UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE GEOCIÊNCIAS
FACULDADE DE GEOFÍSICA

ANDREY MARCOS SOUZA DA SILVA DE LIMA

IMPROVING DIFFRACTION PATTERN RECOGNITION USING KNN

BELÉM
2017

ANDREY MARCOS SOUZA DA SILVA DE LIMA

IMPROVING DIFFRACTION PATTERN RECOGNITION USING KNN

Undergraduate thesis submitted to Faculty of Geophysics of the Universidade Federal do Pará for obtaining a Bachelor of Science degree in Geophysics.

Supervisor: Prof. Dr. Daniel Leal Macedo

BELÉM
2017

Dados Internacionais de Catalogação-na-Publicação (CIP)
Biblioteca do Instituto de Geociências/SIBI/UFPA

Lima, Andrey Marcos Souza da Silva de, 1993 -

Improving diffraction pattern recognition using kNN / Andrey Marcos Souza da Silva de Lima. – 2017.

46 f. : il. ; 30 cm

Inclui bibliografias

Orientador: Daniel Leal Macedo

Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Pará, Instituto de Geociências, Faculdade de Geofísica, Belém, 2017.

1. Prospecção sísmica. 2. Difração. 3. Geofísica. I. Título.

CDD 22. ed.: 622.1592

ANDREY MARCOS SOUZA DA SILVA DE LIMA

IMPROVING DIFFRACTION PATTERN RECOGNITION USING KNN

Undergraduate thesis submitted to Faculty of Geophysics of the Universidade Federal do Pará for obtaining a Bachelor of Science degree in Geophysics .

Approval Date: 10/03/2017
Final Grade:
Examining Committee:

Prof. Dr. Daniel Leal Macedo (Orientador)
Doutor em Geofísica
Universidade Federal do Pará

Prof. Dr. Edelson da Cruz Luz (Membro)
Doutor em Geofísica
Instituto Federal do Pará

Prof. Dr. Marcos Welby Correa Silva (Membro)
Doutor em Geofísica
Universidade Federal do Pará

ANDREY MARCOS SOUZA DA SILVA DE LIMA

2D SEISMIC MODELING OF DIFFRACTIONS USING MADAGASCAR TO AMPLITUDE ANALYSIS AND CALCULUS OF GEOMETRIC PARAMETERS TO IMPROVE DIFFRACTION PATTERN RECOGNITION USING KNN

Trabalho de Conclusão de Curso apresentado à Faculdade de Geofísica do Instituto de Geociências da Universidade Federal do Pará, como requisito parcial à obtenção de grau de Bacharel em Geofísica.

Data da defesa: 24 de março de 2017.
Conceito: E
Banca Examinadora:

Prof. Daniel Leal Macedo - Orientador
Doutor em Ciências e Engenharia de Petróleo
Universidade Federal do Pará

Prof. Jessé Carvalho Costa - Membro
Doutor em Geofísica
Universidade Federal do Pará

Prof. José Jadsom Sampaio de Figueiredo - Membro
Doutor em Ciências e Engenharia de Petróleo
Universidade Federal do Pará

This work is dedicated to my parents, Andrea Lima and Valdeci Lima, for always prioritizing my education and for always work so that I could just study. I love you.

ACKNOWLEDGEMENTS

I thank God for opportunities, joys, security and defeats (without those I would not be the same).

I thank my parents, Andrea and Valdeci because they have been working for years so that I could study and only study.

To my family especially my grandmother Helena and my brothers Anderson, Adriany and Yuri.

To my girlfriend and best friend Daniela for all the hours of studies and joys.

To the professors of the Faculty of Geophysics of UFPA for passing their knowledge throughout the graduation.

To Capes for the Sandwich Graduation in the United States throughout the Science Without Borders Program (CsF), an experience that helped me grow as a person and academically.

To the professors of the University of Oklahoma Kurt J. Marfurt, Matthew J. Pranter, John D. Pigott, Xiaowei Chen and Shannon Dulin for the knowledge learned during the exchange program at OU. Thanks also to senior researcher Dr. Bob Hardage for the opportunity to learn at Exploration Geophysics Laboratory and for the VSP classes. To Michael De Angelo and Alexander Klokov for their teachings and advices at BEG.

To my advisor, Prof. Dr. Daniel Leal Macedo, for the patience and all the teaching passed during his classes and orientation meetings. Thank you for your willingness to teach and improve the education of our faculty. Thank you for believing in my potential.

To the Sociedade Brasileira de Geofísica for the scholarship and for financing this research, through their scientific initiation .

To all others who, directly or indirectly, contributed to this work my sincerely thank you.

ABSTRACT

This work reinforces the importance of using pattern recognition in order to classify seismic events such as diffractions, edge diffractions, reflections and void points. Their identification and processing can be used for the construction of velocity models and the imaging of geological structures. The diffraction operator responses are analyzed using an algorithm to calculate two pairs of three parameters that characterize an event. The k nearest neighbor method (kNN) is used to classify these events as diffractions, reflections, edge diffractions and void points based on their diffractor operator. Since the kNN method uses a measure of distance, this work compares the classification using Euclidean and Mahalanobis distances. The results showed that e10-e6-e3 domain using Mahalanobis distance is the best combination to better cluster and classify events.

Keywords: Seismic Modeling. Seismic Diffractions. kNN.

RESUMO

Este trabalho reforça a importância do reconhecimento de padrão para classificar eventos sísmicos como por exemplo difratores, difratores de canto, refletores e pontos vazios. A identificação e processamento de difrações podem ser utilizados para a construção de modelos de velocidade e imagens de estruturas geológicas. As respostas do operador de difração são analisadas usando um algoritmo para calcular dois pares de três parâmetros que caracterizam um evento. O método kNN é usado para classificar estes eventos como difratores, refletores, difrações de canto e pontos vazios baseados em seus operadores de difrações. Visto que o método kNN usa uma medida de distância, esse trabalho compara a classificação usando as distâncias Euclideana e de Mahalanobis. Os resultados mostram que o domínio e10-e6-e3 usando a distância de Mahalanobis é o melhor domínio para agrupar e classificar os eventos.

Palavras-chave: Modelagem Sísmica. Difrações Sísmicas. kNN.

LIST OF ILLUSTRATIONS

CONTENTS

1 INTRODUCTION

Applied geophysics is concerned to investigate bodies in subsurface through the observation of their effects in the physical fields and the wave propagation phenomena (LUIZ; SILVA, 1995). It is as if one would be "hearing the geology" embedded in the physical fields that cross the matter using non-invasive techniques. The seismic method aims to study the earth using wave propagation that we call seismic waves. Observable seismic events are used to estimate the propagation velocities of the subsurface. When seismic waves propagate through a medium, geological structures appear in the seismic data as reflections, refractions, diffractions and other events. Among these events, diffractions are the ones that reveal more information about small geological features (such as channels, faults and fractures) and small changes in seismic reflectivities (FOMEL; LANDA; TANER, 2006). When we migrate a diffraction event and it collapses, it means that we used the correct velocity model. Furthermore, the collapse of a diffraction event is more sensitive to errors in the migration velocity model when compared to reflection events.

Analyzing diffractions became a goal since the kinematic and dynamic features of them have been mainly used to velocity analysis (SAVA; BIONDI; ETGEN, 2005); (FOMEL; LANDA; TANER, 2006); (NOVAIS; COSTA; SCHLEICHER, 2008);(COIMBRA et al., 2013). Before using diffraction to image geological structures, it is necessary to separate them from reflections. There are several works on this topic such as the Plane Wave Destruction (CLAERBOUT; GUBBINS, 1994) and (FOMEL, 2009) that separates those events by the calculation of the local slope. Recently, Schleicher extracted from data the inverse of the local slope to improve the PWD algorithm (SCHLEICHER et al., 2009). Klokov using radon transform separated diffractions from reflections in the dip-angle domain (KLOKOV; FOMEL, 2013) and Coimbra separated in depth domain (GONZALEZ, 2014).

