SM2_material **– Description of material**

In the drilling process of sites U1480 and U1481, the collected rock samples were subjected to techniques to capture the geophysical properties Gamma Ray Attenuation bulk density (GRA), Moisture and Density (MAD), Magnetic Susceptibility (MS), Natural Gamma Radiation (NGR), P-Wave Velocity Logger System (PWL), Red Green Blue channels (RGB) and Reflectance Spectrophotometry and Colorimetry (RSC).

GRA is a geophysical property acquired by the sensor Whole-Round Multisensor Logger (WRMSL) using a Cs-137 collimated source and NaI scintillation detector, whose values are calibrated using water standard and aluminum standard (BLUM, 1997; MCNEILL et al., 2017). In this article, GRA has a unit of measurement in g/cm^3 in a single feature used.

MAD is the geophysical property that measures the relationship between mass and volume. Core samples of approximately 8 cm² are processed by combining 4 methods (A, B, C, D). The wet mass, by combining methods A, B, C and dry mass (A, B, C, D) is determined by the use of movement signals by the Mettler-Toledo balance (BLUM, 1997; MCNEILL et al., 2017. In this article, MAD has a measurement unit in $wt\%$, g/cm, and $vol\%$ with 6 features used.

MS is the measure of the degree to which a material can be magnetized at a specific external location. In practice, the susceptibility measurement requires calibration considering factors established by the core geometry, type of material, and core profiling type (BLUM, 1997; MCNEILL et al., 2017). In this article, MS has measurement unit in $instr.units$ in a single feature used.

NGR is acquired through scintillation detectors arranged along with the core interval at every 20 cm distance. Electromagnetic waves (Gamma Rays) act on the frequency between 1019 and 1021 Hz emitted spontaneously by an atomic nucleus during the radioactive decay, divided into packages called Photons. Photons are related to the energy transported (E) and the wavelength (λ) or frequency V. (BLUM, 1997; MCNEILL et al., 2017). In this article, NGR has measure units in cps in a single feature used.

PWL is the reading value of the ultrasonic P-wave velocity, transmitted in a 500 kHz wave pulse, through the core section at an interval and specific repetition. The sonic speed is the relation between the distance covered by the rocky material and the signal's duration in the same material (in seconds) (BLUM, 1997; MCNEILL et al., 2017). In this article, PWL has the unit of measure in m/s in a single feature used.

RGB is the measure of the values of the Red, Green, and Blue channels extracted from the digitization of the core, with reading intervals between 0.04 cm to 1 cm and values ranging from 0 to 255 (0 is completely black and 255 is entirely white) (MCNEILL et al., 2017). In this article, RGB has the unit of measure in R, G, B with three features.

RSC is the measurement of spectral counts in the range of 380 to 925 nm, using the light spectrophotometer, acting on the entire spectrum of the visible (BLUM, 1997; MCNEILL et al., 2017). In this article, RSC measures units in Reflectance L*, Reflectance a*, and Reflectance b* with three features used.

2.1 Machine Learning and Random Forest

Machine Learning (ML) is a set of computational techniques capable of inducing the computer to discover new knowledge (GÉRON, 2017). Its conception integrates finding hidden knowledge or in a more agile and practical way than traditional human techniques. ML can be classified into two main categories regarding training and type of supervision: supervised and unsupervised (GÉRON, 2017; GOLLAPUDI, 2016). In this article, the records of geophysical properties extracted from IODP-Expedition 362 are supervised learning, data classification, and the RF method. The records extracted from the images, in their design, are selected in the category of unsupervised learning using the SLIC Superpixel. It is important to note that for the images, the dataset is organized in a training format supervised by the addition of a specific label of the type or group of lithology in each segmented region.

The RF method in the supervised learning category was based on the decision trees method. It is a set learning technique organized in the combination of many trees, based on different data samples and resource combinations (depth and number of nodes) (SINNOTT and SUN, 2016; KOTU and DESHPANDE, 2019). The RF creates multiple combinations of decision trees and averages, processing each combination of tree and nodes. For each iteration in the execution of the algorithm, the RF method selects the best combination (model) of training configuration. It stores its result (bagging) for later use during its execution with the nodes' choice and configuration using the Gini or Entropy functions.

The RF method is applied in several areas, especially in lithological classification (XIE et al., 2020; AO et al., 2020; BRESSAN et al., 2020; KUMAR et al., 2019).

2.2 Interpolation

The method allows estimating or approximating a value or range of values from a set of specific multivariate data related to a given theme, based on the context of sampling or experiment (CELANT and BRONIATOWSKI, 2016; FARRELL, 2018). In geology, interpolation processes are necessary due to the characteristics of the data collected on the rocks, either by the type of lithology, the range of data collected, or the geophysical properties used.

According to CHEN et al., 2018 the geophysical properties used in this work are of irregular sparse data. Irregular because they are collected at irregular depths between all the properties analyzed. The reading intervals are to site U1480: RGB: 1 cm, GRA, MS, RSC and PWL: 2.5 cm, NGR: 10 cm and MAD: non-standard and site U1481: RGB: 0.04 cm, GRA, MS, RSC, and PWL: 2.5 cm, NGR: 10 cm and MAD: non-standard. They are sparse because the data reading intervals create large gaps in depths without records in several cases. Based on the type of data sampled, it was defined to use the following interpolation methods: Linear, Quadratic, Cubic, Spline, Slinear, Akima, Pchip and Piecewise.

