International Journal of
Geo-Information

Article

Exploratory Data Analysis of Synthetic Aperture Radar (SAR) Measurements to Distinguish the Sea Surface Expressions of Naturally-Occurring Oil Seeps from Human-Related Oil Spills in Campeche Bay (Gulf of Mexico)

Gustavo de Araújo Carvalho [1,*], Peter J. Minnett [2], Fernando Pellon de Miranda [1], Luiz Landau [1] and Eduardo Tavares Paes [3]

[1] Programa de Engenharia Civil (PEC), Laboratório de Métodos Computacionais em Engenharia (LAMCE), Laboratório de Sensoriamento Remoto por Radar Aplicado à Indústria do Petróleo (LabSAR), Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE), Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, RJ 21941-909, Brazil; pellon@labsar.coppe.ufrj.br (F.P.M.); landau@lamce.coppe.ufrj.br (L.L.)

[2] Department of Ocean Sciences (OCE), Rosenstiel School of Marine and Atmospheric Science (RSMAS), University of Miami (UM), Miami, FL 33149, USA; pminnett@rsmas.miami.edu (P.J.M.)

[3] Laboratório de Ecologia Marinha e Oceanografia Pesqueira da Amazônia (LEMOPA), Instituto Socioambiental e dos Recursos Hídricos (ISARH), Universidade Federal Rural da Amazônia (UFRA), Belém, PA 66077-830, Brazil; etpaes@gmail.com (E.T.P.)

* Correspondence: ggus.ocn@gmail.com (G.A.C.)

Received: 30 September 2017; Accepted: 13 November 2017; Published: 6 December 2017

Abstract: An Exploratory Data Analysis (EDA) aims to use Synthetic Aperture Radar (SAR) measurements for discriminating between two oil slick types observed on the sea surface: naturally-occurring oil seeps versus human-related oil spills—the use of satellite sensors for this task is poorly documented in scientific literature. A long-term RADARSAT dataset (2008–2012) is exploited to investigate oil slicks in Campeche Bay (Gulf of Mexico). Simple Classification Algorithms to distinguish the oil slick type are designed based on standard multivariate data analysis techniques. Various attributes of geometry, shape, and dimension that describe the oil slick Size Information are combined with SAR-derived backscatter coefficients—sigma-(σ_o), beta-(β_o), and gamma-(γ_o) naught. The combination of several of these characteristics is capable of distinguishing the oil slick type with ~70% of overall accuracy, however, the sole and simple use of two specific oil slick's Size Information (i.e., area and perimeter) is equally capable of distinguishing seeps from spills. The data mining exercise of our EDA promotes a novel idea bridging petroleum pollution and remote sensing research, thus paving the way to further investigate the satellite synoptic view to express geophysical differences between seeped and spilled oil observed on the sea surface for systematic use.

Keywords: Exploratory Data Analysis (EDA); sea surface monitoring; oil slick type differentiation; oil seep; oil spill; remote sensing; Synthetic Aperture Radar (SAR); RADARSAT-2

1. Introduction

Oil floating on the surface of the ocean is commonly detected using instruments flying onboard airplanes or satellites [1]. Currently, the most useful systems to detect oil-contaminated areas at sea are active radars, i.e., Synthetic Aperture Radars (SAR) operating in the microwave region of the electromagnetic spectrum [2]. As some environmental phenomena can generate signatures similar to oil in SAR imagery, this technology may yield false positives, in which the non-unique signature of oil caused by false targets can induce to ambiguous interpretations [3]. The so-called "look-alike features (or radar look-alikes)" range from atmospheric phenomena (e.g., rain cells) and oceanographic features (e.g., upwelling zones), amongst others [4]. Many scholars have described the use of SAR measurements to distinguish oil from look-alike features, and indeed, a broad scientific basis is devoted to the detection and identification of oil in SAR imagery, e.g., [5].

A supplementary application for satellite measurements is the recognition of the oil slick type. Herein, for the sake of brevity and simplicity, the term oil slick indicates the sea surface expression of oil naturally seeped out of the seafloor (i.e., oil seep) or spilled after man-made activities (i.e., oil spill). The latter accounts for operational, accidental, or illegal spills from platforms or ships—i.e., discharge of water containing petroleum products. Information about the oil slick type is an improvement requirement for several purposes, for example, environmental monitoring, emergency response, etc. [6]. Reliable surveillance based on satellite remote sensors capable of distinguishing the oil slick type can add accuracy to oil spill forecast systems, as well as enhance tridimensional mathematical models that backtrack (i.e., hindcast) oil seeps observed on the sea surface to its seafloor origin [7].

Environmental and economical issues associated with oil slicks are a constant concern to the oil and gas exploration and production industry. As such, the satellite synoptic view is an attractive option to distinguish the oil slick type that can lead to considerable scientific advances. From the standpoint of environmental monitoring programs, it can reduce ambiguities about the source of the observed oil (i.e., seeped or spilled). This can develop the relationship between governmental agencies and oil and gas exploration and production industry, thus reducing political uproar. From an economic standpoint, it can lead to offshore discoveries bringing invaluable information to explore active petroleum systems. Therefore, if the distinguishing of seeps and spills becomes possible, satellite information can directly assist in the search to find new oil fields in offshore exploration frontiers, a constant goal of a crucial sector for the world's economic development.

Oil seeps and oil spills are both products of mineral oil floating on the sea surface, therefore, it is expected that their surface signatures are fairly similar in satellite imagery. Peer-reviewed and grey literature present limited information available about oil slick type differentiation using satellite sensors; to this matter, the reader is encouraged to see [8] and the vast list of references therein. As a result, the objective of the present research focuses on performing an Exploratory Data Analysis (EDA) to evaluate the use of SAR measurements to distinguish the sea surface expressions of oil seeps from oil spills to a useful level of confidence for systematic use.

In the course of achieving this objective, three scientific questions are sought about the sea surface expression of oil slicks observed in the Campeche Bay region (Gulf of Mexico—Figure 1), as determined by digital classification of satellite imagery:

1. Does seeped oil floating on the ocean surface have a SAR backscatter signature distinctive enough to distinguish it from anthropogenically-spilled oil?
2. Can the oil slick Size Information be used to distinguish seeps from spills?
3. Which characteristics lead to the generation of a system capable of distinguishing between seeped and spilled oil?

To address these matters, a long-term RADARSAT-2 dataset (2008–2012) is exploited to perform a data mining of selected oil slicks' characteristics. Therefore, an additional goal is devised to design innovative qualitative–quantitative Classification Algorithms to distinguish man-made from natural oil slicks.

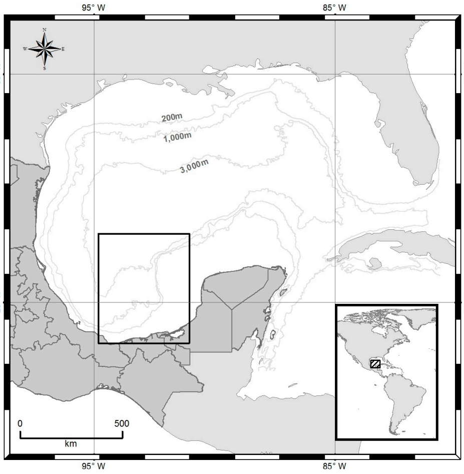


Figure 1. Gulf of Mexico highlighting the Campeche Bay region. The rectangle illustrates the region from which most oil slicks (98%) have been observed [9,10]. Isobaths of 200 m, 1000 m, and 3000 m are also shown. Courtesy of Adriano Vasconcelos.

2. Dataset

An environmental monitoring observed the occurrence of oil slicks for more than a decade (2000–2012) in Campeche Bay—this has been carried out by Pemex (Petróleos Mexicanos; Mexico City, Mexico) to support its decision-making processes, thus demonstrating the operational usefulness SAR measurements to execute effective oil slick mapping. Such monitoring was based on the interpretation of SAR scenes of both RADARSAT satellites analyzed by domain specialists with support of ancillary meteo-oceanographic data from Earth Observing System (EOS) sensors. All logged oil slicks underwent a post-processing validation, in which internal Pemex field reports were considered—i.e., the sea surface source recognized by the satellite image analyses were corroborated by additional field observations.

The long-term RADARSAT dataset exploited by Pemex was made available for this investigation and it is herein referred to as Campeche Bay Oil Slick Satellite Database (i.e., CBOS-Data). It is good to point out that the CBOS-Data does not account for look-alike features and that oil seepage sites on the sea surface were identified with at least three different observations about the same location, thus indicating oil seep cases. For a thorough description of the Pemex's operational environmental monitoring strategy and its produced dataset (i.e., CBOS-Data), the reader is encouraged to see a D.Sc. Thesis [8], a peer-reviewed article [9], and a conference paper [10]—these three references also provide a description of the occurrence and spatio-temporal distribution of the observed oil seeps and oil spills in the region.

To guarantee the quality of circumstances to cope with the objectives of the present research, our EDA only considers oil slicks observed between 2008 and 2012 with RADARSAT-2 (16-bit and VV-polarized—i.e., both ScanSAR Narrow beam modes: SCNA and SCNB). This mostly aims to circumvent technical specification differences (e.g., radiometric depth and polarization) between the two single-polarized RADARSAT measurements, as all analyzed RADARSAT-1 scenes are 8-bit and HH-polarized. Even though RADARSAT-2 provides a range of polarimetric information (i.e., linear cross-, dual parallel cross-, and full quad-polarized), our EDA focuses on using single linear parallel polarimetric SAR measurements from Pemex's provided information—i.e., CBOS-Data. As such, the dataset exploited in our EDA consists of 277 RADARSAT-2 scenes and 4916 oil slicks, of which 2021 (41%) have been identified as oil seeps and 2895 (59%) as oil spills. Herein, this dataset is referred to as Campeche Bay Oil Slick Modified Database: CBOS-MOD. The entire content of the CBOS-MOD has been used to perform the proposed data mining exercises, as well as to design the innovative qualitative-quantitative Classification Algorithms, thus ensuring best statistical reliability of our EDA.

3. Methods

The Methods' structure of our EDA is divided in two main Sections: Workable-Database Preparation (Section 3.1) and Multivariate Data Analysis (Section 3.2). Key concepts are emphasized providing enough detailed information to enable any knowledgeable scientist to replicate the results of our EDA with any SAR dataset containing identified oil slicks [8].

3.1. Workable-Database Preparation

3.1.1. RADARSAT Re-Processing

The RADARSAT-2 images, provided in SGF-format (SAR Georeferenced Fine Resolution), are given in uncalibrated grey-level count, i.e., Digital Numbers (DNs). These are enough for qualitative usage, but they are not recommended to cross-compare time series of SAR images [11,12]. Conversely, SAR backscatter coefficients are essential for quantitative analyses. The use of sigma-naught (σ_o), beta-naught (β_o), or gamma-naught (γ_o) permit the comparison of data acquired with the same sensor from different dates or beam modes, as well as the evaluation of data acquired with various sensors [13–15]. The conversion from DN to σ_o, β_o, and γ_o has been completed using the PCI Geomatica (version 2014, PCI Geomatics; Markham, ON, Canada). Various radiometric-calibrated image products are used in our EDA to describe the radar signal backscatter strength of individual oil slicks. Table 1 lists the twelve forms of σ_o, β_o, and γ_o that are given with the amplitude of the received radar beam (C_1) and in dB units (C_2), both with and without the application of the Frost filter with 3 × 3 window [16]. A thirteenth image product gives the incidence angle values per pixel: INC.ang.

Table 1. Radiometric-calibrated image products: twelve forms of SAR backscatter signature using sigma-naught (σ_o), beta-naught (β_o), and gamma-naught (γ_o), and the incidence angle (INC.ang).

Image Products	Frost Filter	σ_o	β_o	γ_o
C_1 in Amplitude	Without	SIG.amp	BET.amp	GAM.amp
	With	SIG.amp.FF	BET.amp.FF	GAM.amp.FF
C_2 in dB	Without	SIG.dB	BET.dB	GAM.dB
	With	SIG.dB.FF	BET.dB.FF	GAM.dB.FF
Incience angle (INC.ang)				

$C_1 = ([\{DN^2\} + B]/A)$. $C_2 = [(10 \times cc) \times Log_{10}(C_1)]$. DN: Digital Number. A: Range-dependent gain that varies per pixel in range direction. B: Constant offset nominally set to zero for SGF products. cc: 2 for the amplitude (or 1 for the intensity) of the received radar beam.

