original draft preparation, M.W.-B.; writing—review and editing, J.M.; visualization, M.W.-B. Both authors have read and agreed to the published version of the manuscript.

Funding: This research was financed from the AGH University of Science and Technology grant no. 16.16.140.315 in 2021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from the INTEROCEANMETAL Joint Organization (https://iom.gov.pl/) and are available from the authors with the permission of IOM.

Acknowledgments: The authors thank Tomasz Abramowski, General Director of the INTEROCEANMETAL Joint Organization, for access to data and kind permission for data processing and publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hein, J.R.; Mizell, K.; Koschinsky, A.; Conrad, T.A. Deep-ocean mineral deposits as a source of critical metals for high- and green-technology applications: Comparison with land-based resources. *Ore Geol. Rev.* **2013**, *51*, 1–14. [CrossRef]
2. Petersen, S.; Krätschell, A.; Augustin, N.; Jamieson, J.; Hein, J.R.; Hannington, M.D. News from the seabed—Geological characteristics and resource potential of deep-sea mineral resources. *Mar. Policy* **2016**, *70*, 175–187. [CrossRef]
3. Milinovic, J.; Rodrigues, F.J.L.; Barriga, F.J.A.S.; Murton, B.J. Ocean-floor sediments as a resource of rare earth elements: An overview of recently studied sites. *Minerals* **2021**, *11*, 142. [CrossRef]
4. Toro, N.; Robles, P.; Jeldres, R.I. Seabed mineral resources, an alternative for the future of renewable energy: A critical review. *Ore Geol. Rev.* **2020**, *126*, 103699. [CrossRef]
5. Toro, N.; Jeldres, R.I.; Órdenes, J.A.; Robles, P.; Navarra, A. Manganese nodules in chile, an alternative for the production of co and mn in the future—A review. *Minerals* **2020**, *10*, 674. [CrossRef]
6. Watzel, R.; Rühlemann, C.; Vink, A. Mining mineral resources from the seabed: Opportunities and challenges. *Mar. Policy* **2020**, *114*, 103828. [CrossRef]
7. Hein, J.R.; Koschinsky, A.; Kuhn, T. Deep-ocean polymetallic nodules as a resource for critical materials. *Nat. Rev. Earth Environ.* **2020**, *1*, 158–169. [CrossRef]
8. Abramowski, T.; Stoyanova, V. Deep-Sea Polymetallic Nodules: Renewed Interest as Resources for Environmentally Sustainable Development. In Proceedings of the SGEM2012 Conference, Albena, Bulgaria, 17–23 June 2012; Volume 1, pp. 515–522. [CrossRef]
9. Pérez, K.; Villegas, Á.; Saldaña, M.; Jeldres, R.I.; González, J.; Toro, N. Initial investigation into the leaching of manganese from nodules at room temperature with the use of sulfuric acid and the addition of foundry slag—Part II. *Sep. Sci. Technol.* **2021**, *56*, 389–394. [CrossRef]
10. Toro, N.; Saldaña, M.; Castillo, J.; Higuera, F.; Acosta, R. Leaching of manganese from marine nodules at room temperature with the use of sulfuric acid and the addition of tailings. *Minerals* **2019**, *9*, 289. [CrossRef]
11. Usui, A.; Hino, H.; Suzushima, D.; Tomioka, N.; Suzuki, Y.; Sunamura, M.; Kato, S.; Kashiwabara, T.; Kikuchi, S.; Uramoto, G.-I.; et al. Modern precipitation of hydrogenetic ferromanganese minerals during on-site 15-year exposure tests. *Sci. Rep.* **2020**, *10*, 1–10. [CrossRef]