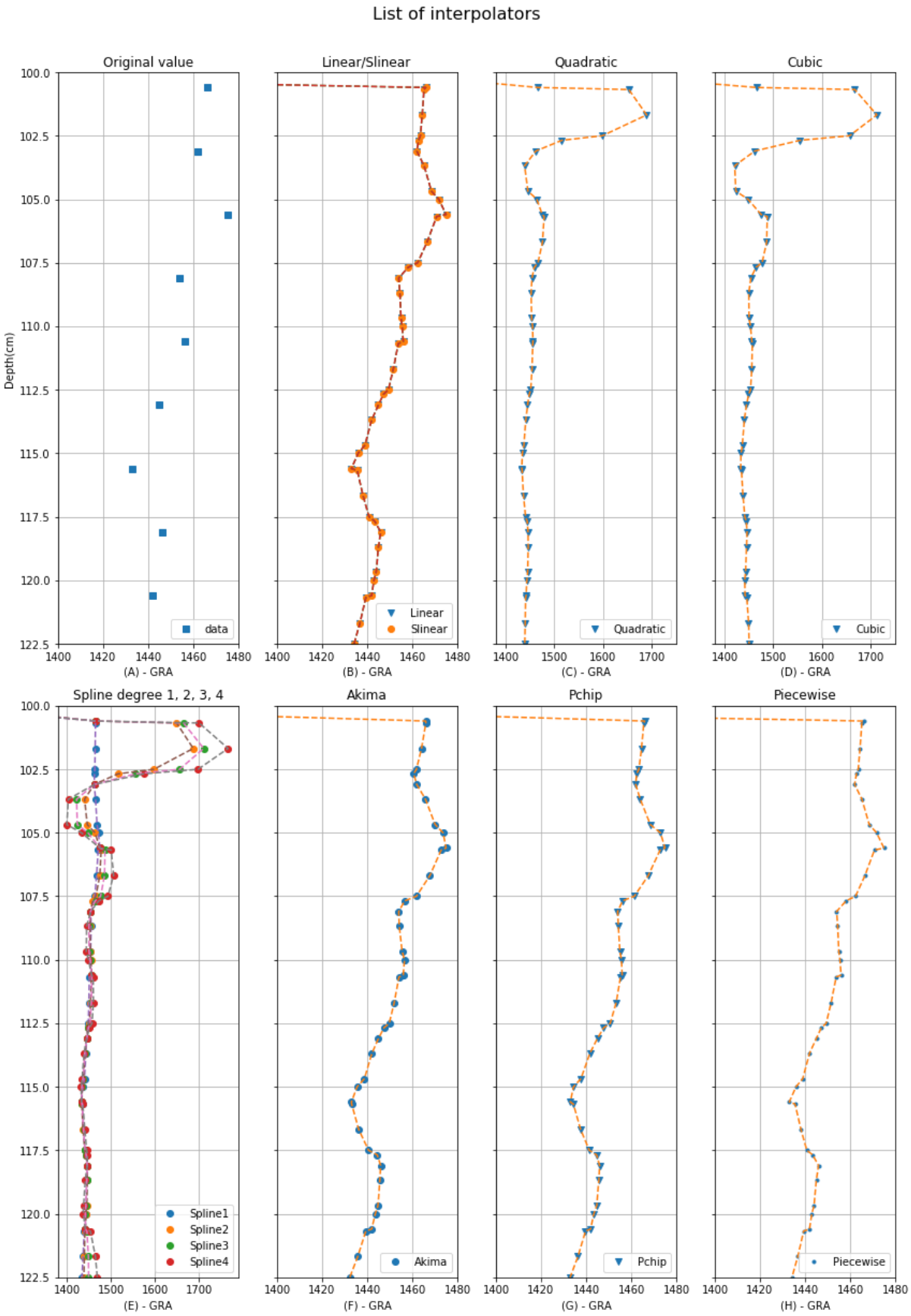

Figure 1 List of interpolators. Data example: GRA geophysical property, site U1480, hole E, Core 1H, Section 2 (range 100 cm to 122.5 cm). The depth is measured in cm according to the shipping definitions. The geophysical properties GRA, MS, RSC, PWL, NGR, MAD and RGB have specific

range and reading values (column y of the figure). (A) are the points of reading the original data. (B) are the data interpolated by the Linear and Slinear interpolator. (C) are the data interpolated by the Quadratic interpolator. (D) are the data interpolated by the Cubic interpolator. (E) are the data interpolated by the Spline interpolator degree 1, 2, 3 and 4. (F) are data interpolated by the Akima interpolator. (G) are data interpolated by the Pchip interpolator. (H) are data interpolated by the Piecewise interpolator.

Linear

Linear interpolation or Interpolation 1-D is the process of interpolation between two points in the same dimension (SALOMON, 2006; PROYAN and KISELEV, 2010). A line is used between two neighboring samples and the appropriate point is calculated along that line according to the interval defined in the depth line (cm) as shown in the Fig. 1 - B e Fig. 2.

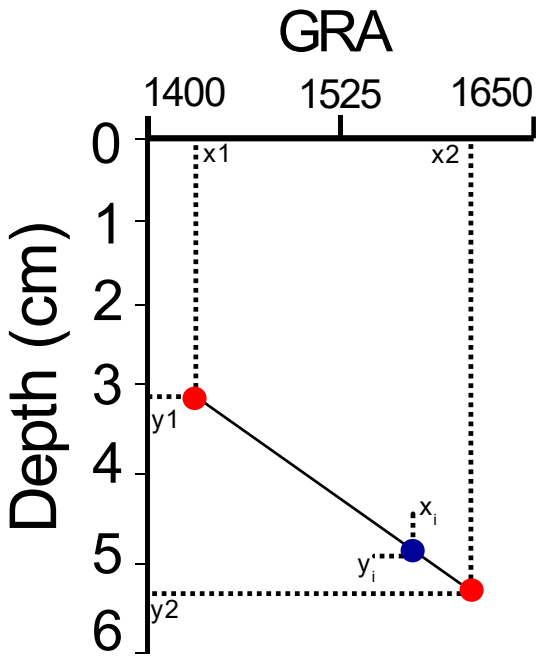

Figure 2 Value interpolation model using linear interpolator. Example for GRA geophysical property, site U1480, hole E, Core 1H, Section 2 (range 0cm to 6cm). The depth is measured in cm according to the shipping definitions. The geophysical properties GRA, MS, RSC, PWL, NGR, MAD and RGB have specific range and reading values (column y).

Column x is defined as the value of the geophysical property, in this case it is GRA, and line y is the depth value. y1, y2, x1, x2 are original values obtained by the IODP-Expedition. The equation for calculating the interpolation value x_i at depth x_i is defined as:

$$x_i = y1 + (x_i - x1)(\frac{y2-y1}{x2-x1}),\ y1 < x_i < y2. \quad (1)$$

Quadratic and Cubic

Quadratic interpolation and cubic interpolation are performed using polynomials of degrees 2 and 3 at different points p'$_0$, p'$_1$, ..., p'$_n$ (Fig. 1 - C, D) each with polynomial function $fp_{p'_n}$ defined by (PROYAN and KISELEV, 2010; BUZZI-FERRARIS and MANENTI, 2010):

$$fp_{p'_n} = a_0 + a_1 x + a_2 x^2, \text{ for degrees 2} \quad (2)$$

$$fp_{p'_n} = a_0 + a_1 x + a_2 x^2 + a_3 x^3, \text{ for degrees 3} \quad (3)$$

Where a_0, a_1, a_2 and a_3 are the pairs of coefficients to be calculated, in this case, depth and value of the geophysical property with interpolation flow from the smallest to the largest in relation to the reading value of the geophysical property and the depth, and with quantity reading values of the geophysical property greater than two units per core and section.

The list of original points (geophysical properties) are organized in a matrix where each ready (matrix line with three point interpolation (for degrees 2) and four point interpolation (for degrees 3)) represents the calculation of the $fp_{p'_n}$ function resulting in a matrix of non-zero determinant and admitting a single result.

Spline

Spline interpolator is composed of a set of polynomial functions that are connected by certain nodes in certain sections of a curve defining that at each point of interpolation two polynomials connect and their first derivatives must have the same value as all derivatives (K- 1) must be continuous (SALOMON, 2006;

LYCHE and MØRKEN, 2018). Spline acts in the processing of curves and smoothing being related to its data set (points) and its functions (formulas).

The interpolator is classified according to the combination of the type of degree (K) of the polynomials present, its complexity in the analysis of the combination of points and smoothing of the nodes or the curve in the interpolated section, with the cubic spline being the most used. In this article, combinations of the spline in the degree of the polynomial were used ($K - 1, 2 \leq K \leq 5$).

Spline degree 1 or Spline Linear consists of several linear polynomials ($Sp_{(x)}$) joined in order to achieve continuity between the original points and the interpolated points. Fig. 1 - E and Fig 3 represents visual analysis in the application of the grade 1 spline and Eq. 4 describes how the operation is performed.