3.1.2. New Slick-Feature Attributes

Most new slick-feature attributes are bridged from elemental characteristics found in the literature; however, these have been suggested to distinguish oil slicks from look-alike features. Herein, they are explored to distinguish seeps from spills. A particular set of basic statistical measures is experimentally used to describe the oil slicks' characteristics. Four Attribute Types (Tables 2 and 3) are investigated:

- **1st) Contextual Information:** The Category attribute informs the oil slick type of each polygon: seeps or spills. Two other attributes provide the geographical location of the oil slick's centroid: latitude (cLAT) and longitude (cLONG). A pair of satellite overpass attributes provides imaging time (SARtime) and observation date (SARdate).
- **2nd) SAR Scene Descriptors:** The satellite scene is described by the imaging configuration (i.e., beam mode: Bmode) and the incidence angle (INC.ang). Regarding the latter, different basic statistical measures are also calculated, as shown when the 4th Attribute Type (i.e., SAR backscatter signature) are introduced.
- **3rd) Size Information:** From the two basic morphological features originally present in the CBOS-Data, i.e., area (Area) and perimeter (Per), seven ratios are derived to describe the

oil slicks' geometry, shape, and dimension (Table 2). The first is the area to perimeter ratio (AtoP). Fiscella et al. [17] suggest using the perimeter to area ratio (PtoA). Fiscella et al. [17] and Singha et al. [18] recommend using a dimensionless normalized perimeter to area ratio (PtoA.nor). While small PtoA.nor values are related to simple geometry, larger values come from oil slicks with more complex geometries [19]. A dimensionless complexity measure is given by [20]: COMPLEX.ind. In addition, Bentz et al. [21] uses another dimensionless descriptor to illustrate how compact (i.e., close to a circle) is a sea-surface feature: COMPACT.ind. Two other indices have been utilized by [22]: SHAPE.ind and FRACTAL.ind. While these indices yield values close to the unit for regular forms (i.e., circular or square, respectively), larger numbers represent form irregularity [23]. The total number of pixels inside each oil slick polygon (LEN) is also provided.

Table 2. Oil slick descriptors (53): 1st Attribute Type (Contextual Information: 6), 2nd Attribute Type (SAR Scene Descriptors: 37), and 3rd Attribute Type (Size Information: 10).

1st Attribute Type: Contextual Information §				
1	Category	Spill or Seep		Oil Slick Type *
2	Class	See [8]		
3	cLAT	Latitude (° N)		Spatial Location
4	cLONG	Longitude (° W)		
5	SARtime	Overpass Time		Temporal Location
6	SARdate	Overpass Date		
2nd Attribute Type: SAR Scene Descriptor				
1	Bmode § *	Beam mode		
2–37 **	INC.ang	Incidence angle of the radar beam		
3rd Attribute Type: Size Information (Geometry, Shape, and Dimension)				
1	LEN	Number of pixels inside the oil slick polygon.		
2	Area §	km^2		
3	Per §	km	Perimeter	
4	AtoP	km	Area to Per ratio	Area/Per
5	PtoA	km^{-1}	Per to Area ratio	Per/Area
6	PtoA.nor	*	Normalized PtoA	Per/[(2 × (Pi × Area))ˆ(1/2)]
7	COMPLEX.ind	*	Complexity Index	(Per^2)/Area
8	COMPACT.ind	*	Compact Index	(4 × Pi × Area)/(Per^2)
9	SHAPE.ind	km^{-1}	Shape Index	[0.25 × Per]/[Areaˆ(1/2)]
10	FRACTAL.ind	*	Fractal Index	[2 × ln(0.25 × Per)]/[ln(Area)]

§ Present in the Campeche Bay Oil Slick Satellite Database (i.e., CBOS-Data). * Dimensionless quantity. ** Includes the 36 statistical measures calculated for the 4th Attribute Type (Table 3).

- **4th) SAR Backscatter Signature:** All pixels within individual polygons are utilized to calculate the different basic statistical measures experimentally used to characterize the oil slick SAR backscatter signature (Table 3). These are separately calculated for the twelve radiometric-calibrated image products (Table 1). As suggested by [18,21], an arithmetic mean (AVG) of all pixels of each oil slick is computed. Three other central tendency representations are also considered: median (MED), mode (MOD), and mid-mean (MDM). To analyze the SAR backscatter signature spread, six dispersion measures are considered: standard deviation (STD), coefficient of dispersion (COD), variance (VAR), total range (RNG), average absolute deviation (AAD), and median absolute deviation (MAD). The coefficient of variation (COV) evaluates the relative relationship between dispersion and central tendency, thus comparing the degree of variation of data with different units and different meanings. COV originally involves the ratio between STD and AVG, but herein other pair-values are also given (Table 3). Different authors recommend exploring COV, e.g., [18,24] refer to it as power-to-mean ratio, whereas [20] describe it as the oil slick's homogeneity, and for [21] it depicts the oil slick's heterogeneity. However, Miranda [25] emphasizes that oil slicks with different spatial structures (i.e., with completely different pixel configuration) can have identical average or median values and the same standard deviation. Minimum (MIN) and maximum (MAX) values of the pixels inside each polygon are also used as supplementary quantities.

Information from the area surrounding oil slicks may play an essential role in the classification between oil slicks and look-alike features [17,20,26]. Nonetheless, herein, no information from the background oil-free surface around oil slicks has been taken into consideration. This decision aims to evaluate as simple as possible range of variables solely accounting for information within the oil slicks polygons' limits.

Table 3. Oil slick descriptors (432): 4th Attribute Type (SAR Backscatter Signature). Basic statistical measures (36) used to describe the oil slicks' characteristics. Separately calculated for the different forms (12) of σ_o, β_o, and γ_o: shown in Table 1: SIG.amp, SIG.amp.FF, SIG.dB, SIG.dB.FF, BET.amp, BET.amp.FF, BET.dB, BET.dB.FF, GAM.amp, GAM.amp.FF, GAM.dB, GAM.dB.FF.

4th Attribute Type: SAR Backscatter Signature				**36 × 12 = 432**
Basic Statistical Measures				**12 × 12 = 144**
1	1–12	**AVG**	Average	Central Tendency (4)
2	13–24	**MED**	Median	
3	25–36	**MOD**	Mode	
4	37–48	**MDM** *	Mid-Mean	
5	49–60	**STD**	Standard Deviation	Dispersion (6)
6	61–72	**COD** **	Coefficient of Dispersion	
7	73–84	**VAR**	Variance	
8	85–96	**RNG**	Total Range	
9	97–108	**AAD** ***	Average Absolute Deviation	
10	109–120	**MAD** ****	Median Absolute Deviation	
11	121–132	**MIN** §	Minimum	
12	133–144	**MAX**	Maximum	
COV = Coefficient of Variation				**24 × 12 = 288**
13	145–156	**COV.STD/AVG** ¥	**1st combined COV set** (STD divided by Central Tendency)	
14	157–168	**COV.STD/MED**		
15	169–180	**COV.STDMOD**		
16	181–192	**COV.STD/MDM**		
17	192–204	**COV.COD/AVG**	**2nd combined COV set** (COD divided by Central Tendency)	
18	205–216	**COV.COD/MED**		
19	217–228	**COV.COD/MOD**		
20	229–240	**COV.COD/MDM**		
21	241–252	**COV.VAR/AVG**	**3rd combined COV set** (VAR divided by Central Tendency)	
22	253–264	**COV.VAR/MED**		
23	265–276	**COV.VAR/MOD**		
24	277–288	**COV.VAR/MDM**		
25	289–300	**COV.RNG/AVG**	**4th combined COV set** (RNG divided by Central Tendency)	
26	301–312	**COV.RNG/MED**		
27	313–324	**COV.RNG/MOD**		
28	325–336	**COV.RNG/MDM**		
29	337–348	**COV.AAD/AVG**	**5th combined COV set** (AAD divided by Central Tendency)	
30	349–360	**COV.AAD/MED**		
31	361–372	**COV.AAD/MOD**		
32	373–384	**COV.AAD/MDM**		
33	385–396	**COV.MAD/AVG**	**6th combined COV set** (MAD divided by Central Tendency)	
34	397–408	**COV.MAD/MED**		
35	409–420	**COV.MAD/MOD**		
36	421–432	**COV.MAD/MDM**		

* MDM: Mean of the interquartile range. ** COD: Subtract the 1st from the 3rd interquartile, divided by their sum. *** AAD: Mean of the absolute difference of each value minus the mean. **** MAD: Median of the absolute difference of each value minus the median. § Negative Values Scaling eliminates 9 MIN variables. ¥ Original COV ratio.

3.1.3. Data Treatment Processes

The new slick-feature attributes (Tables 2 and 3) have both non-numeric values (i.e., qualitative variables) and numeric values (i.e., quantitative variables). The multivariate data analysis techniques applied herein (see Section 3.2.) requires variables to be quantitative descriptors. To guarantee that all slick-feature attributes have discrete or continuous numerical metric values with no categorical variables, four Data Treatment Processes are applied:

1. **Log_{10} Transformation:** Various non-linear transformations (e.g., square, cubic, or fourth root, as well as natural and base 10 logarithms) were tested to bring non-symmetric distributions to, or at least close to, a normal distribution pattern. Through the visual analyses of their histograms, Log_{10} had best results to stabilize the data, thus decreasing the effect of outliers and reducing skewness. The exception is FRACTAL.ind with its negative to positive range that was transformed as the cubic root.
2. **Negative Values Scaling:** Because log functions cannot be applied to negative values, all pixels inside the oil slicks were inspected before such transformation. This mostly relates to the SAR Backscatter Signature (4^{th} Attribute Type: Table 3), as the other Attribute Types all have positive values (except FRACTAL.ind). This specific Data Treatment Process consists of subtracting the minimum negative pixel value within the oil slick (PIXmin) from each pixel inside the oil slick (PIX) and adding one to it: PIXpos = [PIX − PIXmin + 1], where PIXpos is the new positive value. This simply adds two constants to each pixel, i.e., PIXmin and 1. The pixel with the minimum negative value becomes equal to 1, as PIX = PIXmin gives PIXpos = 1. Although this changes the values of the pixels inside some oil slicks, the relationship among pixel values within individual oil slicks is preserved [27,28].
3. **Ranging Standardization:** While some quantitative attributes are dimensionless (e.g., dB), others have incompatible units. To compare the oil slick characteristics in the subsequent data mining exercise of our EDA, a common scale is recommended [29–31]. Of the various available methods, there is no dearth of standardization; however, Milligan and Cooper [32] advocated that Ranging could be more effective than other methods (e.g., z-score). The Ranging standardization, besides bounding the magnitude of the attributes between zero (0) and one (1), also equalizes the variability of the attributes to a common scale: Xranged = [X − Xmin]/[Xmax − Xmin], where Xranged is the new Ranged value, X is the actual attribute value for a particular oil slick, Xmin and Xmax are, respectively, the minimum and maximum values of this attribute among all oil slicks [27]. For instance, if taking the SIG.amp.AVG value of one oil slick: subtract the minimum AVG for this radiometric-calibrated image product among all 4916 oil slicks, then the result is divided by the maximum AVG of all 4916 oil slicks minus the minimum AVG of all 4916 oil slicks. The Xranged variables assume non-negative values, and only one oil slick has Xranged = 0 and another one Xranged = 1, respectively when X = Xmin and when X = Xmax.
4. **Dummy Variables:** Certain qualitative attributes, such as string variables (Category and Bmode) and spatio-temporal variables (SARtime, SARdate, cLAT, and cLONG), have been converted to a specific indicator type: "dummy variable" [30]. For convenience, these have values of one (1) or zero (0), thus referring to its presence (Yes) or absence (No), respectively. These types of binary-coded variables attempt to decompose complex variabilities (or hidden knowledge) into more useful information.

3.2. Multivariate Data Analysis

3.2.1. Attribute Selection Methods

The attributes' relevance is explored for three reasons: to eliminate redundant (or irrelevant) variables, to select more representative variables, and to reduce the variable-hyperspace dimension. A lower degree of dependence from one attribute to another reduces the messiness of the subsequent data mining exercise of our EDA (see Section 3.2.). Two Attribute Selection Methods (i.e., R-mode), having variable selection strategies different in essence, are explored mostly based on their simplicity:

- **Unweighted Pair Group Method with Arithmetic Mean (UPGMA):** The resemblance between variables is revealed on a semi-automatic manner, in which an equally weighted pair wise group relationship among variables is used to form groups [27]. A variable is attributed to a group that has a larger similarity measure, i.e., Pearson's r correlation coefficient, with all other variables of that same group [30,31]. A free scientific data analysis software package (PAST: PAleontological STatistics version 2.17c; Oslo, Norway) was utilized [33]. Rooted-tree dendrograms (i.e., diagrams of relationships) with a bootstrapping of 100 replicates help to find hierarchical relationships (i.e., affinity) among variables. Due to the large subjectivity while using dendrograms to form groups, two user-defined thresholds are explored [34]. Their choice aims to guarantee repeatability, as different thresholds give different groups of variables. The first threshold subjectively uses the Cophenetic Correlation Coefficient (CCC) that is automatically calculated in PAST to specify the clusters; its value varies depending upon dataset [35,36]. The second threshold is simply the fixed similarity value of 0.5. These two thresholds are used to draw horizontal lines (i.e., phenon lines) across the dendrograms, and when such lines cross a branch, a group is formed; from each group of similar variables only one variable is selected [37,38]. The Attribute Selection occurs twice for each UPGMA implementation, one for each phenon line, i.e., CCC and 0.5. This variable selection gives preference to an arbitrary user-defined strategy based on the simplicity of their meanings and on their calculation form, in which simple variables are preferred over complicated ones—e.g., basic variables in lieu of ratios (e.g., Area versus AtoP), C_1 (amplitude) compared with C_2 (dB), SAR backscatter signature without rather than with Frost filter, and central tendency is preferable to dispersion metrics. The variables choice follows the order presented on Tables 1–3.
- **Correlation-Based Feature Selection (CFS):** The CFS Attribute Selection is a fully automated process based on a heuristic variable selection strategy [39,40]. The "Merit" of various groups of selected variables is evaluated to select attributes with low inter-correlation but highly correlated to the categories being distinguished [41,42]. In essence, a Merit is calculated as a measure of the usefulness of the selected best possible group of attributes. A scientific machine learning open source package (WEKA: Waikato Environment for Knowledge Analysis version 3.6.12; Waikato, New Zealand) was utilized. The user specifies an evaluation function and a search algorithm [43–45]. The former is the method by which the groups of attributes are evaluated (Backward Sequential Selection: CfsSubsetEval), and the latter improves the evaluation function (best-first searching strategy: BestFirst).