12. Usui, A.; Nishi, K.; Sato, H.; Nakasato, Y.; Thornton, B.; Kashiwabara, T.; Tokumaru, A.; Sakaguchi, A.; Yamaoka, K.; Kato, S.; et al. Continuous growth of hydrogenetic ferromanganese crusts since 17 Myr ago on Takuyo-Daigo Seamount, NW Pacific, at water depths of 800–5500 m. *Ore Geol. Rev.* **2017**, *87*, 71–87. [CrossRef]
13. Zawadzki, D.; Maciąg, Ł.; Abramowski, T.; McCartney, K. Fractionation trends and variability of rare earth elements and selected critical metals in pelagic sediment from abyssal basin of NE Pacific (clarion-clipperton fracture zone). *Minerals* **2020**, *10*, 320. [CrossRef]
14. Abramowski, T.; Urbanek, M.; Baláž, P. Structural Economic assessment of polymetallic nodules mining project with updates to present market conditions. *Minerals* **2021**, *11*, 311. [CrossRef]
15. Parianos, J.; Lipton, I.; Nimmo, M. Aspects of estimation and reporting of mineral resources of seabed polymetallic nodules: A contemporaneous case study. *Minerals* **2021**, *11*, 200. [CrossRef]
16. Sharma, R. Deep-sea mining: Current status and future considerations. In *Deep-Sea Mining: Resource Potential, Technical and Environmental Considerations*; Sharma, R., Ed.; Springer Science and Business Media LLC: Berlin, Germany, 2017; pp. 3–21.
17. Abramovski, T.; Stefanova, V.P.; Causse, R.; Romanchuk, A. Technologies for the processing of polymetal-lic nodules from clarion clipperton zone in the Pacific Ocean. *J. Chem. Technol. Metal.* **2017**, *52*, 258–269.
18. Vu, N.H.; Kristianová, E.; Dvořák, P.; Abramowski, T.; Dreiseitl, I.; Adrysheva, A. Modified leach residues from processing deep-sea nodules as effective heavy metals adsorbents. *Metals* **2019**, *9*, 472. [CrossRef]

19. Clarion-Clipperton Fracture Zone Exploration Areas for Polymetallic Nodules (Interoceanmetal Joint Organization). Available online: https://www.isa.org.jm/map/interoceanmetal-joint-organization (accessed on 31 October 2020).
20. Mucha, J.; Wasilewska-Błaszczyk, M. Estimation accuracy and classification of polymetallic nodule resources based on classical sampling supported by seafloor photography (Pacific Ocean, Clarion-Clipperton Fracture Zone, IOM Area). *Minerals* **2020**, *10*, 263. [CrossRef]
21. Sterk, R.; Stein, J.K. Seabed mineral deposits: An overview of sampling techniques and future developments. In Proceedings of the Deep Sea Mining Summit, Aberdeen, Scotland, 9–10 February 2015; p. 29.
22. Schoening, T.; Jones, D.O.B.; Greinert, J. Compact-Morphology-based poly-metallic Nodule Delineation. *Sci. Rep.* **2017**, *7*, 1–12. [CrossRef] [PubMed]
23. Felix, D. Some problems in making nodule abundance estimates from seafloor photographs. *Mar. Min.* **1980**, *2*, 293–302.
24. Handa, K.; Tsurusaki, K. Manganese nodules: Relationship between Coverage and Abundance in the Northern Part of Central Pacific Basin. *Geol. Surv. Jpn.* **1981**, *15*, 184–217.
25. Lipton, I.; Nimmo, M.; Stevenson, I. *NORI Area D Clarion Clipperton Zone Mineral Resource Estimate. Deep Green Metals Inc. Pacific Ocean*; AMC Project 318010; AMC Consultants Pty Ltd.: Perth, WA, Australia, 2019.