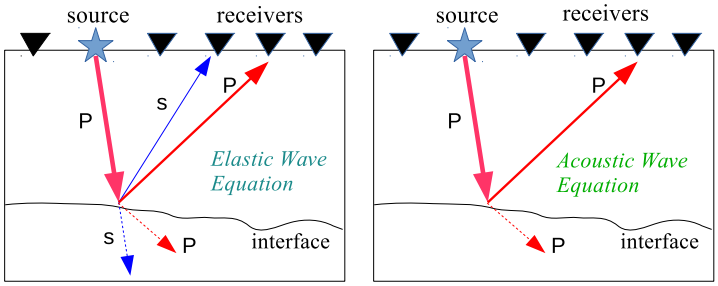
We present a methodology to identify and classify seismic diffraction events using a set of discriminating parameters and pattern recognition algorithm. In a recent study (FIGUEIREDO et al., 2013), diffractions were separated from not diffractions (reflections and void points) using pattern recognition. We went a step further by separating diffractions, reflections, edges diffractions and void points using kNN (k Nearest Neighbor) which is a pattern recognition method that needs parameters to characterize and differentiate features. The differentiation was possible by analyzing two pairs of three parameters in two new domains.

2 THEORY

2.1 Wave Equation

Geophysics uses measurements of the variations of physical fields in the Earth surface made at specific points. As we know the Earth is complex, but we approximate it by fewer parameters than the reality (see Fig.1 for an example). In the real subsurface, the wave propagates in three dimensions and when it finds an interface, it scatters and new wave modes are generated. Also, the energy is absorbed throughout the propagation and the P and S modes change with direction, which characterize a heterogeneous viscoelastic medium. However, we assume simplified versions of it by, for instance, removing the absorption effect from P and S waves, resulting in an elastic anisotropic medium. If we disconsider the anisotropic effects we end up with the isotropic elastic wave equation. The simplest way to model the wave propagation in the subsurface is considering the medium as acoustic with constant density. This is the case in this work. Below, the isotropic, constant density wave equation.

Figure 1 – Approximation of an elastic medium (left side) by an acoustic medium (right side). The general case would be a viscoelastic medium since the energy is absorbed and the frequency content decreases.



Source: From author.

$$\frac{1}{V^2}U_{,tt} - U_{,jj} = S. \tag{2.1.1}$$

where U is the displacement, V is the propagation velocity and S is a source.

2.1.1 Numerical Solution

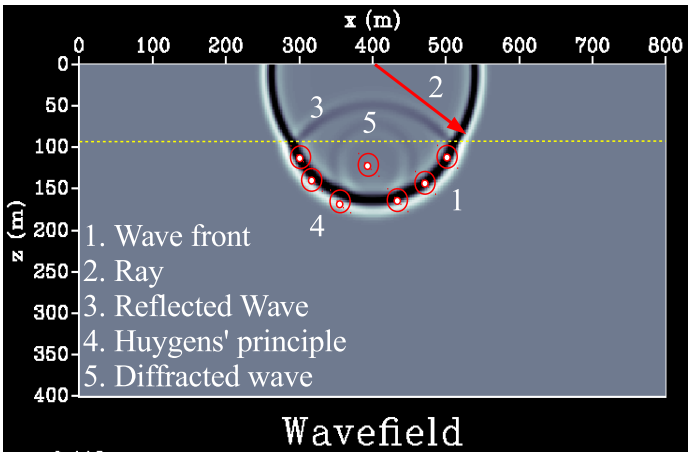
It is important to be aware that for a simple Earth's model such as the homogeneous case, the equation 2.1.1 has an analytical solution. In a model where velocity changes smoothly the solution may be found by an asymptotic approximation which leads us to the ray theory.

However, for more complex media, the solution becomes more difficult to be found and numerical approximations are used. Among the numerical methods available in the literature, the one important in this work is the finite differences that basically uses finite differences to approximate the time and spatial derivatives of the wavefield in equation 2.1.1

2.1.2 Basic Concepts In Seismic

To understand the workflow of this work is necessary to remember some of the basic concepts in seismic (see Fig. 2). The first one is wavefront which is the geometric place where all points are in phase. Ray is the direction perpendicular to the wavefront. The Snell's law describes how the wave is reflected and refracted in an interface (see yellow line in Fig. 2). The last one is Huygens' principle which is used to understand diffractions. It states that the wavefront may be seen as made of smaller secondary sources and if there is a scatter point embedded in a homogeneous medium, the punctual energy is going to be separated from the rest of the wavefront and, a hyperbolic event known as diffraction will appear in the seismogram. This concept is very important to understand the migration concept because (SCHLEICHER; TYGEL; HUBRAL, 2007) define Huygens' surfaces as the hyperbola events present in the time domain (seismogram).

Figure 2 – Basic Seismic Concepts. Wavefront is the geometric place where all points are in phase. Ray is the direction perpendicular to the wavefront. In 3 we see the reflected wavefield governed by Snell's law and 4 is the Huygens' principle which states that the wavefront may be seen as made of smaller secondary sources. In 5 we can see the diffracted wave which has low amplitude, but carries information from small scale geologic features.



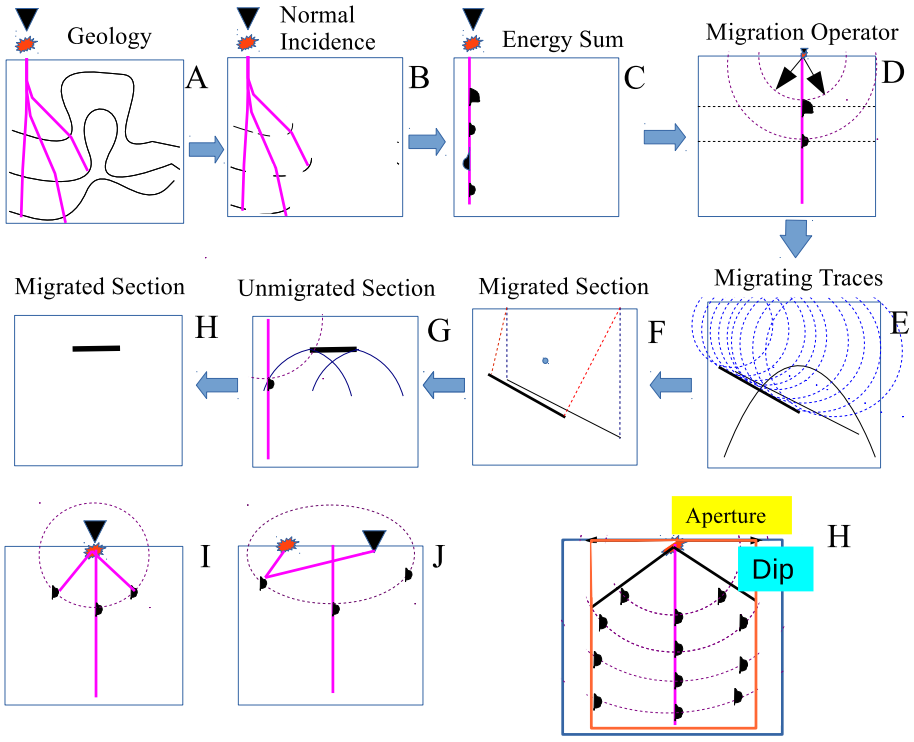
Source: From author

2.2 **Kirchhoff Migration**

The objective of seismic migration is repositioning the events in their true geological position. Among the types of migration available in the literature, the one used in this work was the Kirchhoff migration and it can be seen as smearing the trace amplitudes along isochrons in the image domain, or equivalently, it can also be seen as the sum of amplitudes over Huygens' curves which are the same as diffraction hyperbolas.

To understand the concept behind time migration, let's use the isochrons idea summarized in figure 3. When seismic waves propagate in subsurface, they illuminate the geology in many directions and a zero offset experiment is assumed many times as in A. However, the zero offset experiment only accounts for normal incidence as in B. These rays only illuminate perpendicular to the reflectors and the amount of information is smaller and appear in the seismic trace summed as only one trace as in C. We know that the energy came from anywhere from a region that, in the zero offset experiment, is a circle as in D. The migration operator spreads the energy in this circular region and sum as in E. The result of this summation are reflectors shorter in length and steeper. The migration operator collapses diffractions as in F by taking the energy down as as in G and put upwards to its true position as in H. The migration operator can be a circle if the experiment is zero offset as in I or an ellipse if the offset is different from zero and source and receivers are in the same level as in J. How far the migration operator goes up is determined by the dip parameter and how far it goes laterally is determined by the aperture parameter (see H).