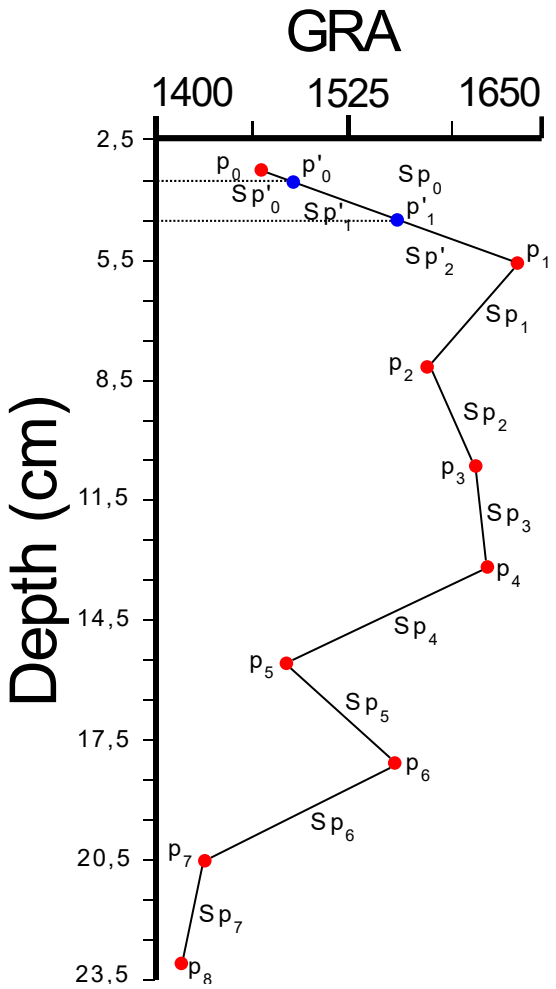

Figure 3 Value interpolation model using Spline degree 1 interpolator. Example for GRA geophysical property, site U1480, hole E, Core 1H, Section 2 (range 2.5 cm to 23.5 cm). The p_i (0 to 8) points are original values of the property reading on the expedition. The p'_i (0 to 1) points are values to be interpolated according to the defined depth interval setting. Sp_i and Sp'_i are segments between the original points and points calculated according to the quantity and degree of polynomials using, resulting in the shape of the connection (line or curve).

$$Sp'_{(x)} = y_i + m_i(x - p_i) = y_i + \frac{y_{i+1} - y_i}{p_{i+1} - p_i}(x - p_i) \quad (4)$$

where $Sp'_{(x)}$ is the segment to be calculated according to the depth value, x is the geophysical property value, m_i is the slope value of the curve calculated by the derivative, y_i is the depth value (in cm) and p_i is the point value (interpolated value found).

Importantly, $Sp_{(x)}$ is an interval between min (x) and max (x), continuous between max (x) and min (x), and min (x) = $p_0 < p_1 < p_2$ ... = max (x), such that $Sp_{(x)}$ is a linear polynomial at each interval between points p_i.

Splines of higher degree such as degree 2, 3 and 4 are applied at the approach level when it requires greater smoothness in calculating the curve. By the degree of the polynomial (K), it is necessary to determine the value of K-1 control points on the position of $Sp'_{(x)}$ resulting in an accuracy in the angle and shape of the curvature of the new segment. The $Sp'_{(x)}$ function for splines of higher degree follows the linear spline pattern by adjusting the quantity and degree of polynomials in each $Sp_{(x)}$ segment.

Slinear

The Slinear interpolation method acts in a similar way to the linear interpolator and spline interpolator degree 1 (Fig. 1 - B) using the same calculation bases (equations). Its main characteristic is the interpolation of values according to a pattern (within a defined range) of the starting and ending points (SALOMON, 2006). Outliers are discarded, making it impossible to continue interpolation. Slinear depends on the values being in a logical sequence of operation, their interpolation order follows the flow $x \rightarrow x$, where x is the value of the geophysical property with an initial value in relation to the non-zero depth.

Akima

Akima or Sub-Spline interpolator is a method of interpolation in the form of a cubic polynomial in parts acting in a univariate manner similar to a Spline degree 3 (AKIMA, 1986; FEDOROV, 2013), according to the equation:

$$fp_{p'_i} = a_i + b_i\,(x - x_i) + c_i\,(x - x_i)^2 + d_i(x - x_i)^3 \quad (5)$$

Where a_i, b_i, c_i and d_i are determined through the derivative of each point p_i. x is the value of geophysical property and x_i is the value of the geophysical property to be calculated at the point p'_i.

The method is based on a function composed of a set of polynomials of degree 3 and applied to successive data point intervals. It acts in order to estimate the first derivative of the function at each point (respective slope of the curve) based on the analysis of up to six reference points according to Fig. 1 - F and Fig. 4. The resulting curve adapts to various types of univariate data with the presence of several original random points.

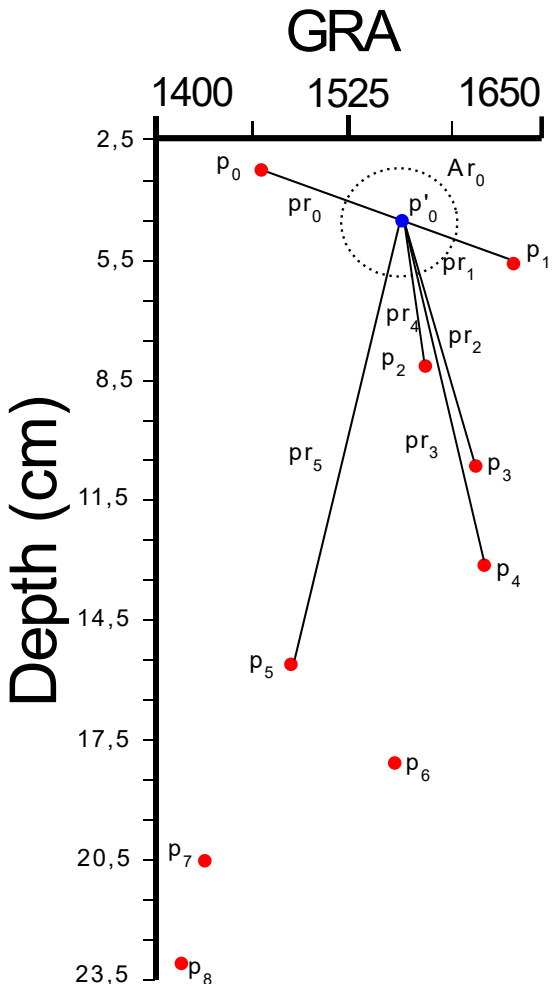

Figure 4 Value interpolation model using Akima interpolator to a point p'_0 with six reference points. Example for GRA geophysical property, site U1480, hole E, Core 1H, Section 2 (range 2.5 cm to 23.5 cm). p_i are points with reading values between depth and geophysical property. p'_i point to be interpolated. Ar_0 is the possible location area for the interpolated point. pr_i are the connections to the reference points p_i.

Contrary to the Spline Degree 3, the Akima method does not require continuity of the second derivative, does not act in smoothing the curve, reduces the oscillation that Spline usually produces and uses the other derivatives as free parameters between the location and the reference points analyzed.