26. Jie, W.L.; Kalyan, B.; Chitre, M.; Vishnu, H. Polymetallic nodules abundance estimation using sidescan sonar: A quantitative approach using artificial neural network. *OCEANS 2017—Aberdeen* **2017**, 1–6. [CrossRef]
27. Yoo, C.M.; Joo, J.; Lee, S.H.; Ko, Y.; Chi, S.-B.; Kim, H.J.; Seo, I.; Hyeong, K. Resource assessment of polymetallic nodules using acoustic backscatter intensity data from the Korean exploration Area, Northeastern Equatorial Pacific. *Ocean Sci. J.* **2018**, *53*, 381–394. [CrossRef]
28. Wasilewska-Błaszczyk, M.; Mucha, J. Possibilities and limitations of the use of seafloor photographs for estimating polymetallic nodule resources—Case study from IOM Area, Pacific Ocean. *Minerals* **2020**, *10*, 1123. [CrossRef]
29. Kuhn, T.; Rathke, M. *Report on Visual Data Acquisition in the Field and Interpretation for SMnN*; Blue Mining Project; Blue Mining Deliverable D1.31; European Commission Seventh Framework Programme, 2017; p. 34. Available online: https://bluemining.eu/download/project_results/public_reports/BLUE-MINING-D1.31b-Final-Report-on-visual-data-acquisition-in-the-field-and-interpretation-for-SMnN.pdf (accessed on 12 March 2021).
30. Jung, M.Y.; Kim, I.K.; Kang, J.K. Analysis of manganese nodule abundance in KODOS Area. *Econ. Environ. Geol.* **1995**, *28*, 429–437.
31. Park, C.-Y.; Park, S.-H.; Kim, C.-W.; Kang, J.-K.; Kim, K.-H. An image analysis technique for exploration of manganese nodules. *Mar. Georesour. Geotechnol.* **1999**, *17*, 371–386. [CrossRef]
32. Tsune, A. Effects of size distribution of deep-sea polymetallic nodules on the estimation of abundance obtained from seafloor photographs using conventional formulae. In Proceedings of the Eleventh OceanMining and Gas Hydrates Symposium, Big Island, HI, USA, 21–27 June 2015; International Society ofOffshore and Polar Engineers: Kona, HI, USA, 2015; p. 7.
33. Sharma, R.; Khadge, N.; Sankar, S.J. Assessing the distribution and abundance of seabed minerals from seafloor photographic data in the Central Indian Ocean Basin. *Int. J. Remote. Sens.* **2012**, *34*, 1691–1706. [CrossRef]
34. Balaz, P.; Krawcewicz, A.; Abramowski, T. *30 Years of Deep Seabed Exploration*; Interoceanmetal Joint Organization: Szczecin, Poland, 2017.
35. Dreiseitl, I. Deep Sea Exploration for Metal Reserves—Objectives, Methods and Look into the Future. In *Deep See Mining Value Chain: Organization, Technology and Development*; Abramowski, T., Ed.; Interoceanmetal Join Organization: Szczecin, Poland, 2016; pp. 105–117.
36. *Statgraphics*; (Version Centurion XVII); Statpoint Technologies, Inc.: Warrenton, VA, USA.
37. Kotliński, R. Relationships between Nodule Genesis and Topography in the Eastern Area of the C-C Region. In *Establishment of A Geological Model of Polymetallic Nodule Deposits in the Clarion-Clipperton Fracture Zone of The Equatorial North Pacific Ocean. Proceedings of the International Seabed Authority's Workshop, Nadi, Fiji, 13–20 May 2003*; International Seabed Authority: Kingston, Jamaica, 2009; pp. 203–221.
38. *A Geological Model of Polymetallic Nodule Deposits in the Clarion-Clipperton Fracture Zone*; ISA Technical Study; Technical Study: No. 6; ISA, International Seabed Authority: Kingston, Jamaica, 2010; p. 211.
39. Dobson, A.J.; Barnett, A.G. *An Introduction to Generalized Linear Models*; CRC Press: Boca Raton, FL, USA, 2008.
40. LaRose, D.T. *Data Mining Methods and Models*; Wiley: Hoboken, NJ, USA, 2005.
41. Benoit, K. Linear Regression Models with Logarithmic Transformations. Available online: http://www.kenbenoit.net/courses/ME104/logmodels2.pdf (accessed on 15 June 2020).
42. Chilès, J.-P.; Delfiner, P. *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed.; John Wiley & Sons: Noboken, NJ, USA, 2009.