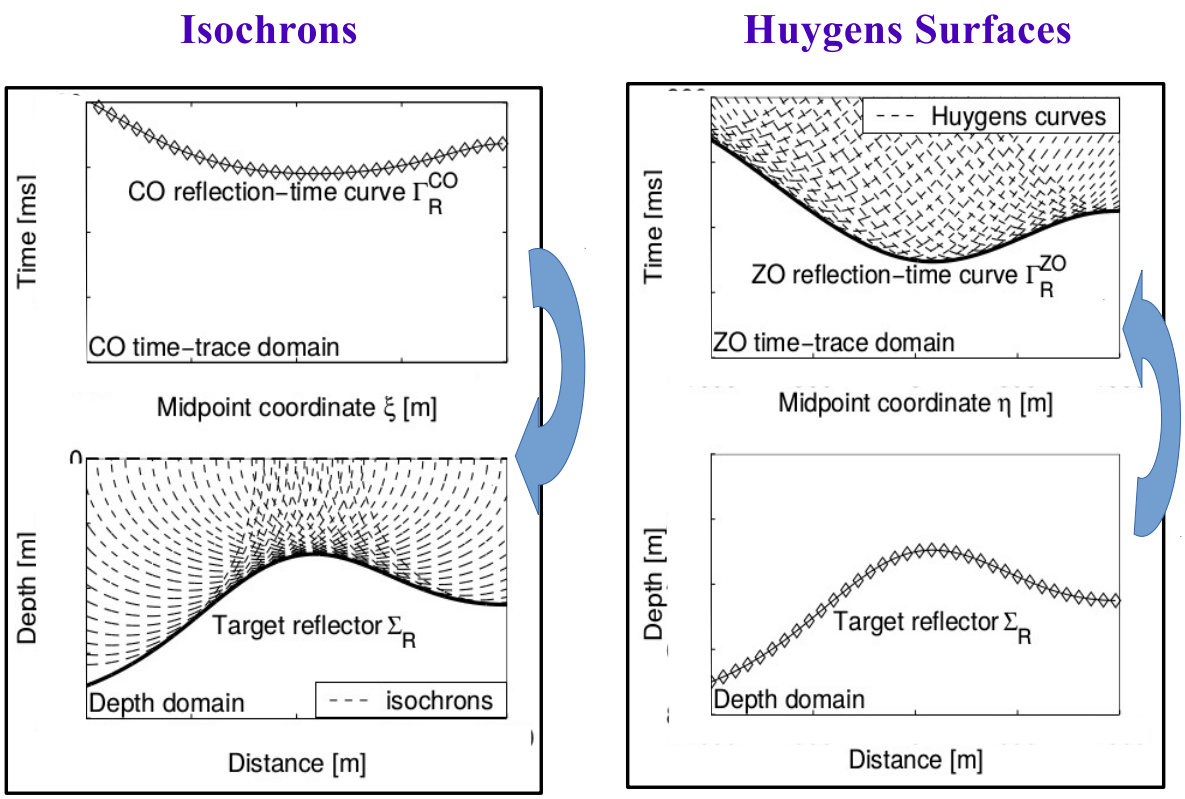
Figure 3 – Migration steps for Isochrons idea. Since we know the energy came from anywhere in a circular surface as in D, the migration operator spreads the energy along this surface and stacks it. The result are steeper and shorter reflectors and collapsed diffractions (if the velocity model is right). The migration operator depends on the aperture and dip parameters to be less noisy and reduce the amount of migration smiles.



Source: From author

In figure 4 (left side) the isochrons idea is used to migrate a whole time section. Each point in time domain is going to generate an isochron that is represented by dashed lines. If it is done for every point in time domain, those points which correspond to an actual reflection event will generate a family of isochrons whose envelope is going to define a reflector in subsurface. On the other hand, if Huygens' surfaces are used (right side of Fig. 4), each point in depth domain will generate a surface (hyperboloid) that will be used to stack events in time. Since a reflector can be seen as a continuous distribution of diffractors, events will be summed in time domain correctly.

Figure 4 – Two equivalent ways of seeing Kirchhoff migration. In the left you can see the isochrons idea where each point in the time domain is spread along isochrons and the envelope from the isochrons' family will define the reflector. On the right the Huygens' surfaces idea where each point in the depth domain will define a Huygens' surface that will be used to sum the amplitudes along this curve.



Source: Modified from (SCHLEICHER; TYGEL; HUBRAL, 2007)

The mathematical description of the migration operator using Huygens' surfaces is shown in the next section.

2.2.1 Migration Operator

The diffraction operator $D(M,x_m,h)$ at a generic image point M is defined as the vector of all seismic amplitudes to be stacked by a Kirchhoff migration based in a Huygens' surface for point M (SCHLEICHER et al., 2009). If x is the horizontal coordinate and τ the traveltime along the diffraction curve, the diffraction operator is given by.

$$D(M,x_m,h) = W(M,x_m,h)\partial_t U(x_m,t)|_{t=\tau(M,x_m,h)} \tag{2.2.1}$$

The prestack migration operator would be the sum over this diffraction operator given by

$$I(M,h) = \int_{Af} d^2x_m W(M,x_m,h)\partial_t U(x_m,t)|_{t=\tau(M,x_m,h)} \tag{2.2.2}$$

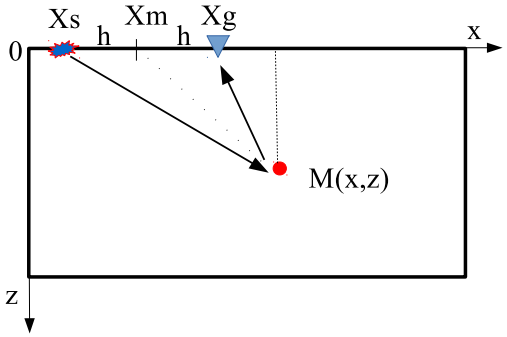
$$I(M,h) = \int_{Af} d^2x_m D(M,x_m,h) \tag{2.2.3}$$

Where Af is the seismic aperture, $W(M,x_m,h)$ is the weight function (in the simplest case $W(M,x_m,h) = 1$) and τ is defined as

$$\tau(M,x_m,h) = \sqrt{\left(\frac{t_o(x,z)}{2}\right)^2 + \left(\frac{x-x_m+h}{V_{rms}}\right)^2} + \sqrt{\left(\frac{t_o(x,z)}{2}\right)^2 + \left(\frac{x-x_m-h}{V_{rms}}\right)^2} \tag{2.2.4}$$

Where τ is the Huygens' surfaces described in figure 4, M is the image point, x is horizontal coordinate of M, z is the depth component of M, x_m is the midpoint and h is the half offset of the source-receiver pair, $t_o(x,z)$ is the two-way time in zero offset configuration with $x_m = x$ so that $t_o(x,z) = \tau(x,z,x,0)$ and V_{rms} is the RMS velocity. It is possible to see from figure 5 that $x_s = x_m - h$ and $x_g = x_m + h$.

Figure 5 – Illustration of the parameters from equation 2.2.4.



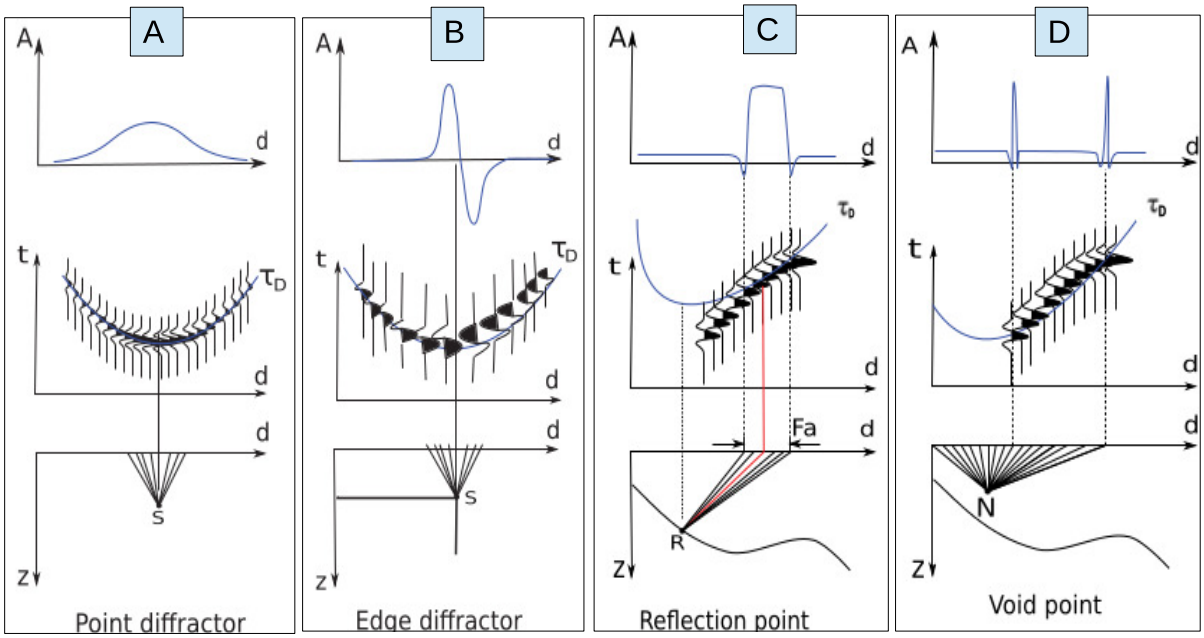
Source: From author

2.2.2 Diffraction Operator And Diffraction Moveout

In (TABTI; GELIUS; HELLMANN, 2004) they introduced amplitude analysis in the elementary diffractions within the tangency region to a reflection event. Their study will influence directly our kNN algorithm because the diffraction operators for each event was assumed to have either A, B, C or D forms (see Fig. 6). The traveltime associated with a reflection point (point R in Fig. 6-C) has a high concentrated amplitude. The traveltime associated with a void point is represented in D. Note that it can cross reflection points and it will have two picks or random picks. The traveltime associated with diffraction points is illustrated in A, being smoother than reflections. The traveltime associated with edge diffractions is