Pchip

Piecewise Cubic Hermite Interpolating Polynomial (Pchip) is an interpolator that uses data in the cubic Hermite format by dividing the interpolation into equal parts or subsets of cubic polynomials. Pchip has a more suitable application in relation

to cubic Spline if the data has flat and steep sections preserving the geometry of the data location and the monotonicity between the points (RABBATH and CORRIVEAU, 2019; FRITSCH, 1982). Pchip identifies 4 different points and analyzes their slopes (through derivatives degree 1, 2 and 3) and the average between the connections of the points through a linear interpolator returning the value of the interpolation function as defined in the equation fp_x:

$$fp_x = d_0 + d_1 * x + d_2 * x^2 + d_3 * x^3 \quad (6)$$

Where x is the point to be calculated, d_0, d_1, d_2 and d_3 are derived from the four points analyzed, respecting the value with the respective sign (negative or positive).

The main characteristics of the Pchip are in the sense that it analyzes the slope of the four points analyzed using its derivatives, respects the monotonicity between the data intervals, presents less overshoot / undershoot than the Spline interpolator and presents itself as a method to perform curve adjustments for all data points (Fig. 1 - G).

Piecewise

Piecewise interpolator interpolates the value between two distinct irregular points using polynomials in parts and respective derivatives of the points (FEDOROV, 2013). It uses the same equation as the Spline interpolator, the definition of the degree of the polynomial is dynamic and adapts to the degree necessary to cover all derivatives at the analyzed points (or interval).

Piecewise seeks to use an equal number of derivatives (polynomial degree) at each point in the interval. If the number of possible derivatives is different, it will

use the lowest value or the value of the righter point of the segment (curve). An exception will be reported if it is not possible to calculate the derivatives of the analyzed points or the number of derivatives for interpolation results in high values making an appropriate calculation impossible (Fig. 1 - H).

Operation flow of the interpolators

It is important to highlight that each interpolator analyzes the original data range of the geophysical properties and performs the operation flow for interpolation according to the location characteristics between depth and number of features present by dataset in relation to the core and section. The table 1 shows the operation flow of each interpolator based on the geophysical properties used in this article.

Table 1: Operating flow of interpolators.

Interpolator	Operation Flow
1. Linear	$x \leftrightarrow x$
2.1 Quadratic	$x \rightarrow x$
2.2 Cubic	$x \rightarrow x$
3. Spline	$x \leftrightarrow x$
4. Slinear	$x \rightarrow x$
5. Akima	$x \rightarrow x$
6. PChip	$x \leftrightarrow x$
7. Piecewise	$x \rightarrow x$

The operation flow described in the interpolators comprises the depth range of the hole, core and section respectively with the sequence of features of the GRA, MAD, MS, NGR, PWL, RGB and RSC geophysical properties.

The Linear interpolator being a 1-D dimension interpolator processing data between two variables (two points p_i) covers the creation of new data interpolated in the flow $x \leftrightarrow x$ of greater or lesser value of x, where x is the value of the geophysical property in relation to depth.

Quadratic, Clubic and Piecewise interpolators are interpolators that, in their design, use polynomials in the interpolation processing, requiring initial values for the interpolation flow by performing the operation in the $x \rightarrow x$ direction where x is the geophysical property value in relation to non-zero depth and reading value greater than two units per core and section.

Slinear interpolator, as already described, by the initial and final data range characteristic follows the operation flow $x \rightarrow x$ where x is the value of the geophysical property with initial value in relation to the non-zero depth.

Spline and Pchip interpolator cover interpolation with polynomial functions in calculating the segment between the points, but the continuity between the other points will occur through the processing of a linear polynomial reaching an operation flow of the interpolation equal to the Linear Interpolator.

Akima interpolator uses a set of polynomials in parts grouped according to the distribution of the points and their connections. In addition to the characteristic, it requires reference points when designing the interpolation without making it impossible to perform the operation. The operation flow follows the logical sequence $x \rightarrow x$ where x is the value of the geophysical property with an initial value in relation to the non-zero depth.

The validation of the interpolated data will occur by applying the classification method according to the defined training and testing configuration, seeking to identify the best accuracy result in the interpolated data range.

Examples of the application of numerical interpolation techniques in seismic data (CHAI et al., 2020; TURCO, AZEVEDO and HEROLD, 2019), in mineral

exploration (GUO et al., 2018), and oil and gas exploration (NGUIMBOUS-KOUOH and MANGUELLE-DICOUM, 2019).

2.3 Image Data Annotation

Visual marking or human annotation on image, video, or audio indicating an area of interest. Each area of interest is defined as a region being automatically or manually indicated and can have the shapes of rectangle, circle, ellipse, polygons, point and polyline (DUTTA and ZISSERMAN, 2019; TORRALBA, RUSSELL and YUEN, 2010). The textual description of each region of interest is the key point of annotating the image.

This description includes marking tags indicating the relevant information for further processing. Semi-structured data type is used as output, which allows changes to the structure and content at run time, forming a dynamic data structure integrated with the analyzed data source. Semi-structured data can be in .json, .xml or .csv file format.

```
" 362-u1480e-2h-2-
a_shlf7853651_20160813145856.jpg6748938": {
    "filename": " 362-u1480e-2h-2-
a_shlf7853651_20160813145856.jpg",
    "size": 6748938,
    "regions": [
     {
       "shape_attributes": {
         "name": "polygon",
         "all_points_x": [
           400,
           1630,
           1660,
           1120,
           710,
           410,
           400
         ],
         "all_points_y": [
           1590,
           1600,
           4930,
           4840,
           4890,
           4960,
           1590
         ]
       },
       "region_attributes": {
         "Lithology": "Medium-sand_sandstone",
         "Depth": "0.5_17.0"
       }
     }
   ],
   "file_attributes": {}
 }
```


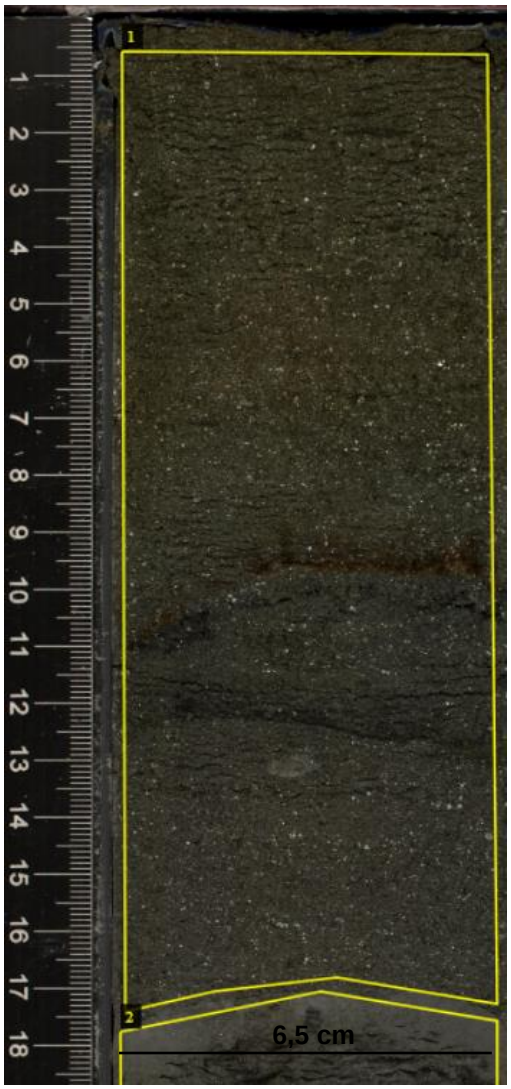

Figure 5 Example of .json file format content of the main tags used in this study. The "name" tag defines the shape of the region of interest. The tags “all_points_x” and “all_points_y” are X and Y pixel coordinates of the analyzed image forming the polygon with 7 coordinates. The tag "region_attributes" includes the attributes defined for "Lithology" and "Depth". Attribute "Depth" is measured in cm, in the format: start_and separated by ‘_’. The example image belongs to site U1480, Hole E, Core 2h, Section 2. (DUTTA and ZISSERMAN, 2019).