A summary of the analyzed 11 original sets, as well as the 33 optimal subsets of selected variables, i.e., UPGMA (CCC and 0.5) and CFS, is presented in Figure 2. These are collectively referred to as 44 data sub-divisions. It is expected that these particularly fruitful attribute-wise evaluations reveal hidden complexities into information capable of distinguishing the oil slick type.

Although literature suggests not using images given in DNs to compare SAR images [11,12,46], herein, an assessment explores DNs to quantify the consequences of using DNs. Only Frost filter and Antenna Pattern Compensation (APC) were applied. DN signatures are expressed by the same set of basic statistical measures calculated experimentally used to describe SAR backscatter signature (4^{th} Attribute Type: Table 3), with the same Data Treatment Processes.

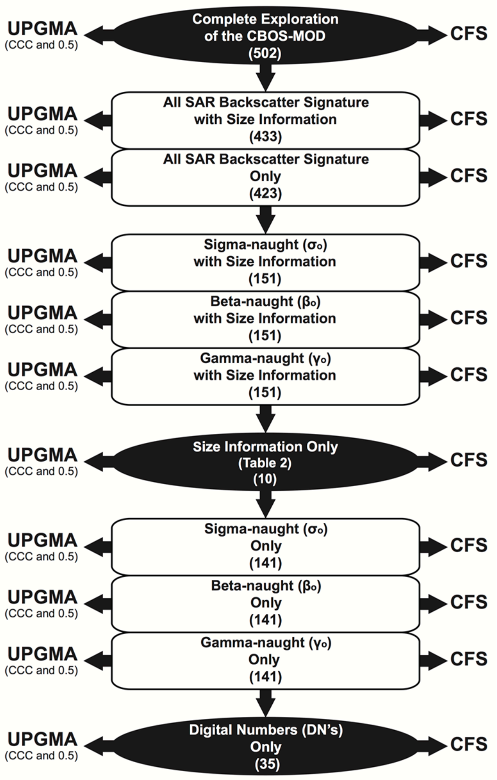

UPGMA (CCC and 0.5)
Complete Exploration of the CBOS-MOD (502)
CFS
All SAR Backscatter Signature with Size Information (433)
All SAR Backscatter Signature Only (423)
Sigma-naught (σₒ) with Size Information (151)
Beta-naught (βₒ) with Size Information (151)
Gamma-naught (γₒ) with Size Information (151)
Size Information Only (Table 2) (10)
Sigma-naught (σₒ) Only (141)
Beta-naught (βₒ) Only (141)
Gamma-naught (γₒ) Only (141)
Digital Numbers (DN's) Only (35)

Figure 2. Summary of 44 data sub-divisions: 11 original sets (shown in the middle along with the number of variables; see also Table 5-1) and 33 optimal subsets of selected variables (shown in both sides). These have undergone two Attribute Selection Methods: UPGMA (Unweighted Pair Group Method with Arithmetic Mean) and CFS (Correlation-Based Feature Selection). The UPGMA implementation employs two user-defined thresholds to form groups on the analysis of dendrograms: CCC (Cophenetic Correlation Coefficient) and 0.5 (fixed similarity value). The CFS automatically provides the selected variables.

3.2.2. Principal Components Analysis (PCA)

PCA reduces the variable-hyperspace dimension, helping on the interpretation of large multivariate datasets. It also assures the fulfillment of the Discriminant Analysis pre-requisite by avoiding multi-collinearity among the explored hypothetical variables, i.e., Principal Components (PCs). PAST (version 3.06; Oslo, Norway) was utilized to run the PCAs, which were applied to the 44 data sub-divisions.

Even though some guidelines are available to decide the number of PCs to account for, there is no "rule of thumb" to determine the best "stopping rule" [47,48]. Two cut-offs were explored herein [49]. The first one is simply referred to as "Scree Plot", as it analyzes the scree plot, but the PC-selection occurs based on the random model curve of expected eigenvalues, i.e., "broken stick". The broken stick is also plotted on the scree plot and PCs falling below this curve, i.e., to its right side, are not considered [50,51]. Furthermore, if the broken stick curve crosses the bootstrapping eigenvalue error bars, i.e., is inside the 95% confidence interval, this PC is also not considered [30,52]. The second is a simpler and direct method based on the so-called Kaiser-Guttman criterion, thus referred to as "Kaiser Criterion", which discards any PC after a specific lower bound: eigenvalue of 1 [49,53]. Only one of these two cut-offs was used to select meaningful PCs. This occurred based on the analysis of the global picture of the percentages of variance concentrated on the PCs (see [8]). The score values of the selected meaningful PC were subjected to the subsequent data mining exercise of our EDA.

Loadings expressing the relationship between variables (rows) and PCs (columns) verify the importance (i.e., meaning) of the original variables on each selected PC, thus determining if there were variables with higher weight (i.e., loadings) influencing each PC. Bootstrapping error bars were plotted per variable, and those having bars not crossing the abscissa are deemed to influence the PC [52]. Peres-Neto et al. [47] suggest 1000 row-wise bootstrapping replicates for the eigenvalues 95% confidence interval; however, as some PCAs took more than 10 h to run, only 500 replicates were used.

3.2.3. Correlation Matrix

Correlation matrices have been used to verify residual inter-variable correlation that might still be present after the Attribute Selection and PCA. These matrices provide the parametric Pearson's r correlation coefficient, p (uncorrelated), to determine the relationship among all original values of the CBOS-MOD variables (no PCA applied), as well as among the scores of the selected PCs (with PCA applied).

3.2.4. Discriminant Function

Discriminant Analyses are linear combinations among the selected characteristics that seek the highest discriminating power to minimize the probability of miss-discrimination [31,49]. Discriminant Functions are mathematical features that arrange objects among reported groups [33], in which the dependent variable is the oil slick type (i.e., seep or spill). Each analyzed sample is a member of only one oil slick type, i.e., simple binary discrimination. In PAST (version 2.17c; Oslo, Norway), Discriminant Functions were tested using the 44 data sub-divisions as input.

The accuracy of the Discriminant Functions is assessed via Confusion Matrices, i.e., simple 2-by-2-tables, in which one can quantify misclassification chances by using a cross validation estimate [27,54]. Table 4 illustrates adapted Confusion Matrices following a wide-open character design proposed by [55,56]. Even though the Confusion Matrix is a well-known tool for analyzing classification performance, a two-fold reason is given to introduce basic details about its usage. Firstly, usually the nomenclature of these metrics may vary, e.g., [57–59], etc. Secondly, usually only the Overall Accuracy is reported, but [60] demonstrated that this may lead to a quality gap on the information provided that can mislead the user [61,62]. To this matter, various metrics should come into play to report the effectiveness of the Discriminant Functions. As a starting point, one should answer a first set of four questions focusing on the lines of Table 4, frequently referred to as Producer's Accuracy (PA) and Commission Error (CE). A second set of four questions is given to interpret the columns of Table 4, converging to what is usually called as User's Accuracy (UA) and Omission Error (OE).

3.2.5. Oil Slick Classification Algorithm

The Classification Algorithm design implies not only the understanding of the Overall Accuracy (i.e., Confusion Matrix diagonal), but it also requires a full line- and column-wise inspection of Table 4, i.e., Sensitivity-Specificity balance and a trade-off between the Positive and Negative Predictive Values. The former is obtained by looking into lines (PA/CE), whereas the latter is achieved exploring columns (UA/OE). Steadiness is sought among line and column information. This is because lines give the reference frame of what was previously known (i.e., spills and seeps): how many of the known oil slick samples are correctly (or incorrectly) identified? Conversely, the column reference frame indicates how many oil slick samples identified by the algorithm are correctly (or incorrectly) identified? While the first question (about lines: PA/CE) measures the success of identifying known samples, the second question (about columns: UA/OE) provides a metric of the success of the algorithm to identify oil slick samples.

Table 4. Confusion Matrices utilized to assess the Discriminant Function accuracy. Adapted from [55,56]. Producer's Accuracy (PA), Commission Error (CE), User's Accuracy (UA), and Omission Error (OE) are given as frequency of occurrence (§) and in percentages (¥).

§	Algorithm Outcome Seeps	Algorithm Outcome Spills			§	Algorithm Outcome Seeps	Algorithm Outcome Spills		
Actual Seeps	GOOD Seeps	BAD Seeps	All Actual Seeps		Actual Seeps	A	B	A + B	
Actual Spills	BAD Spills	GOOD Spills	All Actual Spills		Actual Spills	C	D	C + D	
	All Algorithm Seeps	All Algorithm Spills	All Actual Slicks			A+C	B+D	A + B + C + D	
				Overall Accuracy					Overall Accuracy
¥	Algorithm Outcome Seeps	Algorithm Outcome Spills			¥	Algorithm Outcome Seeps	Algorithm Outcome Spills		
Actual Seeps	Sensitivity	Actual Seeps	100%		Actual Seeps	$\frac{A \times 100}{A + B}$	$\frac{B \times 100}{A + B}$	100%	
Actual Spills	False Positive	Specificity	100%		Actual Spills	$\frac{C \times 100}{C + D}$	$\frac{D \times 100}{C + D}$	100%	
¥	Algorithm Outcome Seeps	Algorithm Outcome Spills			¥	Algorithm Outcome Seeps	Algorithm Outcome Spills		
Actual Seeps	Positive Predictive Value	Invert Neg Predictive Value			Actual Seeps	$\frac{A \times 100}{A + C}$	$\frac{B \times 100}{B + D}$		
Actual Spills	Invert Pos Predictive Value	Negative Predictive Value			Actual Spills	$\frac{C \times 100}{A + C}$	$\frac{D \times 100}{B + D}$		
	100%	100%				100%	100%		
PA/CE Q1)	**How many known oil seep samples are correctly identified?** Sensitivity: A, also given in percentage by (A × 100)/[A + B].								
PA/CE Q2)	**How many known oil spill samples are correctly identified?** Specificity: D, also given in percentage by (D × 100)/[C + D].								
PA/CE Q3)	**How many known oil seep samples are misidentified?** False Negative cases: B, also given in percentage by (B × 100)/[A + B]. Coupled with Sensitivity.								
PA/CE Q4)	**How many known oil spill samples are misidentified?** False Positive cases: C, also given in percentage by (C × 100)/[C + D]. Linked to Specificity.								
UA/OE Q1)	**How many oil seeps identified by the Function are indeed known oil seeps?** Positive Predictive Value: A, also given in percentage by (A × 100)/[A + C].								
UA/OE Q2)	**How many oil spills identified by the Function are indeed known oil spills?** Negative Predictive Value: D, also given in percentage by (D × 100)/[B + D].								
UA/OE Q3)	**Of samples identified by the Function as oil seeps, how many are oil spills?** Inverse of the Positive Predictive Value: C, also given in percentage by (C × 100)/[A + C].								
UA/OE Q4)	**Of samples identified by the Function as oil spills, how many are oil seeps?** Inverse of the Negative Predictive Value: B, also given in percentage by (B × 100)/[B + D].								

4. Results

The Results presented herein mirror the organization structure of the Methods.

4.1. Workable-Database Preparation

4.1.1. RADARSAT Re-Processing

Each processed RADARSAT-2 scene is about 27 GB, as it accounts for several image products (Table 1), and together, the 277 scenes occupy almost 7.5 TB of disk space.

4.1.2. New Slick-Feature Attributes

Typical values (i.e., represented by a set of basic qualitative-quantitative statistics: minimum, maximum, average, and standard deviation) of the slick-feature attributes are presented in [8], along with histograms with their frequency distributions.

4.1.3. Data Treatment Processes

The skewness of most data transformed data attributes is reduced after the application of the Negative Values Scaling and the Log_{10} Transformation. The exception is the FRACTAL.ind, which has a cubic root transformation applied to it, and cLAT and cLONG, which did not change much in comparison to the frequency distributions prior to the Data Treatment Processes.