illustrated in B. Note that it has a change in polarity. This work is based on these four operators since we use them as a priori knowledge to our algorithm. In the methodology is possible to compare the theoretical events of figure 6 with the real events from our study.

Figure 6 – The four diffraction operators used as a priori knowledge within the kNN classifier. Each one of the four points (point diffractor, edge diffractor, reflection point and void point) has a different diffraction operator form.



Source: Modified from (FIGUEIREDO et al., 2013)

2.3 **Pattern Recognition**

Pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes and it is an integral part of most machine intelligence systems built for decision making (THEODORIDIS; KOUTROUMBAS, 2009). One of the most simple algorithms in machine learning is the kNN which means k Nearest Neighbors. The algorithm for the so-called nearest neighbor rule is summarized as follows. Take a set of classes w_i with M elements, and a set of training vectors t_i with N elements previously classified into the M classes. Given an unknown feature vector x and a measurement distance:

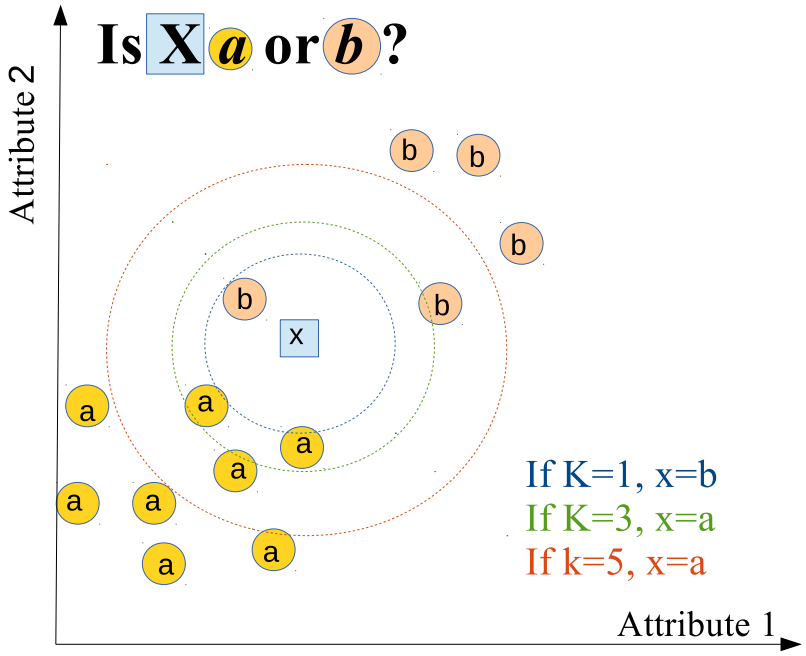
1) Out of the N training vectors, identify the k nearest neighbors, regardless of class label. k is chosen to be odd in order to avoid ties, and in general not to be a multiple of the number of classes M.

2) Out of these k samples, identify the number of vectors, k_i, that belong to class w_i, i=1,2,...M. Obviously, $\Sigma k_i = k$.

3) Assign x to the class w_i with the maximum number k_i of samples.

For example, consider a feature described by two attributes and classified into two classes: a or b. The algorithm tries to classify an unknown feature X as a or b depending on the closest neighbors by calculating the distance from this feature X to all already classified neighbors (training vectors) and the ones that have the smaller distances are the closest. This concept is better understood if we look at figure 7. If we look to the closest feature of X, we classify it as b, so if $k = 1$ we look only the closest one. Now, if we look to the 3 closest features of X, we have b, a and a. Since a appears more frequently, we classify X as a, so if $k = 3$ we look the 3 closest ones and the one with more repetitions is chosen. The same idea is applied when we choose $k = 5$: look at the 5 closest ones and count the ones that have more repetitions. Thus, both the number of neighbors k and the distance used to search for the nearest neighbors influence the classification process (CHOPRA; HADSELL; LECUN, 2005).

Figure 7 – The kNN algorithm. The features are described by two attributes. Training features are divided into two classes 'a' and 'b'. The rule is to look the closest ones and the ones with more repetitions defines the class that feature X will be assigned.



Source: From author.

2.3.1 Euclidean Vs Mahalanobis Distances

The two distances used in this work are Euclidean and Mahalanobis distances. The Euclidean distance is the "ordinary" (i.e. straight-line) distance between two points in Euclidean

space. Its mathematical definition is:

$$De(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T (\vec{x}-\vec{y})} \tag{2.3.1}$$

where $\vec{x_i} = (x_1, x_2, x_3, ..., x_p)^T$ and $\vec{y_i} = (y_1, y_2, y_3, ..., y_p)^T$

The Mahalanobis distance is a measure of the distance between two points which take into account the distribution of points where they belong. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away a point is from the mean of the distribution (MAHALANOBIS, 1922).

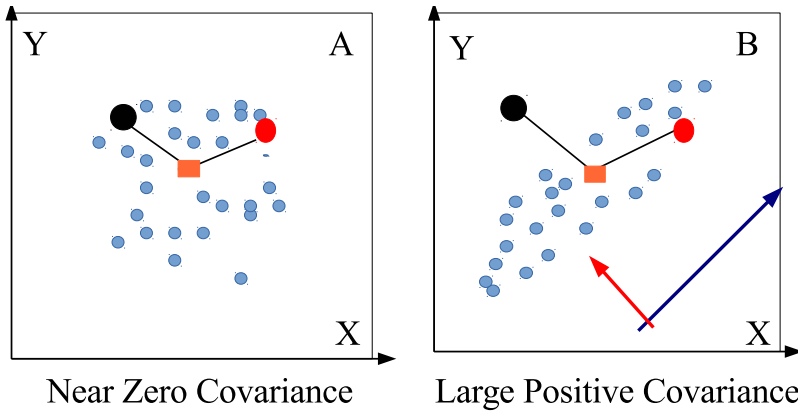
$$Dm(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1} (\vec{x}-\vec{y})} \tag{2.3.2}$$

where S is the covariance matrix, defined by

$$S = \begin{bmatrix} var(x_1) & cov(x_1,x_2) & ... & cov(x_1,x_n) \\ cov(x_2,x_1) & var(x_2) & ... & cov(x_2,x_3) \\ cov(x_3,x_1) & cov(x_3,x_2) & ... & cov(x_3,x_n) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ cov(x_n,x_1) & cov(x_n,x_2) & ... & cov(x_n,x_n) \end{bmatrix} \tag{2.3.3}$$

In a distribution with large covariance (measure of how much two random variables vary together, see Appendix B for more details), the Euclidean distance can lead to a misclassification. Figure 8 illustrates a case where one wants to know how much the black and red dots belong to the group by, for example, measuring the distance to the mean value of the group (orange square). In A and B, both black and red dots are equally distant from the mean. Because of that, in B the kNN would not considerate the fact that the red dot follows the trend of the distribution. Thus, the Euclidean distance is not the best measure of distance in cases which the data has some degree of covariance.