The json output file type is widely used in markup structures that require quick, secure and easy handling (Fig. 5). It is a format widely used on the internet and compatible with unstructured databases. (HAMOUDA and ZAINOL, 2019; STREKALOVA and BOUAKKAZ, 2017; MARTINEZ-MOSQUERA, NAVARRETE and LUJAN-MORA, 2020). The file name is defined as a primary initial tag designated as an access key. The tags "filename", "size", "regions" and "file_attributes" define the primary structure with information from the analyzed

file (image, video or audio). The subtag “shape_attributes” stores information on areas of interest with pixel X and Y coordinates, in addition to specific user-defined attributes, which in this case are “lithology” and “depth” attributes.

The annotation of images together with ML are appropriate tasks for the visual processing of images, being included in procedures in the health area such as disease identification (ELAZIZ et al., 2020; SHI et al., 2019) and geology such as mineral identification and seismic processing (RAN et al., 2019; MAITRE, BOUCHARD and BÉDARD, 2019)

2.4 Segmentation SLIC Superpixel

Simple Linear Interative Clustering (SLIC) or SLIC Superpixel is a method of grouping or segmenting pixels, based on the k-means method, acting more quickly and reducing the complexity of tasks in image processing. The grouped areas are called Superpixel regions or segmented regions having their characteristic identification, location, and separation as for future processing (ACHANTA et al., 2010; ACHANTA et al., 2012; KAVZOGLU and TONBUL, 2018). In Fig. 6, segmentation of an example image from the lithology Medium-sand_sandstone.

Figure 6 Example of lithology image Medium-sand_sandstone, original file: 362-u1480e-2h-2-a_shlf7853651_20160813145856.jpg, segmented by the SLIC Superpixel resulting in 26 regions. Apx: 115,07 cm². Original image size: 1700 x 32007 pixels Resolution: 96 DPI. Bits intensity: 24. The grouping of pixels takes place in an appropriate way to the process, covering the entire image, separating characteristics and textures relevant to each lithology.

The computational logic for the operation of the SLIC Superpixel method begins with the definition of the maximum number of K groups with approximately the same pixel size. For color images in the CIELAB space, the procedure starts with the step of defining the initial cluster centroid $C_i = [L_i a_i b_i x_i y_i]$ where C_i are arranged in a regular G pixel grid. To generate SLIC Superpixels of approximate sizes, the G= √(N/K) where the centroids are located in a lower gradient pixel region, avoiding the centroid being close to the edge of the Superpixel or having a noisy pixel (ACHANTA et al., 2012; KAVZOGLU and TONBUL, 2018).

When grouping pixels, each pixel i is associated with the cluster centroid C_i nearest, whose search region is related to the distance measurement Da which determines the cluster center closest to each pixel analyzed, as seen in Eq. 1 and Fig. 7.

$$Da = d_{Lab} + \left(\frac{d_{x,y}}{G}\right)m \tag{7}$$

Where Da is the distance between the centroid C_i and the pixel i, m is the compression between the maximum and minimum color distance in the Superpixel, d_{Lab} are the color values in the CIELAB space, G is grid interval and $d_{x,y}$ is the value of the row and column in pixels.

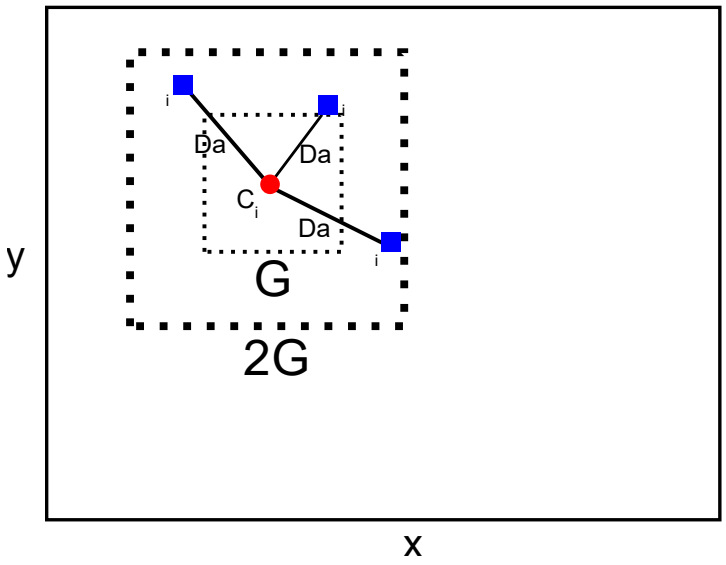

Figure 7 SLIC processing model in the allocation of a pixel i in the centroid C_i according to the distance Da in the region 2G x 2G. x and y are defined as rows and columns of the image, measured in pixels. The size x, y is variable and depends on the image analyzed.

As the size of the expected Superpixel space region is approximated to G x G, the search for similar pixels is done in a 2G x 2G region around the centroid C_i. The main point to speed up the analysis is here because the processing is limited

to the size of the research region reducing the calculation of a and increasing the speed of the grouping in relation to the conventional analysis using K-means in which each pixel i is compared with all centroid C_i. At the end, after assigning all pixels to the centroid, the adjustment step is processed where for each centroid C_i the average value of its location in relation to the total grouped pixel is calculated. The SLIC method acts to minimize the error of grouping and location of pixels in relation to the centroid through testing and updating steps using the Euclidean Norm or L2 Norm.

The distance Da (Eq. 1) represents processing in the CIELAB color space, whose range of values varies in combinations of L_i which is the luminance of black to white, a_i as being the combination of green to red and b_i as being the combination of blue to yellow, with the pixel position in line and column (x, y). The range of CIELAB values varies with the type, size of the image and number of colors present.

SLIC based on the K-means method, being a faster iterative grouping, more efficient in using computational memory, exhibits adherence between the targeted regions and improves the performance of the segmentation algorithm.

2.4.1 Texture and Image color

Images collected in IODP-Expedition 362 are of the RGB type. The image's size is calculated through the resolution of the pixel quantity and the number of colors per pixel. IODP images, the resolution used is 96 Dots Per Inch (DPI) and 24 bits of color in the 3 channels. The size of the image (width and height) is variable

and depends on the sample size collected at the time of drilling (MCNEILL et al., 2017).