The application of these processes has five main consequences: (1) The Negative Values Scaling gives all pixels inside individual oil slicks that had negative pixel values a positive value; (2) The MIN attribute of 9 radiometric-calibrated image products lost its meaning and were removed, as PIXpos of several oil slicks became 1, as PIX = PIXmin gives PIXpos = 1; (3) The frequency of distribution of all variables has been brought to a Gaussian distribution with the Log_{10} Transformation; (4) All qualitative variables are coded with values lying between 0 and 1 after the Ranging Standardization; and (5) The 7 qualitative attributes have been replaced by 33 new dummy variables that are binary-coded to 1 or 0.

Consequently, the CBOS-MOD that had 485 different oil slick descriptors (Table 2: 53 and Table 3: 432) has been further refined after these processes, thus accounting for 502 variables: $485 - 9 - 7 + 33$.

4.2. Multivariate Data Analysis

4.2.1. Attribute Selection Methods

Table 5-1 presents the CCC (UPGMA) and Merit (CFS) values of the 11 original sets, whereas Table 5-2 reviews the number of selected variables of the 33 optimal subsets. Even though CCC values are quite similar, there is a small variation on the separate investigations of σ_o, β_o, and γ_o (with and without Size Information) that gives a slight variation in their number of selected variables. A higher CCC value stands out when only Size Information is considered: 0.9143.

The CFS-Merit decreasing trend reveals important information about the oil slick type predictability (Table 5-1). The CFS implementation using the all CBOS-MOD variables (502) mostly selected categorical contextual dummy variables (13 out of 15) and its Merit is the highest as possible: 1.00. The considerable CFS-Merit drop (0.131) when such variables are removed (i.e., all SAR backscatter signature with Size Information: 433) depicts the complexity of the exploratory nature of this research: the use of radiometric and size variables to distinguish seeps from spills is indeed not an easy task.

If Size Information variables are involved, there is a higher CFS-Merit as compared to when they are not considered, e.g., all SAR backscatter signatures with (0.131) and without (0.103) size. The same pattern is observed on the separate analysis of each SAR backscatter signature, e.g., σ_o with size (0.127) versus only σ_o (0.099). When using only Size Information the CFS-Merit (0.112) is higher than when separately exploring the SAR backscatter signature (σ_o, β_o, and γ_o): 0.099, 0.099, and 0.097, respectively. These are strong indications that the sole use of radiometric variables to differentiate the oil slick type is indeed intricate, as well as suggests that by using the Size Information one has a somewhat better chance in differentiating seeps from spills. The CFS-Merit resulted from the analysis of DNs showed the worse value (0.059), thus agreeing with literature that does not recommend the use of DN values to cross-compare time series of SAR images [11,12,46].

The CFS selected attributes are presented on Table 6 (separate analyses of Size Information and DN variables) and on Table 7 (each SAR backscatter coefficients with and without Size Information). A series of other tables with the results of the several CFS implementations is presented in [8].

Table 5. Data mining summary of the 11 original sets and the 33 optimal subsets of selected variables proposed on Figure 3, collectively referred to as to as 44 data sub-divisions. 1) Values of UPGMA-CCC and CFS-Merit. 2) Number of variables. 3) Number of selected Principal Components (PCs). See also Table 9 for full PCA results.

11 Original Sets and 33 Optimal Subsets	1) Values of		2) Number of Variables:				3) Number of Selected PC's:			
	UPGMA	CFS	Original	UPGMA		CFS	Original	UPGMA		CFS
	(CCC)	(Merit)	Sets	(CCC)	(0.5)		Sets	(CCC)	(0.5)	
1. Complete Exploration of the CBOS-MOD	0.8175	1.000	502	59	46	15 #	10	21	18	§
2. ❖ With Size Information	0.8227	0.131	433	†	†	20	8	†	†	6
3. ❖ Only	0.8253	0.103	423	††	††	36	7	††	††	6
4. σ_o with Size Information	0.8262	0.127	151	26	12	26	6	7	3	7
5. β_o with Size Information	0.8031	0.129	151	25	12	15	6	7	3	5
6. γ_o with Size Information	0.8240	0.126	151	26	12	19	6	7	3	5
7. Size Information only	0.9143	0.112	10	6	5	3	2	2	2	2 ¥
8. Only σ_o	0.8319	0.099	141	20	7	24	5	5	2	4
9. Only β_o	0.8130	0.099	141	19	7	21	5	4	2	5
10. Only γ_o	0.8260	0.097	141	20	7	22	6	5	2	4
11. DNs only	0.8020	0.059	35	5	3	11	2	2	2	2

❖ All SAR Backscatter Signature (σ_o, β_o, and γ_o; See also Tables 1 and 3). σ_o Sigma-naught: SIG.amp, SIG.amp.FF, SIG.dB, and SIG.dB.FF. β_o Beta-naught: BET.amp, BET.amp.FF, BET.dB, and BET.dB.FF. γ_o Gamma-naught: GAM.amp, GAM.amp.FF, GAM.dB, and GAM.dB.FF. † Same as σ_o with Size Information. †† Same as σ_o only. § PCA not performed because most categorical are dummy variables. # 13 are categorical contextual dummy variables (see [8]). ¥ Because the Discriminant Function requires at least 2 variables. DNs: Digital Numbers.

The results of the dendrogram analysis are depicted in Figure 3. Figure 3 (1st row left) illustrates the UPGMA dendrogram exploring only Size Information: of the initial 10 attributes (Table 2), the CCC (0.9143) forms six groups and the fixed similarity gives five groups: 6 and 5 variables, respectively (Table 6). Similarly, Figure 3 (1st row right) presents assessment of the DN values: of the initial 35 attributes, the CCC (0.8020) and the fixed similarity thresholds found 5 and 3 groups (variables), respectively (Table 6). Figure 3 also depicts the dendrograms of the separate UPGMA analysis of σ_o (CCC = 0.8262), β_o (CCC = 0.8031), and γ_o (CCC = 0.8240) together with Size Information, whereas Table 7 presents an inventory with the UPGMA selected attributes. Carvalho [8] presents other dendrograms and lists the selected attributes of the 44 data sub-divisions.

4.2.2. Principal Components Analysis (PCA)

A comparison of Table 5-2,3 confirms the PCA dimensionality reduction. After analyzing the global picture of the percentages of variance concentrated on particular PCs among the 44 data sub-divisions a user-defined criterion was used to decide which of the two cut-offs should be used to select the meaningful PCs: if the "Scree Plot" PC concentrated less than 80% of the variance, the "Kaiser Criterion" PC was selected; otherwise, the Scree Plot PC was the chosen cut-off (see [8]). The number of selected PCs varies per data sub-divisions (Table 5-3).

Figure 4 (left panels) demonstrates the Scree Plot PC selection strategy that explores the broken stick analyzing only the Size Information. The 4th row of Figure 4 depicts the bootstrapping eigenvalue error bars touching the broken stick curve for the PCA of the CFS; as such, no PC has been selected with the Scree Plot. Although the Kaiser Criterion selected one PC concentrating 90% of the variance, the Discriminant Function requires at least 2 variables (i.e., in this case, two PCs); therefore, the first two PCs have been selected (Table 5-3). Carvalho [8] illustrates the full proficiency of the broken stick PCs selection strategy using the Scree Plot.

Figure 4 (middle panels) exemplifies scatterplots illustrating the relationship between the first two PCs resulted from the PCA exploring only Size Information: 10, 6, 5, and 3 variables (Tables 2 and 6). It is evident the considerable overlap between the scores of the two populations: seeps (green circles) and spills (red triangles). The same pattern is observed on all other scatterplots, not only for the relationship between the first two PCs, but also for all other combinations of PCs [8].

The top three rows of Figure 4 (middle panels) reveal a peculiar gap in the cloud of data points affecting both populations, i.e., spills and seeps. This is mostly due to the FRACTAL.ind and its bi-modal (multi-modal) frequency distribution [8]. Such gap does not exist in the CFS analysis (Figure 4: 4th row middle panel), in which FRACTAL.ind is not accounted for (Table 6). However, this gap is not as prominent on the scatterplots when the FRACTAL.ind is accounted for on the analysis of Size Information together with the original set of SAR backscatter signature variables, i.e., σ_o, β_o, or γ_o. This is probably because, in this case, each PC is influenced by a much larger number of variables: 433 or 151 (Tables 1 and 3).

Figure 4 (right panels) depicts the influence (i.e., meaning) of each variable to each PC quantified by the inspection of Loadings Plots considering only Size Information: 10, 6, 5, and 3 variables (Tables 2 and 6). Almost all variables largely influence the first PC, the exception is on the CFS (Figure 4: 4th row right) that has the perimeter (Per) being the sole variable to account for the variance on the first PC; the second PC has the opposite occurring. In fact, the only PC of all PCAs that had a preferable influence was indeed this case, as the loadings expressions of all PCs of the 44 data sub-divisions have shown that each PC receives the influence of every single variable [8].

Table 6. Attribute Selection results. Top: Only Size Information (Table 2: 10). UPGMA implementation: Variables in bold (5) were selected using the fixed similarity value of 0.5. See also Figure 3 (1st row left: CCC = 0.9143). CFS implementation: Variables in bold (1) was also selected with the UPGMA implementation. Bottom: Only Digital Numbers (DNs) variables (35). UPGMA implementation: Variables in bold (3) were selected using the fixed similarity value of 0.5. See also Figure 3 (1st row right: CCC = 0.8020). CFS implementation: Variables in bold (3) are the same as those selected with the UPGMA implementation.

Only Size Information	
UPGMA Selected Attributes	
1	LEN
2	AtoP
3	PtoA
4	COMPACT.ind
5	COMPLEX.ind
6	FRACTAL.ind
CFS Selected Attributes	
1	LEN
2	PER
3	SHAPE
Only DNs	
UPGMA Selected Attributes	
1	DN.AVG
2	DN.COD
3	DN.STD/MOD
4	DN.COD/AVG
5	DN.RNG/AVG
CFS Selected Attributes	
1	DN.AVG
2	DN.MDM
3	DN.STD
4	DN.COD
5	DN.RNG
6	DN.COD/MOD
7	DN.RNG/AVG
8	DN.RNG/MED
9	DN.RNG/MOD
10	DN.RNG/MDM
11	DN.MAD/MED

Table 7. Pool of attributes selected by the UPGMA and CFS implementations considering each SAR backscatter coefficient, i.e., sigma-naught (σ_o), beta-naught (β_o), and gamma-naught (γ_o), with and without the Size Information (*). Double-checked variables (**) have also been selected with the UPGMA fixed similarity value of 0.5. Figure 3 portrays the UPGMA dendrograms.

Selection Method		UPGMA			CFS			CFS		
Size Information		Without §			With			Without		
Attribute Inventory		σ_o	β_o	γ_o	σ_o	β_o	γ_o	σ_o	β_o	γ_o
1	LEN				*	*	*			
2	AREA				*	*	*			
3	PER				*	*	*			
4	AtoP									
5	PtoA				*		*			
6	PtoAnor				*					
7	COMPACT.ind				*	*	*			
8	COMPLEX.ind				*		*			
9	SHAPE.ind				*	*	*			
10	FRACTAL.ind				*	*	*			
1	amp.AVG	**	**	**						
2	amp.MOD				*			*		*
3	amp.STD	**	**	**						
4	amp.COD	**	**	**	*	*	*	*	*	*
5	amp.RNG				*	*	*	*	*	*
6	amp.MAX					*		*	*	
7	amp.STD/MOD									*
8	amp.STD/MDM							*		
9	amp.RNG/MED						*			*
10	amp.RNG/MOD								*	
11	amp.RNG/MDM				*			*		
12	amp.MAD/AVG								*	
13	amp.FF.AVG	*	*	*						
14	amp.FF.COD	**	**	**						
15	amp.FF.RNG				*			*	*	
16	amp.FF.VAR/AVG	*	*	*						
17	amp.FF.RNG/AVG		*		*			*	*	*
18	amp.FF.RNG/MDM									*
19	amp.FF.RNG/MOD							*		
20	dB.AVG	**	**	**						
21	dB.MED					*		*	*	*
22	dB.MOD	*	*	*						
23	dB.STD	*	*	*	*	*		*	*	*
24	dB.COD	**	**	**	*		*	*	*	*
25	dB.MAD	*	*	*	*		*	*	*	*
26	dB.STD/AVG				*			*	*	*
27	dB.STD/MED							*	*	*
28	dB.STD/MOD	*	*	*						
29	dB.VAR/AVG	*	*	*		*	*	*	*	*
30	dB.VAR/MED				*			*	*	*
31	dB.VAR/MOD				*			*	*	*
32	dB.RNG/AVG	*	*	*						
33	dB.RNG/MED				*		*	*		*
34	dB.RNG/MOD	*	*	*	*	*	*	*	*	*
35	dB.FF.AVG	**	**	**						
36	dB.FF.COD	*		*						
37	dB.FF.RNG				*		*	*	*	*
38	dB.FF.MIN	*		*						
39	dB.FF.MAX	*	*	*						
40	dB.FF.MAD				*	*	*	*	*	*
41	dB.FF.COD/AVG						*			*
42	dB.FF.COD/MED				*					
43	dB.FF.VAR/AVG		*							
44	dB.FF.RNG/AVG	*		*						
45	dB.FF.RNG/MOD							*	*	*
46	dB.FF.MAD/AVG					*			*	
47	dB.FF.MAD/MOD							*		
57	Total (**)	20 (7)	19 (7)	20 (7)	26	15	19	24	21	22

§ The UPGMA with Size Information is the same but with the attributes of Table 6.