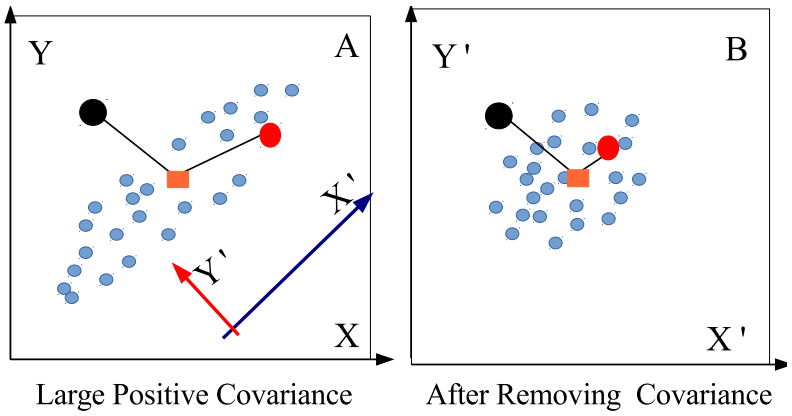
Figure 8 – Euclidean does not take account the distribution of the data.



Source: From Author

The Mahalanobis distance takes into account the covariance of the distribution. It basically removes the covariance by rotating the distribution into its eigen vector basis (blue and red arrows in Fig. 9-A) and scaling the distance by the inverse of the eigen values. After compressing the distribution (Fig. 9-B), it measures the Euclidean distance in this new vector space. Thus, the misclassification problem showed earlier is overcame and the one can, now, see that the red dot belongs to the group more than the black dot.

Figure 9 – Before and after calculus of Mahalanobis distance. Note that the distribution is compressed and Euclidean distance is applied after the remotion of covariance.



Source: From Author

3 METHODOLOGY

The first step was modeling the events. We simulated four geological situations in order to get the training data set and the test data to the our kNN algorithm. The first model contains scatters in a constant velocity background. The second model also contains scatters, but now located in a linear velocity gradient background. The third model is a layered medium with velocity increasing with depth. The fourth model is a more complex situation where edge diffractions are present due to truncated reflectors. All the modeling was done in Madagascar using a space and time 2^{nd} order acoustic finite difference modeler. The second step was to analyze the diffraction panels in the zero offset domain and define parameters to classify the diffractors operators. The analysis and classification were based on the assumption that diffractions operators produce different results for diffractions, reflections, edge diffractions and void points (see Fig. 6). The third step was to use some diffraction operators as the training vectors and test vectors for kNN. The fourth step was the kNN classification using both Euclidean and Mahalanobis distances. The details are explained below.

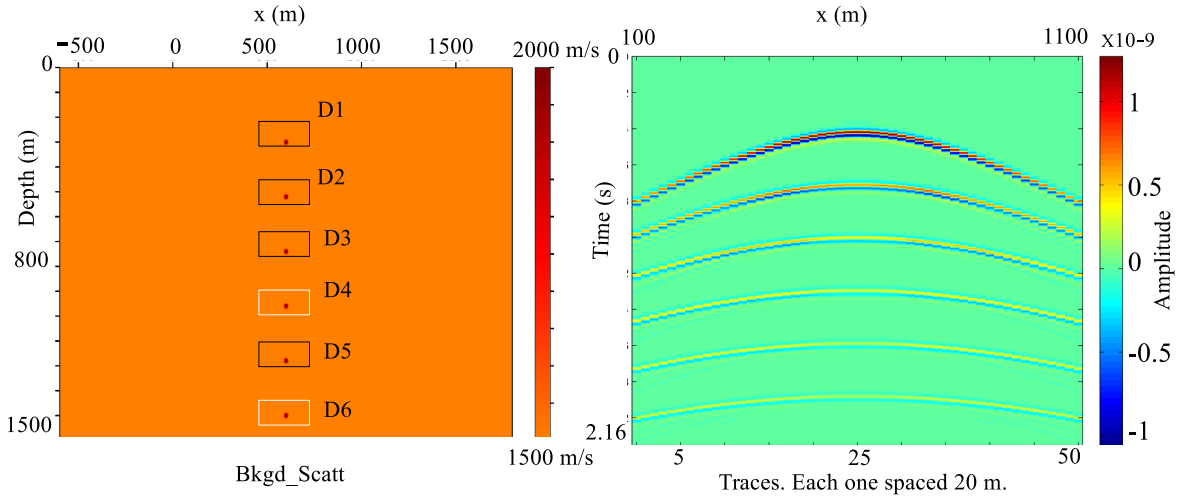
3.1 **Modeling and Data**

The models have 1500 m in depth and ranges from -600 m to 1800 m in horizontal space. The scatters are circular with radius of 10 m. According to Huygens' principle this is the situation where the size of the structure is from the order of the seismic wave wavelength and diffractions are more evident. Usually a seismic wavelength varies from 4 to 100 m depending on the frequency content (YILMAZ, 2015) and in this work the dominant frequency is 25 Hz. The experiment has 51 sources spaced by 20 m and the shot array depth is 10 m. The shot start and end positions are 100 m and 1100 m, respectively. The extra space left on the boards was necessary to avoid board effects such as the residual boundary reflections. For each shot, 11 receivers where placed in a split-spread geometry spaced by 20 m from offset -100 m to 100 m. It is important to emphasize that in the data from all models, the length of the acquisition goes from 0 m to 1200 m in space and the recording time is 2.16 s. The grid spacing was 2 m and the fd time step was 0.4 ms. After the modeling, we resampled the time to 2 ms.

3.1.1 Model 1

The first model is represented in figure 10. We introduced scatter points in the depths of 300 m, 520 m, 740 m, 960 m, 1180 m and 1400 m with a horizontal position of 600 m. The velocity in the background is constant (1500 m/s) and the scatter points have over velocity of 500 m/s with respect to its surrounding.

Figure 10 – First geological setting. Scatters of same size distributed in different depths. On the right, the data displayed in a ZO gather.



Source: From author

The diffraction operators related to points D1, D2, D3, D4, D5 and D6, which coincide with the diffraction points, will be used either as training points (D1, D2, D3 and D5) or test points (D4 and D6) in the kNN pattern recognition step. See table 1 for the exact position of all test points.

Table 1 – Exact positions of the test points for diffractions and reflections.

	X position (m)	Depth position (m)
D4	600	960
D6	600	1400
DG4	600	960
DG6	600	1400
RL2	560	960
RL3	540	1400
RR2	640	960
RR3	660	1400

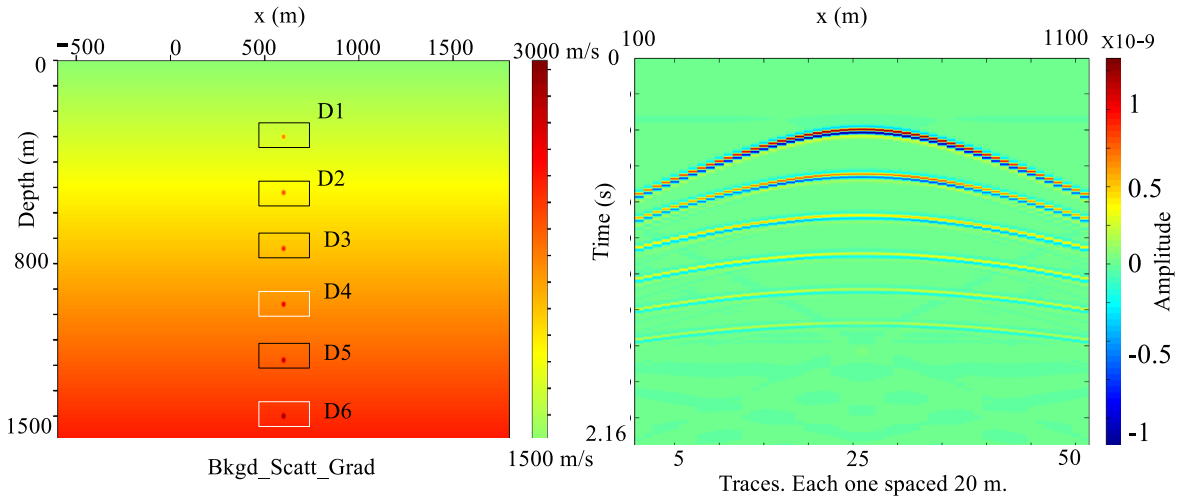
Table 2 – Exact positions of the test points for edge diffractions and void points..

	X position (m)	Depth position (m)
DCL4	340	960
DCL6	340	1400
DCR4	780	960
DCR6	780	1400
N9	120	646
N10	320	1309
N11	780	1425
N12	1600	240

3.1.2 Model 2

The second model is represented in figure 11. The positions of scatters are the same as in model 1. The velocity is a linear velocity gradient from 1500 m/s on the top to 2500 m/s in the bottom and the scatter points have over velocity of 500 m/s. Since we have a gradient background velocity, the scatters velocity changes as well.