The segmented regions follow the RGB image pattern, processed individually, maintaining all the initial characteristics (Fig. 8).

Figure 8 Segmentation region extracted from the image: 362-u1480e-2h-2-a_shlf7853651_20160813145856.jpg, lithology: Medium-sand_sandstone cod_region: 22, mean_intensity color: 10.03, Apx: 3.56 cm², Pixel count: 134368

Each region, the $mean\_intensity$ feature is extracted and stored in an appropriate dataset and calculated by the equation:

$$mean\_intensity = \frac{\frac{\sum_n^1 R}{n} + \frac{\sum_n^1 G}{n} + \frac{\sum_n^1 B}{n}}{3} \quad (8)$$

Where:

$mean\_intensity$ is defined as the average RGB color intensity in the segmented region. R, G, B are the color values ranging from 0 to 255, and n is the total number of pixels in the segmented region. Mean values of $mean\_intensity$ are float numbers and range from 0 to 255.

The brightness values or gray levels for the segmented regions are calculated using the Gray Level Co-occurrence Matrix (GLCM). GLCM is defined as the analysis and tabulation of gray level values at a given displacement in the image segment (HALL-BEYER, 2017). The displacement of the pixel block in the image follows the pattern defined according to Fig. 9.

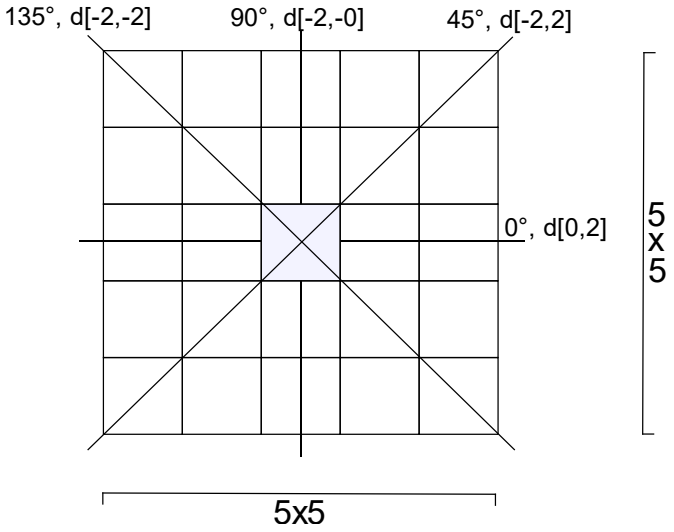

Figure 9 Symmetrical block 5x5 with the pixel of interest indicating the possibilities of angles (0 °, 45 °, 90 °, 135 °) and distance (d) for processing. The block maintains a symmetrical pixel distribution pattern with normalized gray values on the sum scale equal to 1 before calculating the texture properties.

The pixel of interest in GLCM follows the context of binary value (second-order) advancing and processing in groups of two pixels in the original image. The advance of the pixel of interest includes the configuration of two properties: angle and distance. For angle, there are four possible combinations of distribution in the horizontal, vertical and diagonal (0º, 45º, 90º, 135º). Distance comprises the list of the displacement of the pixel pair in the chosen angle defined as an array of initial and final values return a 4-d size matrix with the dimensions GLCM [i, j, d, Ɵ], where i and j are the gray values in the row and column dimensions, d is the distance from the pixel pair and θ is the angle in degrees.

The texture properties calculated using the GLCM are Contrast, Dissimilarity, Homogeneity, Angular Second Moment (ASM), Energy and Correlation.

According to the equations (HALL-BEYER, 2017; ALAZAWI, SHATI and ABBAS, 2019):

$$Contrast = \sum_{i,j=0}^{levels-1} P_{i,j}(i-j)^2 \quad (9)$$

$$Dissimilarity = \sum_{i,j=0}^{levels-1} P_{i,j}|i-j|^2 \quad (10)$$

$$Homogeneity = \sum_{i,j=0}^{levels-1} \frac{P_{i,j}}{1+(i-j)^2} \quad (11)$$

$$ASM = \sum_{i,j=0}^{levels-1} P_{i,j}^2 \quad (12)$$

$$Energy = \sqrt{ASM} \quad (13)$$

$$Correlation = \sum_{i,j=0}^{levels-1} P_{i,j} \left[\frac{(i-\mu i)\,(j-\mu j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right] \quad (14)$$

Where:

i and j are defined as position between row and column of the matrix, P is the GLCM matrix [i, j, d, Θ]. The result of the properties is an array 2-d Property [d, Θ], with property value at the distance d and angle Θ.

$Contrast$ (Eq. 9) is the sum of the measurement related to the distance of the GLCM diagonal, allocating specific weights. They receive weight 0 when i and j are equal, weight 1 when the difference between i and j is equal to 1, weight 4 when the difference between i and j is 2 and thus increase exponentially as the difference between i and j increases.

$Dissimilarity$ (Eq. 10) follows the Contrast calculation pattern by modifying the weight value in a linear way.

$Homogeneity$ (Eq. 11) acts in order to weight the values by the inverse of the Contrast weights, that is, the weights decrease exponentially in relation to the GLCM diagonal. Its value ranges from 0 to 1.

ASM (Eq. 12) calculate the sum of the weight measurement of each position i and j. High values for ASM indicate high pixel ordering with the same values.

$Energy$ (Eq. 13) is calculated using the square root of ASM.

$Correlation$ or Linear Correlation (Eq. 14) is the sum of the linear dependence measure of the gray level values of the GLCM matrix with the closest neighbors. For a symmetric matrix, the correlation value is equivalent to variance (σ^2).

Research in geology how to use of SLIC Superpixel covers few practical applications of this method with studies in the field of rock morphology (MALLADI, RAM and RODRÍGUEZ, 2014), classification of mineral grains in microscopic images (MAITRE, BOUCHARD and BÉDARD, 2019; JIANG et al., 2017) and detection of lithological limits in images from remote sensing platforms (VASUKI et al., 2017).

2.5 Validation and Evaluation

Confusion matrix represents the visualization of the data classification performance according to the division between real data and predicted data, being the main way to tabulate the results of a classification, being present in its structure the metrics of evaluation accuracy, precision, recall and F_1-score (BRESSAN et al., 2020; NISBET, MINER and YALE, 2018), according to the organization shown in Fig. 10. The metrics for evaluating the confusion matrix are employed in a precise and correct way to classify geophysical data with excellent

results in the binary classification of lithologies in IODP-Expeditions (BRESSAN et al., 2020).