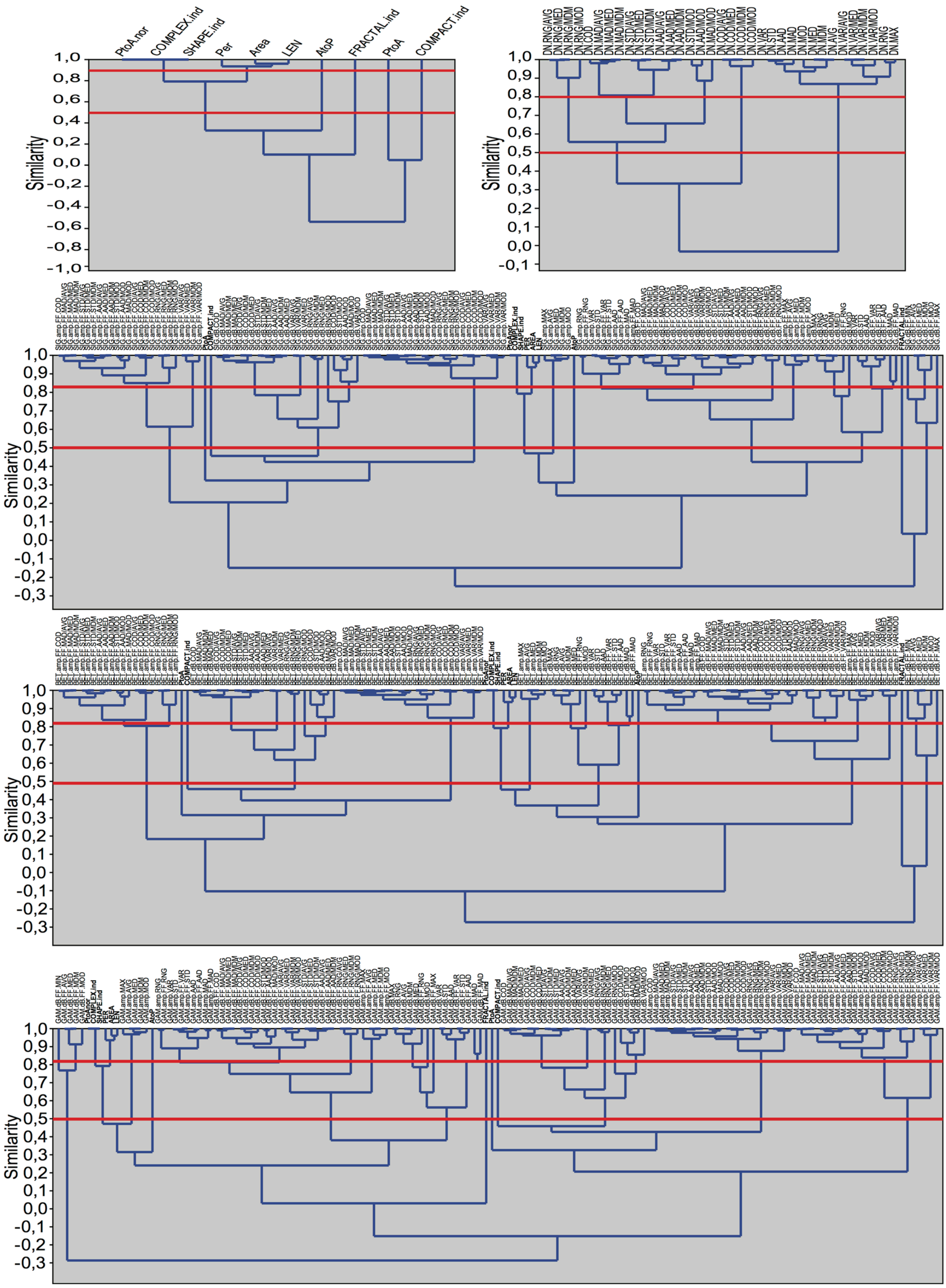


Figure 3. Rooted-tree dendrograms (i.e., diagrams of relationships) of the UPGMA implementation. 1st row left: only Size Information (10). 1st row right: only Digital Number (DN) variables (35). Size Information (10) together with: sigma-naught (σ_o; 141—2nd row), beta-naught (β_o; 141—3rd row), and gamma-naught (γ_o; 141—4th row). Bottom and top horizontal red lines respectively correspond to fixed similarity of 0.5 and Cophenetic Correlation Coefficient (CCC: 0.9143 (Size), 0.8020 (DN), 0.8262 (σ_o), 0.8031 (β_o), and 0.8240 (γ_o)).

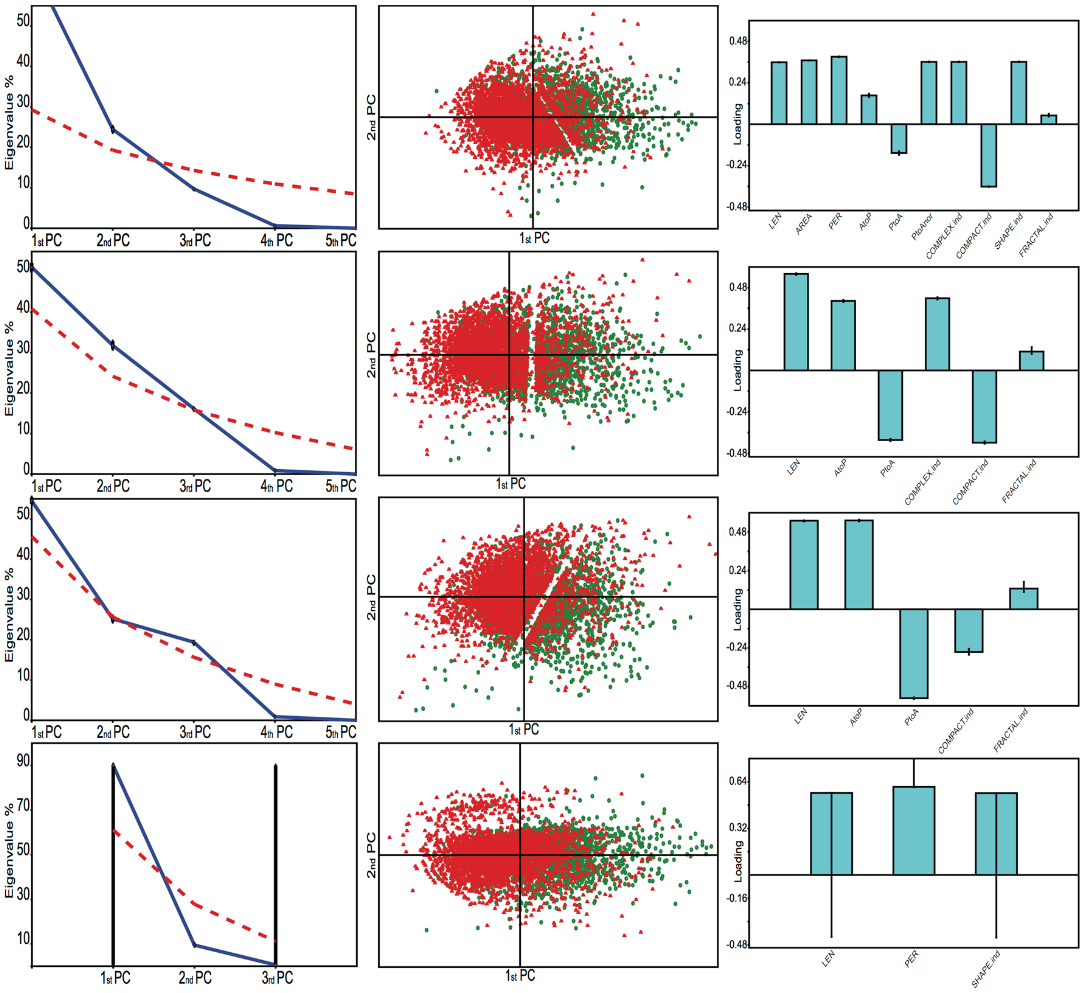


Figure 4. Results of the PCA exploring only Size Information. From top to bottom: original set (Table 2: 10 variables), UPGMA-CCC (Table 6: 6 variables), UPGMA-Fixed (Table 6: 5 variables), and CFS (Table 6: 3 variables). Left panels: Scree Plots (Eigenvalue x Principal Components (PC)), where the dashed red line corresponds to the broken stick curve. Bootstrapping eigenvalue error bars are shown as black lines: 2.5% and 97.5%. Middle panels: Scatterplots of the first two PCs, in which red triangles correspond to oil spills and green circles to oil seeps. Right panels: Loadings Plots.

4.2.3. Correlation Matrix

The correlation matrices were computed in PAST (version 2.17c; Oslo, Norway). Just about all information from the original values of the CBOS-MOD variables (no PCA applied: Table 5-2) showed residual inter-variable correlation among the variables of the 44 data sub-divisions. On the other hand, and as expected, the information from the scores of the selected PCs (with PCA applied: Table 5-3) presented no inter-variable correlation. This confirms that only the PCA information is carried to design the Classification Algorithms.

4.2.4. Discriminant Function

Table 8 summarizes the Discriminant Analyses' findings, where it is possible to notice the correspondence between Hotelling's t2 and Overall Accuracy. The former represents a measure of equality of the means of the groups been discriminated (i.e., spills and seeps), in which larger values correspond to better discrimination of the a priori informed groups. Likewise, the latter is the total classification efficiency that reports the success rate in correctly telling apart a priori known spills and seeps samples. Initially, only the Overall Accuracy is explored; see Section 4.2.5 for the full assessment of the other metrics. The Overall Accuracy somewhat matches the Hotelling's t2 effectiveness. Because the results about the complete exploration of the CBOS-MOD accounted for many categorical contextual dummy variables, its discrimination power almost reached 100%. This corroborates the quality of excellence of the explored dataset, i.e., CBOS-MOD. DN is again the worst, matching expectations [11,12,46]. A slightly better discrimination is observed while using only the Size Information.

Higher Hotelling's t2 is observed for the complete exploration of the CBOS-MOD that had undergone a major data reduction (Table 5-2: 502, 59, and 46 variables, and Table 5-3: 10, 21, and 18 PCs). This was anticipated as many categorical contextual dummy variables are accounted for. On the other hand, DNs (e.g., UPGMA-CCC: 163.0) has the lower Hotelling's t2 by far. The same result observed for its CFS-Merit (Table 5-1: 0.059); important to note that Hotelling's t2 and CFS-Merit are not comparable.

Column-wise, the Hotelling's t2 behavior is fairly similar to the one from the CFS-Merit (Table 5-1). The pattern shown by the sole analysis of Size Information presents higher Hotelling's t2 as compared to when these attributes are not included. Line-wise, higher Hotelling's t2 values are usually observed for the UPGMA-CCC optimal subset. Again, the sole use of Size Information has about the same effectiveness to discriminate seeps from spills, as when these attributes are combined with the SAR backscatter signature. This is an indication that the SAR backscatter signature is not adding much to the whole differentiation process. The UPGMA-Fixed optimal subsets tend to present lower Hotelling's t2, line-wise speaking. With the UPGMA-CCC optimal subset, noted above, it is possible to confirm that the number of variables has some influence on the discrimination process. This might be related not only to the amount of variables, but also to their quality, i.e., whether there is redundant correlated information or not.

Table 8. Discriminant Function results: Hotelling's t2 values and Overall Accuracy of the 44 data sub-divisions proposed on Figure 2.