Figure 11 – Second geological setting. Scatters of same size distributed in the same depths of the model 1 with a linear gradient of velocity varying from 1500 m/s on top to 2500 m/s on the bottom given by equation $v = 0.66d$, where v is the velocity at depth d. On the right, the data displayed in a ZO gather.



Source: From author

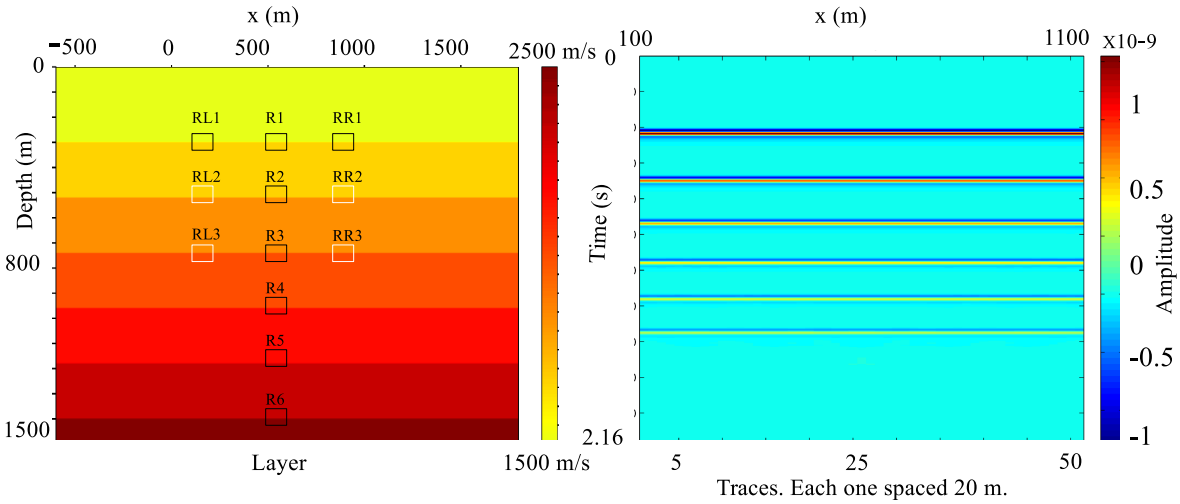
The diffraction operators related to points DG1, DG2, DG3, DG4, DG5 and DG6, which coincide with the diffraction points, will be used either as training points (DG1, DG2, DG3 and

DG5) or test points (DG4 and DG6) in the kNN pattern recognition step. See table 1 for the exact position of all test points.

3.1.3 Model 3

In this model, the top of layers are 300 m, 520 m, 740 m, 960 m, 1180 m and 1400 m. The velocity for each layer is 1500 m/s, 1666 m/s, 1833 m/s, 2000 m/s, 2166 m/s, 2330 m/s and 2500 m/s.

Figure 12 – Third geological setting. The data for this model will be used to train our reflection data set. On the right, the data displayed in a ZO gather.



Source: From author

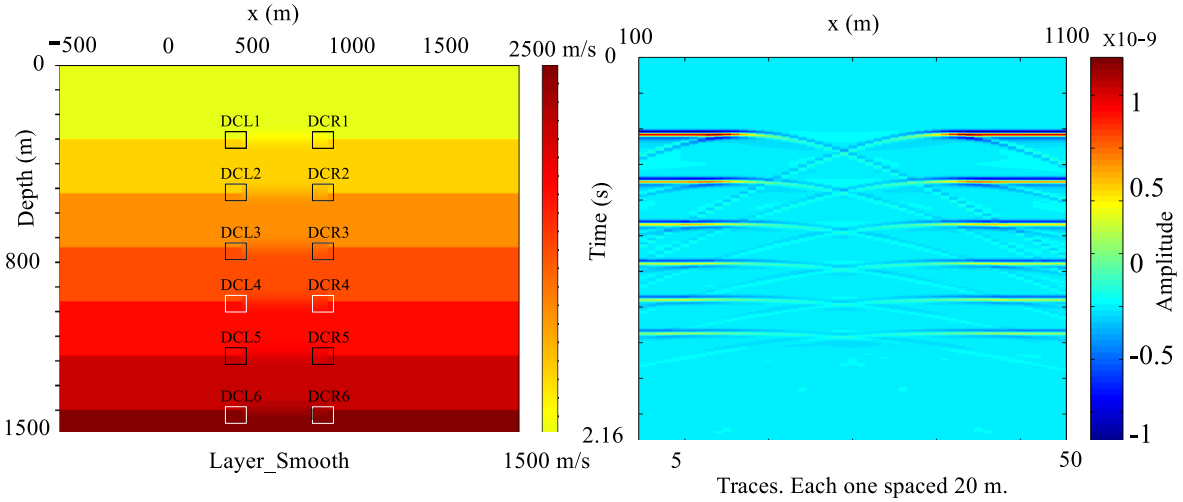
The positions of layers in depth are the same of the diffractions from model 1 and model 2 because we would like to observe if the 6 parameters are similar in both cases. The diffraction operator of points R1, R2, R3, R4, R5, R6, RL1 and RR1 will be used as training points while the diffraction operators of RL1, RL2, RR2 and RR3 will be used as test points. See table 2 for the exact position of all test points.

3.1.4 Model 4

This is a more complex model where the top of layers are 300 m, 520 m, 740 m, 960 m, 1180 m and 1400 m. The velocity for each layer is 1500 m/s, 1666 m/s, 1833 m/s, 2000 m/s, 2166 m/s, 2330 m/s and 2500 m/s. In the center of the model we have introduced a smooth layered media simulating a situation where smooth changes are really close to abrupt changes. The left and right lateral steps that creates the edge diffractions are placed horizontally at positions 350 m and 800 m respectively. The diffraction operator of points DCL1, DCL2,

DCL3, DCL5, DCR1, DCR2, DCR3 and DCR5 will be used as training points while the diffraction operators of DCL4, DCL6, DCR4 and DCR6 will be used as test points. See table 2 for the exact position of all test points.

Figure 13 – Fourth geological setting. The top of layers are 300 m, 520 m, 740 m, 960 m, 1180 m and 1400 m. The velocity for each layer is 1500 m/s, 1666 m/s, 1833 m/s, 2000 m/s, 2166 m/s, 2330 m/s and 2500 m/s. On the right, the data displayed in a ZO gather.

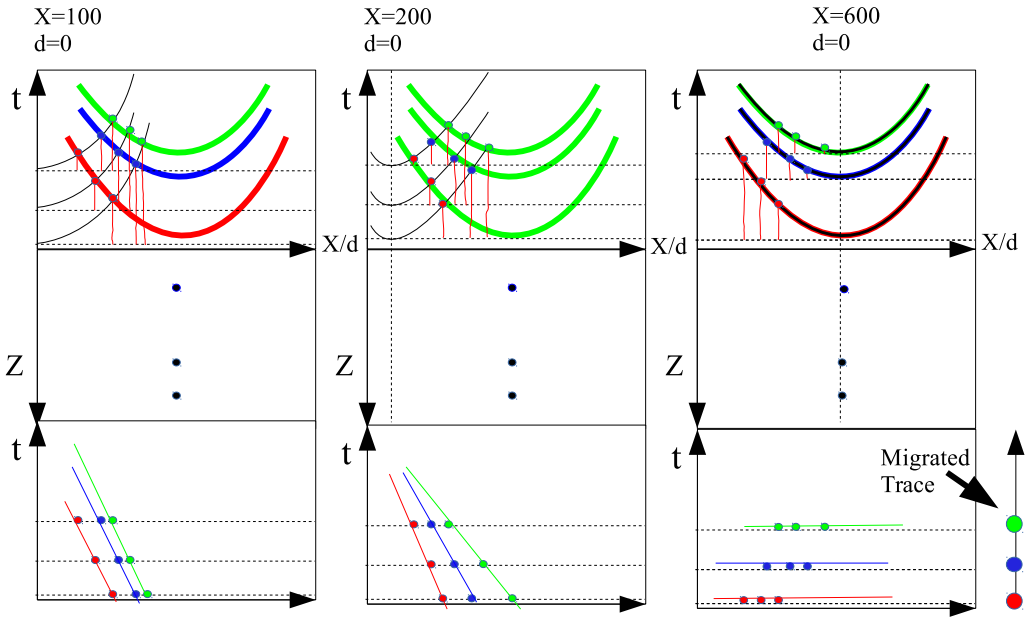


Source: From author

3.2 **Diffraction Moveout and Panels**

We used a Kirchhoff migration algorithm from CREWES (available in www.crewes.org) to create the diffraction panels. This algorithm creates the moveout out panels and then stacks the events. The step to create the moveout panels is illustrated in figure 14. Each point (red, blue and green) represents the amplitudes in the seismic event that will be brought to the apex line (dashed black line) of the correspondent diffraction-traveltime curves (full black lines). If the diffraction-traveltime curve fits an event (as in figure 14, panel x=600) this event will be flattened. Once the moveout is done for a specific "x" position, the moveout panel is stack along the horizontal direction and a time-image trace is produced.