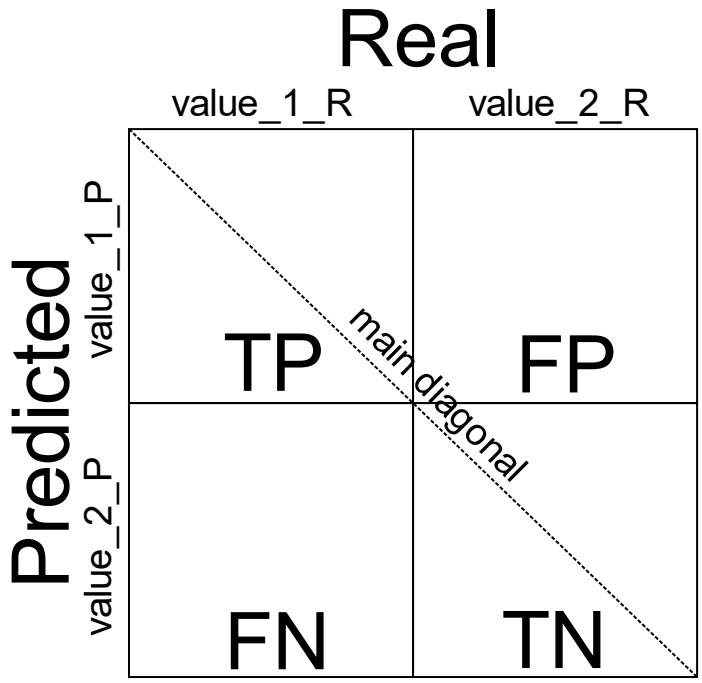

Figure 10 Organization of the confusion Matrix. The rows and columns are divided into predicted data and real data combining TP (True Positive), FP (False Positive), FN (False Negative) and TN (True Negative). The representation in the matrix format is TP [value_1_P: value_1_R], FP [value_1_P: value_2_R], FN [value_2_P: value_1_R] and TN [value_2_P: value_2_R].

The metrics $Accuracy$, $Precision$, $Recall$ and $F_1 - score$ are the results of sampling the classification on the confusion matrix, being defined by the following equations:

$$Accuracy = \frac{TP+TN}{P+N} \quad (15)$$

$$Precision = \frac{TP}{TP+FP} \quad (16)$$

$$Recall = \frac{TP}{TP+FN} \quad (17)$$

$$F_1 - score = 2 * \frac{\text{Precision} * \text{recall}}{\text{Precision} + \text{recall}} \quad (18)$$

$Accuracy$ (Eq. 15) is formed by the division between the sum of true positive and true negative values and the total sum of positive and negative values. The calculated value is presented as a percentage, from 0 to 100%, returning how much the model got right from the possible classification.

$Precision$ (Eq. 16) and $Recall$ (Eq. 17) are formed by the division between TP and the sum of TP and FP values and the sum of TP and FN values, respectively. These two metrics assess the model as to whether the proportion of identifications (class to be classified) is correct.

$F_1 - score$ (Eq. 18) uses values of $Precision$ and $Recall$ calculated by dividing the multiplication of $Precision$ and $Recall$ with the sum of the values of $Precision$ and $Recall$. The calculated value shows a balance between $Precision$ and $Recall$ in the classification of the proposed class and interpreted as the reliability value of the found accuracy (SARKAR, BALI and SHARMA, 2018).

Cross-validation is a method used to evaluate the classification data's performance using metrics from the confusion matrix. Cross-validation processing divides data into equal-sized groups called K groups or K folds. The divided groups are trained and tested in n possible combinations in the same configuration as the machine learning method used on all data (BRESSAN et al., 2020; GÉRON, 2017).

The results of the confusion matrix give rise to new analyzes related to error and success rates. Main values are related to True Positive Rate (TPR) and False Positive Rate (FPR) and visual analysis can be applied using the Receiver Operating Characteristic (ROC) method and its area on the ROC curve called

Area Under the ROC curve (AUC) according to the equations (GÉRON, 2017; SUN et al., 2020):

$$TPR = \frac{TP}{TP+FN} \quad (19)$$

$$FPR = \frac{FP}{FP+TN} \quad (20)$$

The new metric applied in this article is called Area Superpixel (Apx). In this analysis, the images extracted from the IODP-Expedition are analyzed after processing, segmentation and classification of the segmented regions according to the definition of the lithology groups, and the Apx is calculated as defined:

$$Apx = \frac{(A' * a')}{1000} \quad (21)$$

$$ApxTotal = \sum_{n}^{1} Acm \quad (22)$$

Where:

Apx is the resulting total area in cm² of the segmented region. A' is the total number of pixels in the segmented region. a' is the pixel size, in cm, according to the DPI resolution of the analyzed image (ACHARYA and RAY, 2005; GONZALES and WOODS, 2007). In this work, the images are in 96 DPI resolution, the value of a' is 0.02646 cm. $ApxTotal$ of the analyzed image is calculated (Eq. 22) by the sum of the Apx of all the segmented regions.

References

ACHANTA, R. et al., 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34,11. https://doi.org/10.1109/TPAMI.2012.120.

ACHANTA, R. et al., 2010. SLIC Superpixels. EPFL Technical Report 149300, June 2010.

ACHARYA, T., RAY, A. K., 2005. Image Processing. Principles and Applications. Wiley-Interscience.

AKIMA, H., 1986. A Method of Univariate Interpolation That Has the Accuracy of a Third-Degree Polynomial. Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce.

ALAZAWI, S. A., SHATI, N. M., ABBAS, A. H., 2019. Texture features extraction based on GLCM for face retrieval system. Periodicals of Engineering and Natural Sciences, 7, 3. http://dx.doi.org/10.21533/pen.v7i3.787.

AO, Y. et al., 2020. Probabilistic logging lithology characterization with random forest probability estimation. Computers & Geosciences, 104556. doi:10.1016/j.cageo.2020.104556.

BLUM, P., 1997. Physical properties handbook: a guide to the shipboard 583 measurement of physical properties of deep-sea cores, College Station, Texas, 584 USA, http://www-odp.tamu.edu.

BRESSAN, T. S. et al., 2020. Evaluation of machine learning methods for lithology classification using geophysical data. Computers & Geosciences, 139. https://doi.org/10.1016/j.cageo.2020.104475.

BUZZI-FERRARIS, G., MANENTI, F., 2010. Interpolation and Regression Models for the Chemical Engineer: Solving Numerical Problems. Chapter 1. WILEY-VCH Verlag GmbH & Co. KGaA.

CELANT, G., BRONIATOWSKI, M., 2016. Interpolation and Extrapolation Optimal Designs 1: Polynomial Regression and Approximation Theory. Wiley Online Library. DOI: 10.1002/9781119292272.

CHAI, X. et al., 2020. Deep learning for irregularly and regularly missing data reconstruction. Scientific Reports 10, 3302 (2020). https://doi.org/10.1038/s41598-020-59801-x.

CHEN, Y. et al., 2018. The Interpolation of Sparse Geophysical Data. Springer: Surveys in Geophysics, 40, 73-105 (2019). https://doi.org/10.1007/s10712-018-9501-3.

CHEN, X., ZHANG, F., ZHANG, R., 2017. Medical image segmentation based on SLIC superpixels model. International Conference on Innovative Optical Health Science, 10245 (2017). https://doi.org/10.1117/12.2258384.

DUTTA, A., ZISSERMAN, A., 2019. The VIA Annotation Software for Images, Audio and Video. 27th ACM International Conference on Multimedia. https://doi.org/10.1145/3343031.3350535.

ELAZIZ, M. A. et al., 2020. New machine learning method for image-based diagnosis of COVID-19. PLOS ONE. https://doi.org/10.1371/journal.pone.0235187.

FARRELL, P., 2018. Numerical Mathematics. Lecture Notes. Hamburg University of Technology, Version 22, August 2018.

FEDOROV, D. V., 2013. Introduction to Numerical Methods. Version 13.05. GNU Licence.