Hotelling's t^2 Values	**Original Sets**	**UPGMA (CCC)**	**UPGMA (0.5)**	**CFS**
Complete Exploration	17,056.0	89,403.0	94,086.0	§
All SAR Backscatter Signature * with Size Information	1065.5	†	†	1044.5
All SAR Backscatter Signature * only	598.9	††	††	632.4
σ_o with Size Information	891.0	1123.5	769.8	1065.8
β_o with Size Information	908.7	1099.5	722.3	1106.2
γ_o with Size Information	863.1	1115.4	753.4	1028.4
Size Information only	1034.2	1064.7	1086.4	1085.9
σ_o Only	594.1	589.5	447.7	640.9
β_o Only	584.1	464.5	435.7	612.6
γ_o Only	579.2	574.7	432.2	606.7
Digital Numbers (DNs) only	138.0	163.0	138.4	210.5
Overall Accuracy	**Original Sets**	**UPGMA (CCC)**	**UPGMA (0.5)**	**CFS**
Complete Exploration	90.04%	99.98%	99.96%	§
All SAR Backscatter Signature * with Size Information	68.61%	†	†	68.71%
All SAR Backscatter Signature * only	63.63%	††	††	63.93%
σ_o with Size Information	67.64%	69.51%	65.46%	69.18%
β_o with Size Information	68.02%	69.57%	65.58%	69.55%
γ_o with Size Information	67.47%	69.32%	65.79%	68.59%
Size Information only	69.59%	70.00%	70.02%	70.22%
σ_o Only	63.28%	63.16%	63.26%	64.12%
β_o Only	63.53%	63.20%	63.02%	64.46%
γ_o Only	63.16%	63.02%	63.24%	64.18%
Digital Numbers (DNs) only	57.65%	57.89%	56.75%	59.48%

σ_o Sigma-naught: SIG.amp, SIG.amp.FF, SIG.dB, and SIG.dB.FF. β_o Beta-naught: BET.amp, BET.amp.FF, BET.dB, and BET.dB.FF. γ_o Gamma-naught: GAM.amp, GAM.amp.FF, GAM.dB, and GAM.dB.FF. * σ_o, β_o, and γ_o. See also Tables 1 and 3. † Same as σ_o with Size Information. †† Same as σ_o only. § Discriminant Function not performed because most categorical are dummy variables.

The UPGMA-CCC Hotelling's t2 values are somewhat equivalent to those of the CFS. Even though their variables are similar, they are not the same (e.g., Tables 6 and 7). This also corroborates the importance of reducing the dimensionality on the attribute-domain, for instance, by using Attribute Selection and PCA. However, these processes should be applied with caution so as not to lose information (i.e., underestimation) or to include noise (i.e., overestimation) [47–49].

4.2.5. Oil Slick Classification Algorithm

The full set of metrics shown in Table 4 is assessed to design the qualitative-quantitative Classification Algorithms, which are simple-to-use and simple in their formulation. Because the UPGMA-Fixed showed lower discriminating power than that of UPGMA-CCC and CFS on the initial Overall Accuracy analysis (Table 8), only the latter two are presented.

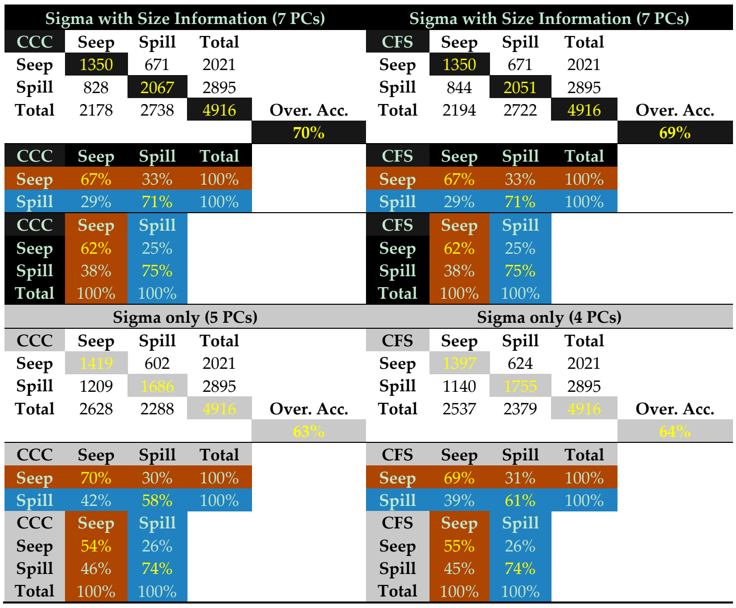
The Discriminant Function results using selected PCs from all SAR backscatter signature variables (Table 3: σ_o, β_o, and γ_o) analyzed with and without Size Information (Table 2) are presented in Table 9. A two-fold encouraging outcome is revealed: (1) The influence of Size Information on the discrimination between spills and seeps: if comparing the top with the lower tables, better results are observed when these variables are accounted for (top) than when they are not (lower); and (2) It confirms that the Overall Average (64%) may not inform the real discrimination performance; for instance, Specificity (59%) and Positive Predictive Value (54%) are not very effective, thus agreeing with [60] that the user should be aware of the real accuracy of the algorithm by examining different metrics.

Table 9. Confusion Matrices of the Discriminant Functions using selected Principal Components (PCs): all SAR backscatter signature variables (Table 3: σ_o, β_o, and γ_o) analyzed together with (top) and without (bottom) the Size Information (Table 2). Left: CCC (Cophenetic Correlation Coefficient). Right: CFS (Correlation-Based Feature Selection). See Table 4 for support.

All SAR Backsc. Sig. with Size Info. (7 PCs)					All SAR Backsc. Sig. with Size Info. (6 PCs)				
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	1395	626	2021		Seep	1324	697	2021	
Spill	917	1978	2895		Spill	841	2054	2895	
Total	2312	2604	4916	Over. Acc.	Total	2165	2751	4916	Over. Acc.
				69%					69%
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	69%	31%	100%		Seep	66%	34%	100%	
Spill	32%	68%	100%		Spill	29%	71%	100%	
CCC	Seep	Spill			CFS	Seep	Spill		
Seep	60%	24%			Seep	61%	25%		
Spill	40%	76%			Spill	39%	75%		
Total	100%	100%			Total	100%	100%		
All SAR Backscatter Signature only (5 PCs)					All SAR Backscatter Signature only (6 PCs)				
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	1430	591	2021		Seep	1375	646	2021	
Spill	1197	1698	2895		Spill	1127	1768	2895	
Total	2627	2289	4916	Over. Acc.	Total	2502	2414	4916	Over. Acc.
				64%					64%
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	71%	29%	100%		Seep	68%	32%	100%	
Spill	41%	59%	100%		Spill	39%	61%	100%	
CCC	Seep	Spill			CFS	Seep	Spill		
Seep	54%	26%			Seep	5 5%	27%		
Spill	46%	74%			Spill	45%	73%		
Total	100%	100%			Total	100%	100%		

The results of the separate analyses of σ_o variables with and without Size Information are presented in Table 10. The improvement observed when Size Information is considered is quite similar to the one observed in Table 9. This reinforces the fact that there is a benefit to using these variables, as all metrics evaluated are improved when they are accounted for. Confusion Matrices for β_o and γ_o with similar results are presented in [8].

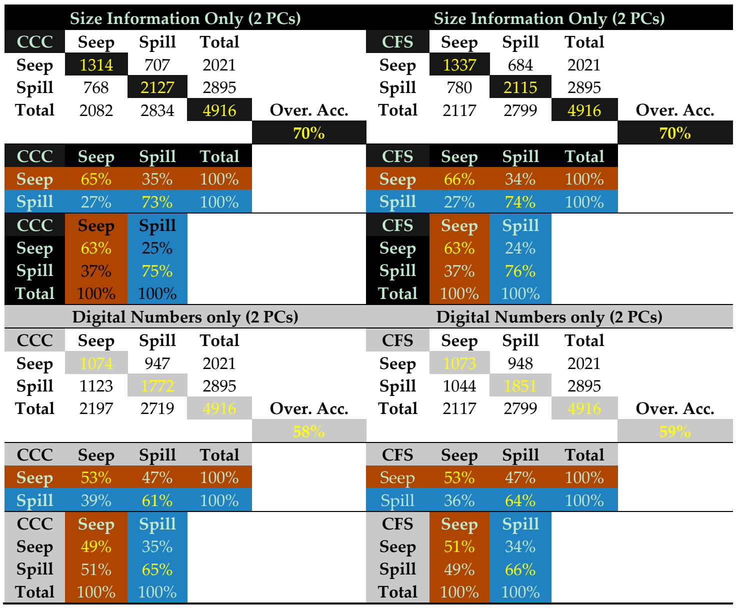
Table 10. Confusion Matrices of the Discriminant Functions using selected Principal Components (PCs): only sigma-naught (σ_o) variables analyzed together with (top) and without (bottom) the Size Information (Table 2). Left: CCC (Cophenetic Correlation Coefficient). Right: CFS (Correlation-Based Feature Selection). Carvalho [8] presents similar Confusion Matrices for the separate analysis of beta-naught (β_o) and gamma-naught (γ_o). See Table 4 for support.



Sigma with Size Information (7 PCs)					Sigma with Size Information (7 PCs)				
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	1350	671	2021		Seep	1350	671	2021	
Spill	828	2067	2895		Spill	844	2051	2895	
Total	2178	2738	4916	Over. Acc.	Total	2194	2722	4916	Over. Acc.
				70%					69%
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	67%	33%	100%		Seep	67%	33%	100%	
Spill	29%	71%	100%		Spill	29%	71%	100%	
CCC	Seep	Spill			CFS	Seep	Spill		
Seep	62%	25%			Seep	62%	25%		
Spill	38%	75%			Spill	38%	75%		
Total	100%	100%			Total	100%	100%		
Sigma only (5 PCs)					Sigma only (4 PCs)				
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	1419	602	2021		Seep	1397	624	2021	
Spill	1209	1686	2895		Spill	1140	1755	2895	
Total	2628	2288	4916	Over. Acc.	Total	2537	2379	4916	Over. Acc.
				63%					64%
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	70%	30%	100%		Seep	69%	31%	100%	
Spill	42%	58%	100%		Spill	39%	61%	100%	
CCC	Seep	Spill			CFS	Seep	Spill		
Seep	54%	26%			Seep	55%	26%		
Spill	46%	74%			Spill	45%	74%		
Total	100%	100%			Total	100%	100%		

The Discriminant Functions results for separate analysis of Size Information (best) and of DNs variables (worst) are presented in Table 11. A two-fold positive outcome is evident: (1) The relevancy of Size Information to distinguish oil seeps from oil spills on the sea surface of the Campeche Bay region; and (2) The corroboration with literature (e.g., [11,12,46] that DN images should not be used to compare time-series of SAR imagery.

Table 11. Confusion Matrices of the Discriminant Functions using selected Principal Components (PCs). Top: Size Information only (Table 2). Bottom: Digital Numbers (DNs) only. Left: CCC (Cophenetic Correlation Coefficient). Right: CFS (Correlation-Based Feature Selection). See Table 4 for support.



Size Information Only (2 PCs)					Size Information Only (2 PCs)				
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	1314	707	2021		Seep	1337	684	2021	
Spill	768	2127	2895		Spill	780	2115	2895	
Total	2082	2834	4916	Over. Acc.	Total	2117	2799	4916	Over. Acc.
				70%					70%
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	65%	35%	100%		Seep	66%	34%	100%	
Spill	27%	73%	100%		Spill	27%	74%	100%	
CCC	Seep	Spill			CFS	Seep	Spill		
Seep	63%	25%			Seep	63%	24%		
Spill	37%	75%			Spill	37%	76%		
Total	100%	100%			Total	100%	100%		
Digital Numbers only (2 PCs)					Digital Numbers only (2 PCs)				
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	1074	947	2021		Seep	1073	948	2021	
Spill	1123	1772	2895		Spill	1044	1851	2895	
Total	2197	2719	4916	Over. Acc.	Total	2117	2799	4916	Over. Acc.
				58%					59%
CCC	Seep	Spill	Total		CFS	Seep	Spill	Total	
Seep	53%	47%	100%		Seep	53%	47%	100%	
Spill	39%	61%	100%		Spill	36%	64%	100%	
CCC	Seep	Spill			CFS	Seep	Spill		
Seep	49%	35%			Seep	51%	34%		
Spill	51%	65%			Spill	49%	66%		
Total	100%	100%			Total	100%	100%		

Because the Multivariate Data Analysis revealed the use of Size Information (i.e., 3rd Attribute Type: Table 2) as best to distinguish the oil slick type, a separate analysis explores the only two basic morphological features originally present in the CBOS-Data: area and perimeter. The same Data Treatment Process has been applied preceding the PCA, and two PCs have been selected prior the Discriminant Function. Table 12 presents quite similar results to those shown in Tables 9–11, thus demonstrating that the sole and simple use of area and perimeter is equally capable of distinguishing human-related oil spills from naturally-occurring oil seeps to a useful accuracy. Therefore, there is no need to evoke more complicated variables for distinguishing the oil slick type.

Table 12. Confusion Matrices of the Discriminant Functions using selected Principal Components (PCs): sole and simple use of two specific oil slicks Size Information, i.e., area and perimeter. See also Table 4 for further information about Confusion Matrixes. See Table 4 for support.

Area and Perimeter (2 PCs)				
Size	**Seep**	**Spill**	**Total**	
Seep	1314	707	2021	
Spill	794	2102	2895	
Total	2108	2808	4916	**Over. Acc.**
				69%
Size	**Seep**	**Spill**	**Total**	
Seep	65%	35%	100%	
Spill	27%	73%	100%	
Size	**Seep**	**Spill**		
Seep	62%	25%		
Spill	38%	75%		
Total	100%	100%		
PA/CE Q1)	65%	Sensitivity		
PA/CE Q2)	73%	Specificity		
PA/CE Q3)	35%	False Negative		
PA/CE Q4)	27%	False Positive		
UA/OE Q1)	62%	Positive Predictive Value		
UA/OE Q2)	75%	Negative Predictive Value		
UA/OE Q3)	38%	Inv. of the Pos. Pred. Val.		
UA/OE Q4)	25%	Inv. of the Neg. Pred. Val.		