Figure 14 – Diffraction Operator. Each black curve represents the diffraction curve and each point represents the amplitudes that will be put into the apex of the hyperbola which results in a moveout panel.

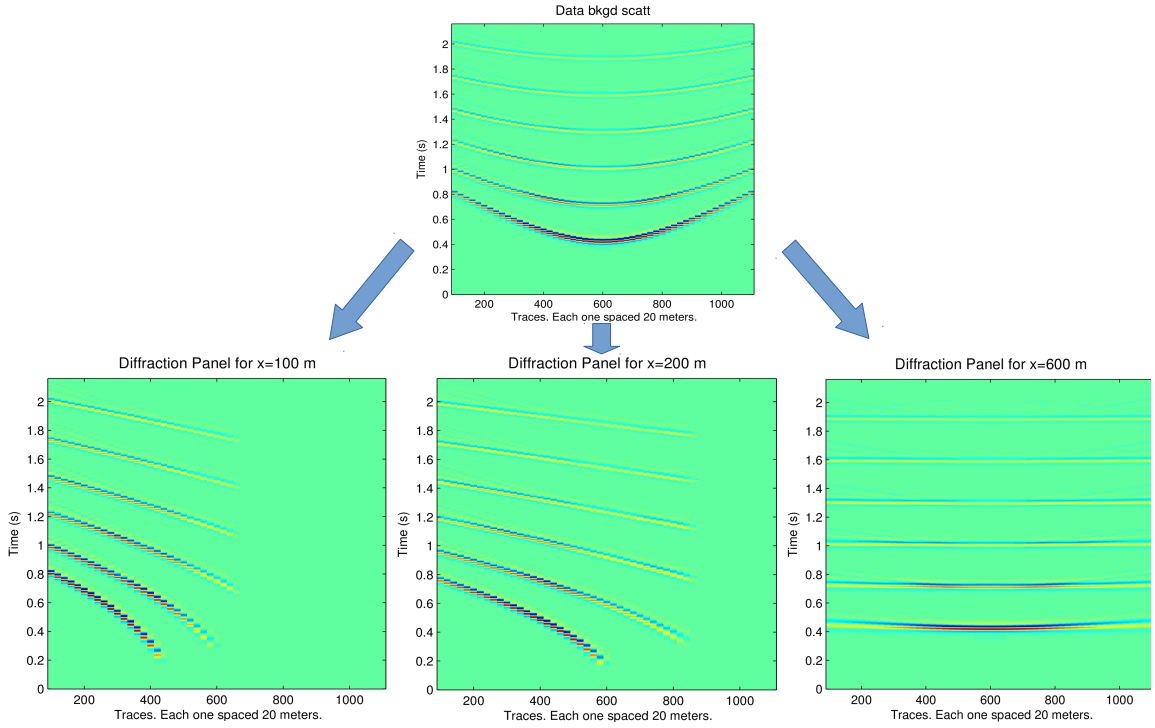


Source: From author

Figure 14 is very important because, before the step procedure, we stop and study the diffraction operators. In figure 15 it is possible to see the data on top and the moveout panels (x=100, x=200 and x=600) from our work below it, as an example of figure 14. Each horizontal line (a fixed time) in the panels, correspond to a diffraction operator of a point in depth domain in the correspondent horizontal position in the model domain. Once the events are flattened, they are used to calculate attributes that can be used for construct correlograms. It is when the human interaction starts with the algorithm because the operator will define which fixed time to observe and set the event as diffraction, reflection, edge diffraction or void point. Details in how it is implemented, refer to Appendix B.

If the diffraction point is at any position in the model, the panel associated can be found by equation $x = (i-1)*20+100$, where x is the position in space and i is the panel. Note that this rule changes if you change the number of sources.

Figure 15 – Moveout panels from our work. When the chosen 'x' position coincides with diffractors positions, all the points are flattened correctly.



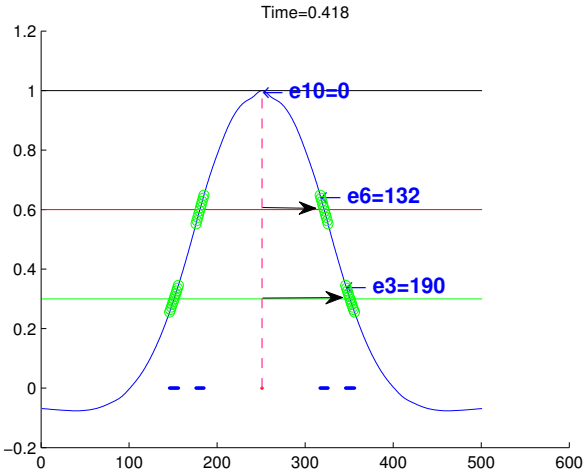
Source: From author

3.3 **Training Vectors and Classification of Parameters**

The separation method used in this work (kNN) needs some attributes to classify events. It was created an algorithm called *finder* to evaluate this attributes (see Appendix B for details). We have to choose parameters that, based on the diffraction operators, differentiate each one of the events which we want to classify: diffractions, edge diffractions, reflections and void points (see Fig. 6). We have have chosen 6 parameters divided in 2 domains. In the first domain, e10-e6-e3, the operators are characterized by (see Fig. 16):

1. $e10$ is the distance from the maximum normalized energy from the origin (defined by the chosen panel)
2. $e6$ is the distance from 60% of the maximum energy from the origin
3. $e3$ is the distance from 30% of the maximum energy from the origin

Figure 16 – Illustration of e10-e6-e3.

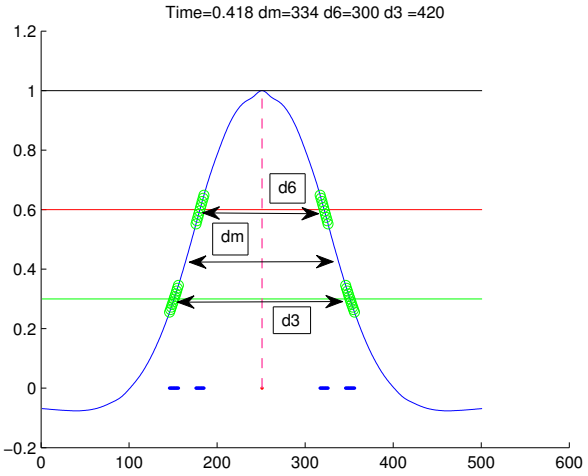


Source: From author

The following domain, dm-d6-d3, is characterized by (see Fig. 17):

1. *dm* is the distance between the two closest inflection points of the maximum energy
2. *d*6 is the distance from the two closest e6 of the maximum energy
3. *d*3 is the distance from the two closest e3 of the maximum energy

Figure 17 – Illustration of dm-d6-d3.



Source: From author

3.4 **kNN Classification**

The implementation of kNN algorithm is the following: Once you have established the training data set, each line of the data test is used to construct a matrix of equal lines (same size

as training). Then it calculates the Euclidean or Mahalanobis distance. A new matrix with these distances is created. We reorder the distances matrix in ascending order and since each line of the training is associated with D1, D2, D3, ..., DG1, ..., R1,.. DC1, DC2, ..., and N8, this index information is carried with the distances. Now, the only thing needed is to look at the k nearest neighbors. If you choose $k = 1$, just look the first neighbor. If $k = 3$ look the 3 closest ones and the more repetitive classification is chosen. The same approach for $k = 5$. The final comparison of both domains and distances are present on table 4.