FRITSCH, F. N., 1982. PCHIP FINAL SPECIFICATIONS. Version 8.5. Lawrence Livermore National Laboratory, August 1982.

GÉRON, A., 2017. Hands-On Machine Learning with Scikit-Learn and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.

GONZALES, R. C., WOODS, R. E., 2007. Digital Image Processing, Third Edition. Pearson.

GOLLAPUDI, S., 2016. Practical Machine Learning. Packt Publishing, Birmingham B3 2PB, UK.

GUO, Z. et al., 2018. Geophysical Field Data Interpolation Using Stochastic Partial Differential Equations for Gold Exploration in Dayaoshan, Guangxi, China. Minerals 2019, 9, 14. https://doi.org/10.3390/min9010014.

HALL-BEYER, M., 2017. GLCM Texture: A Tutorial v. 3.0. University of Calgary. http://dx.doi.org/10.11575/PRISM/33280.

HAMOUDA, S., ZAINOL, Z., 2019. Semi-Structured Data Model for Big Data (SS-DMBD). 8th International Conference on Data Science, Technology and Applications (DATA 2019). DOI: 10.5220/0007957603480356.

JIANG, F. et al., 2017. Grain segmentation of multi-angle petrographic thin section microscopic images. IEEE International Conference on Image Processing (ICIP). https://doi.org/10.1109/ICIP.2017.8297009.

KAVZOGLU, T., TONBUL, H., 2018. An experimental comparison of multi-resolution segmentation, SLIC and K-means clustering for object-based classification of VHR imagery. International Journal of Remote Sensing, 39, 18. https://doi.org/10.1080/01431161.2018.1506592

KOTU, V., DESHPANDE, B., 2019. Chapter 4 – Classification. Data Science (Second Edition). Concepts and Practice. https://doi.org/10.1016/B978-0-12-814761-0.00004-6.

KUMAR, C. et al., 2019. Automated lithological mapping by integrating spectral enhancement techniques and machine learning algorithms using AVIRIS-NG hyperspectral data in Gold-bearing granite-greenstone rocks in Hutti, India. International Journal of Applied Earth Observation and Geoinformation, 86, 102006. doi:10.1016/j.jag.2019.102006

LYCHE, T., MØRKEN, K., 2018. Spline Methods Draft. Department of Mathematics. University of Oslo.

MAITRE, J., BOUCHARD, K., BÉDARD, L. P., 2019. Mineral grains recognition using computer vision and machine learning. Computers & Geosciences, 130, 84-93 (2019). https://doi.org/10.1016/j.cageo.2019.05.009.

MARTINEZ-MOSQUERA, D., NAVARRETE, R., LUJAN-MORA, S., 2020. Modeling and Management Big Data in Databases—A Systematic Literature Review. Sustainability, 12, 634. DOI:10.3390/su12020634.

MALLADI, S. R. S. P., RAM, S., RODRÍGUEZ, J. J., 2014. Superpixels using morphology for rock image segmentation. Southwest Symposium on Image Analysis and Interpretation. https://doi.org/10.1109/SSIAI.2014.6806050.

MCNEILL, L. C. et al., 2017. Expedition 362 methods. Proceedings of the International Ocean Discovery Program, 362. https://doi.org/10.14379/iodp.proc.362.102.2017.

NGUIMBOUS-KOUOH, J. J., MANGUELLE-DICOUM, E., 2019. Evaluating Interpolation Methods by Geostatistical Modeling of the Douala Oil Field Porosity Data (Cameroon). Geoinfor Geostat 7, 1. DOI: 10.4172/2327-4581.1000203.

NISBET, R., MINER, G., YALE, K. D. D. S., 2018. Chapter 11 - Model Evaluation and Enhancement. Handbook of Statistical Analysis and Data Mining Applications (Second Edition) 2018, 215-233. https://doi.org/10.1016/B978-0-12-416632-5.00011-6.

PROYAN, V., KISELEV, Y., 2010. Statistical Methods of Geophysical Data Processing. World Scientific Publishing Co.

RABBATH, C. A., CORRIVEAU, D., 2019. A comparison of piecewise cubic Hermite interpolating polynomials, cubic splines and piecewise linear functions for the approximation of projectile aerodynamics. Defence Technology 15 (5). https://doi.org/10.1016/j.dt.2019.07.016.

RAN, X. et al., 2019. Rock Classification from Field Image Patches Analyzed Using a Deep Convolutional Neural Network. Mathematics, 7, 755. DOI:10.3390/math7080755.

SALOMON, D., 2006. Curves and Surfaces for Computer Graphics. Chapter 2. Springer. https://doi.org/10.1007/0-387-28452-4.

SARKAR, D., BALI, R., SHARMA, T., 2018. Practical Machine Learning with Python. A Problem-Solver's Guide to Building Real-World Intelligent Systems. APRESS. https://doi.org/10.1007/978-1-4842-3207-1.

SHI, J. S. et al., 2019. Effects of annotation granularity in deep learning models for histopathological images. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). https://doi.org/10.1109/BIBM47256.2019.8983158.

SINNOTT, R. O., SUN, H. D., 2016. Chapter 15 - A Case Study in Big Data Analytics: Exploring Twitter Sentiment Analysis and the Weather. Big Data, Principles and Paradigms. https://doi.org/10.1016/B978-0-12-805394-2.00015-5.

STREKALOVA, Y.A., BOUAKKAZ, M., 2017. Semi-structured Data. In: Schintler L., McNeely C. (eds) Encyclopedia of Big Data. Springer, Cham. https://doi.org/10.1007/978-3-319-32001-4_183-1.

SUN, Z. et al., 2020. A Data-Driven Approach for Lithology Identification Based on Parameter-Optimized Ensemble Learning. Energies, 13, 3903. DOI:10.3390/en13153903.

TONG, H. et al., 2019. Purifying SLIC Superpixels to Optimize Superpixel-Based Classification of High Spatial Resolution Remote Sensing Image. Remote Sens, 11(22), 2627. https://doi.org/10.3390/rs11222627.

TORRALBA, A., RUSSELL, B. C., YUEN, J., 2010. LabelMe: Online Image Annotation and Applications. Proceedings of the IEEE, 98, 8 (2010). https://doi.org/10.1109/JPROC.2010.2050290.

TURCO, F., AZEVEDO, L., HEROLD, D., 2019. Geostatistical interpolation of non-stationary seismic data. Computacional Geosciences 23, 665-682 (2019). https://doi.org/10.1007/s10596-019-9812-6.

VASUKI, Y. et al., 2017. An interactive image segmentation method for lithological boundary detection: A rapid mapping tool for geologists. Computers & Geosciences, 100, 27-40. https://doi.org/10.1016/j.cageo.2016.12.001.

XIE, Y. et al., 2020. A Coarse-to-Fine Approach for Intelligent Logging Lithology Identification with Extremely Randomized Trees. Math Geosci (2020). https://doi.org/10.1007/s11004-020-09885-y.

ZHANG, S., YOU, Z., WU, X., 2019. Plant disease leaf image segmentation based on superpixel clustering and EM algorithm. Neural Computing and Applications 31, 1225–1232. https://doi.org/10.1007/s00521-017-3067-8.