5. Discussion

Even though the three SAR backscatter coefficients can easily be transferred from one to another [8], the pool of attributes selected by the UPGMA and CFS implementations indicate the utility of σ_o, β_o, and γ_o is not exactly the same (Table 7). It is also true that because σ_o is directly related to the reflectivity of the ground horizontal range plane, it is a common SAR backscatter coefficient considered in oil slick detection; see [8] and references therein. Nonetheless, the data mining exercise of our EDA has demonstrated that σ_o, β_o, and γ_o present differences among each other (Table 7) and that it is worth analyzing them separately.

Two aspects agreed by the remote-sensing scientific community regarding the use of geometry, shape, and dimension attributes in the use of SAR measurements to identify oil slicks are: (1) Ship spills have distinct Size Information from other oil spills [63]; and (2) These attributes are crucial to distinguish oil slicks from look-alike features [64,65]. Thus, considering the exploratory nature of our Classification Algorithms, it should be emphasized that only a small fraction of the CBOS-Data (3.9%) and of the CBOS-MOD (3.2%) represents ship spills. In addition, the sole use of only two basic oil slicks' morphological features (i.e., area and perimeter) is an innovative way to distinguish oil from oil, i.e., natural from man-made oil slicks—a task poorly documented in the scientific literature—peer-reviewed, as well as grey literature; see [8] and references therein.

The data mining exercise of our EDA promotes a novel idea bridging petroleum pollution and remote sensing research, thus paving the way for further investigations using the satellite synoptic view to express geophysical differences between spilled and seeped oil observed on the sea surface. Despite the substantial amount of work performed in the data mining exercise of our EDA that uses standard multivariate data analysis techniques applied to SAR measurements for the successful distinguishment of sea surface expressions of naturally-occurring oil seeps from human-related oil spills in the Campeche Bay region [8], a list of nine main suggestions that may be further explored to achieve improved accuracies for systematic use is given:

- Information from the area surrounding the oil slicks, i.e., background oil-free surface, could come into play, e.g., damping ratio [66];
- Information about radar beam incidence angles and environmental configurations (e.g., wind conditions) could add knowledge to the differentiation process;
- Fractal dimension (e.g., box-counting or dynamic fractal approaches) could be used to measure, analyze, and classify the oil slicks textures and shapes [67,68];
- The number of individual non-contiguous parts forming oil slick polygons could be explored along with other contextual properties, such as the distance between centroid and the shoreline or the distance from a possible point source, i.e., oil rig complex or oil seep site location on the sea surface [21];
- Other methods instead of Ranging Standardization could be applied during the Data Treatment Process, e.g., z-score [30];
- As an alternative to the proposed automatic Attribute Selection Methods to select groups and variables, visual analyses of the UPGMA dendrograms could be instituted, or even other custom methods, e.g., Ward's [30];
- Different "stopping rule" methods could select meaningful PCs [47,48,69];
- Artificial Neural Networks (ANN), largely used to distinguish between oil slicks and look-alike features (e.g., [70–72]), indeed could be tested to differentiate oil seeps from oil spills using variables from the Attribute Selection Methods used herein, e.g., area and perimeter; and;
- Clustering Analyses (e.g., K-means) could also reveal natural groupings among oil slicks, thus supporting the oil slick type differentiation process.

In addition to the abovementioned, the data mining exercise of our EDA surely serves as an archetype for developing other ways to differentiate the oil slick type [73]. It also paves the way for full polarimetric investigations that could indeed search other ways to differentiate oil seeps from oil spills using SAR measurements (e.g., [74,75]).

6. Conclusions

The data mining exercise of our EDA has successfully demonstrated that it is feasible to use multivariate data analysis techniques, applied to SAR measurements, to distinguish the sea surface expressions of naturally-occurring oil seeps from human-related oil spills observed in Campeche Bay. Indeed, the use of an assorted pool of standard multivariate data analysis techniques (e.g., R-mode Correlation, Principal Components Analysis, and Discriminant Function) has been effectively applied to analyze the remote sensing library of a specific long-term dataset (2008–2012) that consists of 277 RADARSAT-2 scenes and 4916 oil slicks.

The answers to the three scientific questions about the sea surface expression of oil slicks are:

1. Yes. The seeped oil has SAR backscatter signatures (i.e., σ_o, β_o, and γ_o) distinctive enough to distinguish it from anthropogenically-spilled oil;
2. Yes. The Size Information (i.e., attributes describing the geometry, shape, and dimension of the oil slicks) can be used to distinguish seeps from spills, and in fact, the sole and simple use of area and perimeter can also distinguish natural from man-made oil slicks with an overall accuracy of about 70%; and;

3. The synergistic combination of various oil slick characteristics leads to systems capable of distinguishing between seeped and spilled oil, in which different combinations of variables promote similar differentiation; however, the one leading to the most effective classification is represented by the sole and simple use of area and perimeter (70%), whereas the one with the worst capabilities have variables expressed in Digital Numbers (DNs).

The proposed oil slick type Classification Algorithms of our EDA have useful accuracies, and range from uncomplicated algorithms combining the synergy of different variables to very simple ones using only SAR backscatter signature or only Size Information. In fact, the effective differentiation between oil seeps and oil spills achieved using only basic morphological features, such as the area and the perimeter of oil slicks, as determined by digital classification of satellite imagery, is a relevant outcome directly contributing to the remote sensing research, as well as to the activities of the oil and gas exploration and production industry. The field of environment monitoring also benefits from the outcomes of the data mining exercise conducted during our EDA. Conclusively, if the techniques presented herein are capable of distinguishing two types of oil observed on the sea surface (i.e., naturally-occurring oil seeps from human-related oil spills) based on RADARSAT-2 measurements, it is likely they can also produce good results if applied to the problem of discriminating oil from look-alike features (for instance, algal blooms or low wind zone signatures), indeed, solely using SAR measurements.

Acknowledgments: We are thankful to Pemex and to MDA Geospatial Services for making available the high-quality RADARSAT dataset, as well as to the three anonymous reviewers that greatly contributed to improve the manuscript with thoughtful criticism. The Human Resource Program (PRH) of the National Agency of Petroleum, Natural Gas, and Biofuels (ANP) provided the financial support for this research.

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

References

1. Jackson, C.R.; Apel, J.R. *Synthetic Aperture Radar Marine User's Manual*; NOAA/NESDIS; Office of Research and Applications: Washington, DC, USA, 2004. Available online: http://www.sarusersmanual.com (accessed on 13 November 2017).
2. Chan, Y.K.; Koo, V.C. An introduction to synthetic aperture radar (SAR). *Prog. Electromagn. Res. B* **2008**, *2*, 27–60. [CrossRef]
3. Espedal, H.A.; Johannessen, O.M.; Knulst, J. Satellite detection of natural films on the ocean surface. *Geophys. Res. Lett.* **1996**, *23*, 3151–3154. [CrossRef]
4. Johannessen, O.M.; Espedal, H.A.; Jenkins, A.J.; Knulst, J. SAR surveillance of ocean surface slicks. In Proceedings of the Second ERS Application Workshop, London, UK, 6–8 December 1995; pp. 187–192.
5. Topouzelis, K. Oil spill detection by SAR Images: Dark formation detection, feature extraction and classification algorithms. *Sensors* **2008**, *8*, 6642–6659. [CrossRef] [PubMed]
6. Mchugh, S.L. Satellite Synthetic Aperture radar in the Prosecution of Illegal Oil Discharges. Master's Thesis, Memorial University of Newfoundland, St. John's, NL, Canada, 2009; p. 122.
7. Mano, M.F.; Beisl, C.H.; Siqueira, C.Y.S.; Pereira, J.S. Evaluation of remote technologies applied to natural seep mapping and their impact in oil exploration. In Proceedings of the Rio Oil & Gas Expo and Conference, Rio de Janeiro, Brazil, 15–18 September 2014; p. 7.
8. Carvalho, G.A. Multivariate data Analysis of Satellite-Derived Measurements to Distinguish Natural from Man-Made Oil Slicks on the Sea Surface of Campeche Bay (Mexico). Ph.D. Dissertation, Universidade Federal do Rio de Janeiro (UFRJ), COPPE, Rio de Janeiro, Brazil, 2015; p. 285. Available online: http://www.coc.ufrj.br/index.php?option=com_content&view=article&id=5354:gustavo-de-araujo-carvalho (accessed on 13 November 2017).
9. Carvalho, G.A.; Minnett, P.J.; Miranda, F.P.; Landau, L.; Moreira, F. The use of a RADARSAT-derived long-term dataset to investigate the sea surface expressions of human-related oil spills and naturally-occurring oil seeps in Campeche Bay, Gulf of Mexico. *Can. J. Remote Sens.* **2016**, *42*, 307–321. [CrossRef]

10. Carvalho, G.A.; Landau, L.; Miranda, F.P.; Minnett, P.; Moreira, F.; Beisl, C. The use of RADARSAT-derived information to investigate oil slick occurrence in Campeche Bay, Gulf of Mexico. In Proceedings of the XVII Brazilian Remote Sensing Symposium (SBSR), INPE, João Pessoa, Brazil, 25–29 April 2015; pp. 1184–1191. Available online: http://www.dsr.inpe.br/sbsr2015/files/p0217.pdf (accessed on 13 November 2017).
11. Freeman, A. Radiometric calibration of SAR image data. In Proceedings of the XVII Congress for Photogrammetry and Remote Sensing (ISPRS), Washington, DC, USA, 2–14 August 1992; pp. 212–222.
12. El-darymli, K.; Mcguire, P.; Gill, E.; Power, D.; Moloney, C. Understanding the significance of radiometric calibration for synthetic aperture radar imagery. In Proceedings of the IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), Toronto, ON, Canada, 4–7 May 2014; p. 6. [CrossRef]
13. ESA (European Space Agency). 2014. Available online: http://earth.eo.esa.int/polsarpro/ (accessed on 1 August 2016).
14. Thakur, P.K. SAR data processing to extract backscatter response from various features. In Proceedings of the Symposium Tutorials on Polarimetric SAR Data Processing and Applications, International Society for Photogrametry and Remote Sensing (ISPRS), Hyderabad, India, 9–12 December 2014.
15. ASF (Alaska Satellite Facility). MapReady User Manual Remote Sensing Tool Kit, Engineering Group. 2015. Available online: https://media.asf.alaska.edu/uploads/pdf/mapready_manual_3.1.22.pdf (accessed on 13 November 2017).
16. Frost, V.S.; Stiles, J.A.; Shanmugan, K.S.; Holtzman, J.C. A model for radar images and its application to adaptive digital filtering of multiplicative noise. *IEEE Trans. Pattern Anal. Mach. Intell.* **1982**, *4*, 157–166. [CrossRef] [PubMed]
17. Fiscella, B.; Giancaspro, A.; Nirchio, F.; Pavese, P.; Trivero, P. Oil spill monitoring in the Mediterranean Sea using ERS SAR data. In Proceedings of the Envisat Symposium, ESA, Göteborg, Sweden, 16–20 October 2010; p. 9.
18. Singha, S.; Bellerby, T.J.; Trieschmann, O. Satellite Oil Spill Detection Using Artificial Neural Networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 9. [CrossRef]
19. Calabresi, G.; Del Frate, F.; Lichtenegger, I.; Petrocchi, A.; Trivero, P. Neural networks for the oil spill detection using ERS–SAR data. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '99), Hamburg, Germany, 28 June–2 July 1999; pp. 215–217. [CrossRef]
20. Solberg, A.H.S.; Storvik, G.; Solberg, R.; Volden, E. Automatic detection of oil spills in ERS SAR images. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 1916–1924. [CrossRef]
21. Bentz, C.M. Reconhecimento Automático de Eventos Ambientais Costeiros e Oceânicos em Imagens de Radares Orbitais. Ph.D. Dissertation, Universidade Federal do Rio de Janeiro (UFRJ), COPPE, Rio de Janeiro, Brazil, 2006; p. 115.
22. Pisano, A. Development of Oil Spill Detection Techniques for Satellite Optical Sensors and Their Application to Monitor Oil Spill Discharge in the Mediterranean Sea. Ph.D. Dissertation, Università di Bologna, Bologna, Italy, 2011; p. 146.
23. McgarigaL, K.; Marks, B.J. FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure. USDA For. Serv. Gen. Tech. Rep. PNW-351. 1994, p. 134. Available online: http://www.umass.edu/landeco/pubs/mcgarigal.marks.1995.pdf (accessed on 13 November 2017).
24. Solberg, A.H.S.; Volden, E. Incorporation of prior knowledge in automatic classification of oil spills in ERS SAR images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '97), Singapore, 3–8 August 1997; pp. 157–159.
25. Miranda, F.P. Reconnaissance Geologic Mapping of a Heavily-Forested Shield Area (Guiana Shield, Northwestern Brazil). Ph.D. Dissertation, University of Nevada, Reno, NV, USA, 1990; p. 176.
26. Fiscella, B.; Giancaspro, A.; Nirchio, F.; Pavese, P.; Trivero, P. Oil spill detection using marine SAR images. *Int. J. Remote Sens.* **2000**, *21*, 3561–3566. [CrossRef]
27. Sneath, P.H.A.; SokaL, R.R. *Numerical Taxonomy—The principles and Practice of Numerical Classification*; W.H. Freeman and Company: San Francisco, CA, USA, 2017; p. 573. ISBN 0-7167-0697-0.
28. Lane, D.M.; Scott, D.; Hebl, M.; Guerra, R.; Osherson, D.; Ziemer, H. Introduction to Statistics. 2015. Available online: http://onlinestatbook.com/Online_Statistics_Education.pdf (accessed on 13 November 2017).
29. Moita Neto, J.M.; Moita, G.C. Uma introdução à análise exploratória de dados multivariados. *Quím. Nova* **1998**, *21*, 467–469. [CrossRef]

30. Legendre, P.; Legendre, L. Developments in Environmental Modelling. In *Numerical Ecology*, Third ed.; Elsevier Science B.V.: Amsterdam, The Netherlands, 2012; p. 990, ISBN 978-0444538680.
31. Valentin, J.L. *Ecologia Numérica–Uma Introdução à Análise Multivariada de Dados Ecológicos*, Second ed.; Editora Interciência: Rio de Janeiro, Brazil, 2012; 153p, ISBN 978-85-7193-230-2.
32. Milligan, G.W.; Cooper, M.C. A study of standardization of variables in cluster analysis. *J. Classif.* **1988**, *5*, 181–204. [CrossRef]
33. Hammer, Ø.; Harper, D.A.T.; Ryan, P.D. PAST: PAleontological STatistics software package for education and data analysis. *Palaeontol. Electron.* **2001**, *4*, 9. Available online: http://palaeo-electronica.org/2001_1/past/issue1_01.htm (accessed on 13 November 2017).
34. Kelley, L.A.; Gardner, S.P.; Sutcliffe, M.J. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng. Des. Sel.* **1996**, *9*, 1063–1065. [CrossRef]
35. Farris, J.S. On the cophenetic correlation coefficient. *Syst. Biol.* **1969**, *18*, 279–285. [CrossRef]
36. Saraçli, S.; Dogan, N.; Dogan, I. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *J. Inequal. Appl.* **2013**, *2013*, 203. [CrossRef]
37. Sokal, R.R.; Rohlf, F.J. The Comparison of dendrograms by objective methods. *Taxon* **1962**, *11*, 33–40. [CrossRef]
38. NCSS (Number Cruncher Statistical System). Hierarchical Clustering and Dendrograms. NCSS Statistical Software. 2015, Chapter 445, p. 15. Available online: http://ncss.wpengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Hierarchical_Clustering-Dendrograms.pdf (accessed on 13 November 2017).
39. Hall, M.A.; Smith, L.A. Feature subset selection: a correlation based filter approach. In Proceedings of the Fourth International Conference on Neural Information and Intelligent Information Systems, Dunedin, New Zealand, 24–28 November 1997; pp. 855–858.
40. Hall, M.A.; Smith, L.A. Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper. In Proceedings of the 12th International FLAIRS Conference, Orlando, FL, USA, 1–5 May 1999.
41. Hall, M.A. Correlation-Based Feature Selection for Machine Learning. Ph.D. Dissertation, The University of Waikato, Department of Computer Science, Hamilton, New Zealand, 1999; p. 178.
42. Bouckaert, R.R.; Frank, E.; Hall, M.; Kirkby, R.; Reutemann, P.; Seewald, A.; Scuse, D. *WEKA Manual for Version 3–6-0*; The University of Waikato: Hamilton, New Zealand, 2008; p. 212. Available online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.153.9743 (accessed on 13 November 2017).
43. Aha, D.W.; Bankert, R.L. Feature selection for case-based classification of cloud types: An empirical comparison. In *Case-Based Reasoning: Papers From the 1994 Workshop (Technical Report WS-94–01)*; Advancement of Artificial Intelligence (AAAI): Menlo Park, CA, USA, 1994; pp. 106–112. Available online: https://www.aaai.org/Papers/Workshops/1994/WS-94-01/WS94-01-019.pdf (accessed on 13 November 2017).
44. Tetko, I.; Baskin, I.; Varnek, A. Tutorial on Machine Learning, Part 2: Descriptor Selection Bias. 2008. Available online: http://infochim.u-strasbg.fr/CS3/program/Tutorials/Tutorial2b.pdf (accessed on 13 November 2017).
45. Wilt, C.; Thayer, J.; Ruml, W.A. Comparison of Greedy Search Algorithms. In Proceedings of the Third Annual Symposium on Combinatorial Search (SoCS-10), Atlanta, GA, USA, 8–10 July 2010; pp. 129–136.
46. Thompson, A.A.; McLeod, I.H. The RADARSAT-2 SAR processor. *Can. J. Remote Sens.* **2004**, *30*, 336–344. [CrossRef]
47. Peres-Neto, P.R.; Jackson, D.A.; Somers, K.M. Giving meaningful interpretation to ordination axes: Assessing loading significance in principal component analysis. *Ecology* **2003**, *84*, 2347–2363. [CrossRef]
48. Peres-Neto, P.R.; Jackson, D.A.; Somers, K.M. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput. Stat. Data Anal.* **2005**, *49*, 974–997. [CrossRef]
49. Hair, J.F.; Anserson, R.E.; Tatham, R.L.; Black, W.C. Multivariate Data Analysis. In *Análise Multivariada de Dados*, Fifth ed.; Sant'Anna, A.S., Chaves Neto, A., Eds.; Bookman: Porto Alegre, RS, Brazil, 2005; ISBN 0-13-014406-7.
50. Cattell, R.B. The Scree Test For The Number of Factors. *Multivar. Behav. Res.* **1966**, *1*, 245–276. [CrossRef] [PubMed]
51. Jackson, D.A. Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology* **1993**, *74*, 2204–2214. [CrossRef]

52. Hammer, Ø. PAST: Multivariate Statistics. 2015. Available online: http://folk.uio.no/ohammer/past/multivar.html (accessed on 13 November 2017).
53. Kaiser, H.F. A note on Guttman's lower bound for the number of common factors. *Br. J. Stat. Psychol.* **1961**, *14*, 2. [CrossRef]
54. Maimon, O.; Rokach, L. *Data Mining and Knowledge Discovery Handbook*, Second ed.; Springer Science: New York, NY, USA, 2010; p. 1285. [CrossRef]
55. Carvalho, G.A.; Minnett, P.J.; Fleming, L.E.; Banzon, V.F.; Baringer, W. Satellite remote sensing of harmful algal blooms: A new multi-algorithm method for detecting the Florida Red Tide (*Karenia brevis*). *Harmful Algae* **2010**, *9*, 440–448. Available online: https://www.ncbi.nlm.nih.gov/pubmed/21037979 (accessed on 13 November 2017). [CrossRef] [PubMed]
56. Carvalho, G.A.; Minnett, P.J.; Banzon, V.F.; Baringer, W.; Heil, C.A. Long-term evaluation of three satellite ocean color algorithms for identifying harmful algal blooms (*Karenia brevis*) along the west coast of Florida: A matchup assessment. *Remote Sens. Environ.* **2011**, *115*, 18. Available online: http://www.ncbi.nlm.nih.gov/pubmed/22180667 (accessed on 13 November 2017). [CrossRef] [PubMed]
57. Congalton, R.G. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sens. Environ.* **1991**, *37*, 35–46. [CrossRef]
58. Nussmeier, N.A.; Miao, Y.; Roach, G.W.; Wolman, R.L.; Mora-Mangano, C.; Fox, M.; Szekely, A.; Tommasino, C.; Schwann, N.M.; Mangano, D.T. Predictive value of the National Institutes of Health Stroke Scale and the Mini-Mental State Examination for neurologic outcome after coronary artery bypass graft surgery. *J. Thorac. Cardiovasc. Surg.* **2010**, *139*, 901–912. [CrossRef] [PubMed]
59. Lee, M.-J.; Yun, M.J.; Park, M.-S.; Cha, S.H.; Kim, M.-J.; Lee, J.D.; Kim, K.W. Paraaortic lymph node metastasis in patients with intraabdominal malignancies: CT vs PET. *World J. Gastroenterol.* **2009**, *15*, 4434–4438. [CrossRef] [PubMed]
60. Carvalho, G.A. The Use of Satellite-Based Ocean Color Measurements for Detecting the Florida Red Tide (*Karenia brevis*). Ph.D. Thesis, University of Miami (UM/RSMAS/MPO), Miami, FL, USA, 2008; p. 156. Available online: http://scholarlyrepository.miami.edu/oa_theses/116/ (accessed on 13 November 2017).
61. Alberg, A.J.; Park, J.W.; Hager, B.W.; Brock, M.V.; Diener-West, M. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. *J. Gen. Intern. Med.* **2004**, *19*, 460–465. [CrossRef] [PubMed]
62. Theriault, C.; Scheibling, R.; Hatcher, B.; Jones, W. Mapping the distribution of an invasive marine alga (*Codium fragile* spp. tomentosoides) in optically shallow coastal waters using the compact airborne spectrographic imager (CASI). *Can. J. Remote Sens.* **2006**, *32*, 315–329. [CrossRef]
63. Espedal, H.A.; Johannessen, O.M. Cover: Detection of oil spills near offshore installations using synthetic aperture radar (SAR). *Int. J. Remote Sens.* **2000**, *21*, 2141–2144. [CrossRef]
64. Espedal, H.A. Detection of oil spill and natural film in the marine environment by spaceborne synthetic aperture radar. Ph.D. Dissertation, Department of Physics, University of Bergen and Nansen Environmental and Remote Sensing Center (NERSC), Bergen, Norway, 1998; p. 200.
65. Espedal, H. Detection of oil spill and natural film in the marine environment by spaceborne SAR. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '99), Hamburg, Germany, 28 June–2 July 1999; pp. 1478–1480. [CrossRef]
66. Holt, B. SAR imaging of the ocean surface. In *Synthetic Aperture Radar Marine User's Manual*; NOAA/NESDIS; Office of Research and Applications: Washington, DC, USA, 2004; Chapter 2, pp. 25–79. Available online: http://www.sarusersmanual.com (accessed on 13 November 2017).
67. Klinkenberg, B. A review of methods used to determine the fractal dimension of linear features. *Math. Geol.* **1994**, *26*, 23–46. [CrossRef]
68. Bevilacqua, L.; Barros, M.M.; Galeão, A.C.R.N. Geometry, dynamics and fractals. *J. Br. Soc. Mech. Sci. Eng.* **2008**, *30*, 11–21. [CrossRef]
69. Ledesma, R.D.; Valero-Mora, P. Determining the number of factors to retain in EFA: an easy-to-use computer program for carrying out Parallel Analysis. *Pract. Assess. Res Eval.* **2007**, *12*, 11.
70. Garcia-Pineda, O.; MacDonald, I.; Zimmer, B. Synthetic aperture radar image processing using the Supervised Textural-Neural Network Classification Algorithm. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '08), Boston, MA, USA, 8–11 July 2008; pp. IV:1265–IV:1268. [CrossRef]

71. Garcia-Pineda, O.; MacDonald, I.; Zimmer, B.; Shedd, B.; Roberts, H. Remote-sensing evaluation of geophysical anomaly sites in the outer continental slope, northern Gulf of Mexico. *Deep Sea Res. Part II: Top. Stud. Oceanogr.* **2010**, *57*, 1859–1869. [CrossRef]

72. Garcia-Pineda, O.; Zimmer, B.; Howard, M.; Pichel, W.; Li, X.; MacDonald, I.R. Using SAR images to delineate ocean oil slicks with a texture-classifying neural network algorithm (TCNNA). *Can. J. Remote Sens.* **2009**, *35*, 411–421. [CrossRef]

73. Mityagina, M.; Lavrova, O. Satellite Survey of Inner Seas: Oil Pollution in the Black and Caspian Seas. *Remote Sens.* **2016**, *8*, 875. [CrossRef]

74. Song, D.; Ding, Y.; Li, X.; Zhang, B.; Xu, M. Ocean Oil Spill Classification with RADARSAT-2 SAR Based on an Optimized Wavelet Neural Network. *Remote Sens.* **2017**, *9*, 799. [CrossRef]

75. Migliaccio, M.; Nunziata, F.; Buono, A. SAR polarimetry for sea oil slick observation. *Int. J. Remote Sen.* **2015**, *36*, 3243–3273. [CrossRef]