These attributes are used to create correlograms, where e10, e6 and e3 will define a domain where $e10 = x$, $e6 = y$, $e3 = z$ and dm, d6 and d3 will define another domain where $dm = x$, $d6 = y$ and $d3 = z$. The attributes were chosen from the a priori knowledge about the shapes of the events (Fig. 6). These attributes were chosen in order to differentiate the events spatially in the correlogram (to create a cluster for each type of event). After the training points were plotted in the correlogram and classified, any other event to be classified will need its attributes calculated and later, using either Euclidean distance or Mahalanobis distance, to be classified by the method. It is the same problem of classifying X as a or b in figure 7. Here, the $a's$ and $b's$ are diffractions, reflections, edge diffractions and void points and X is any other event that needs to be classified. The algorithm does it by comparing the 6 parameters from X with the 6 parameters from all the training data set. All the steps of the methodology are in figures 18 and 19.

Figure 18 – The methodology summarized.

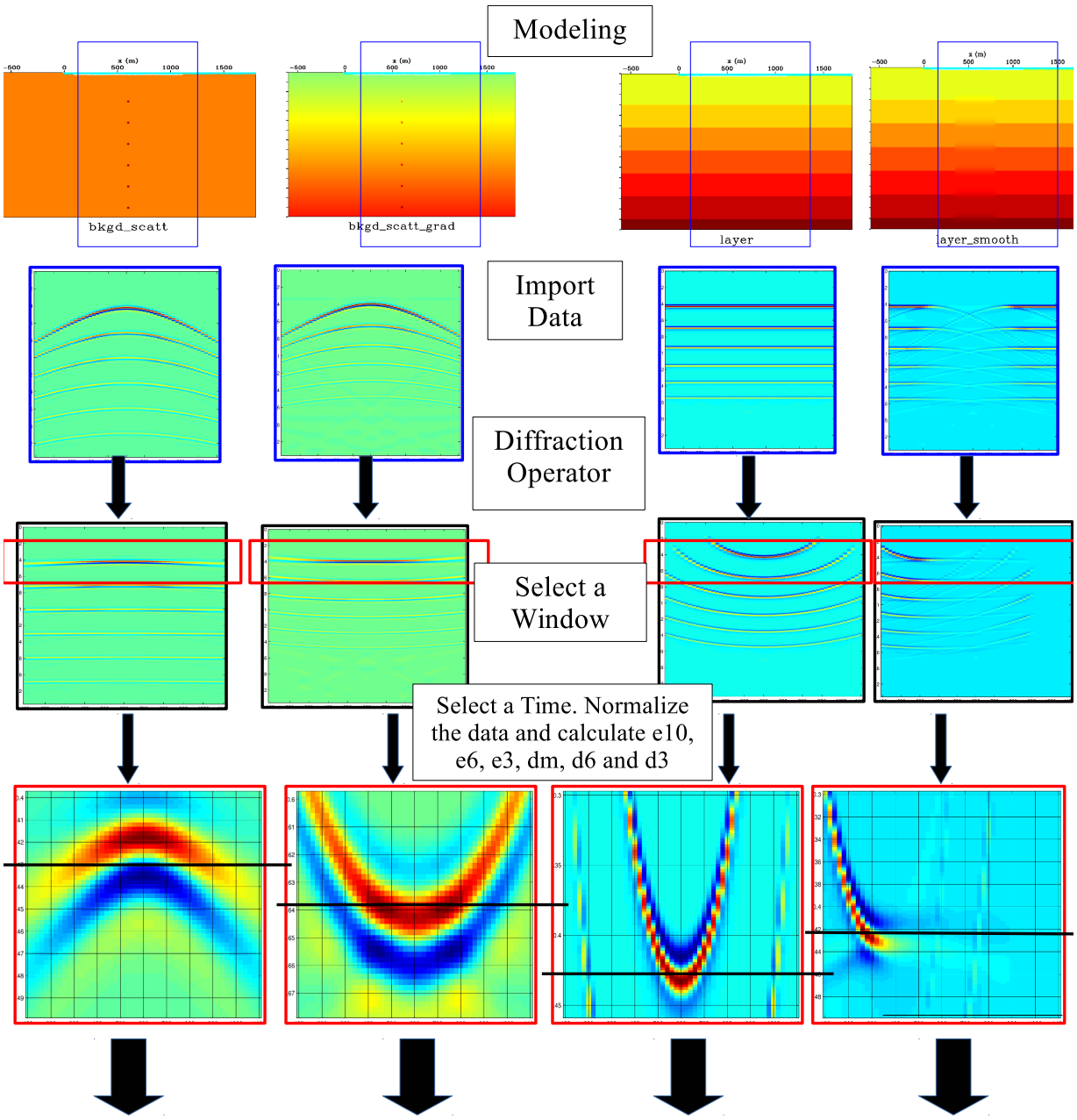
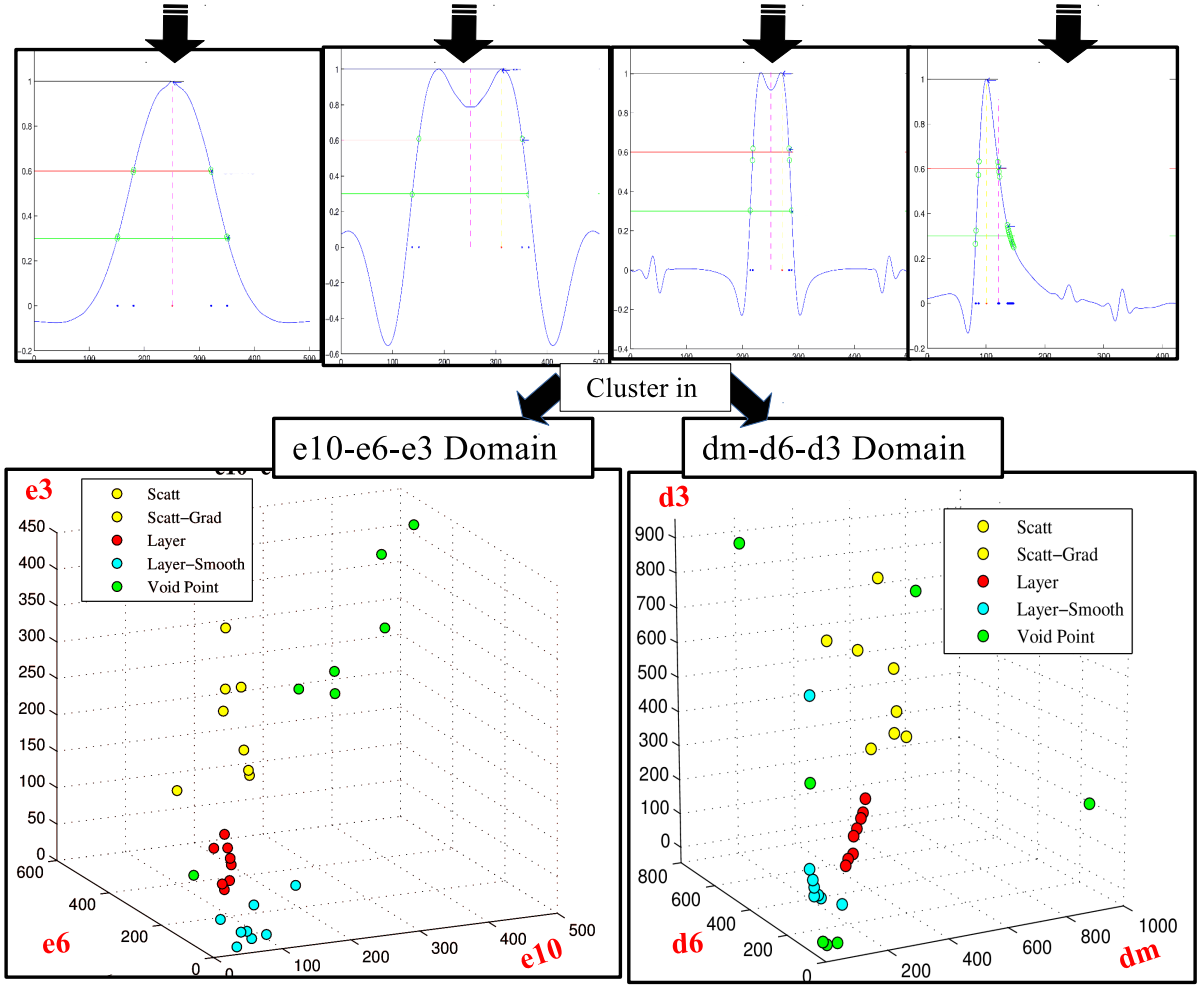


Figure 19 – Continuation of figure 18



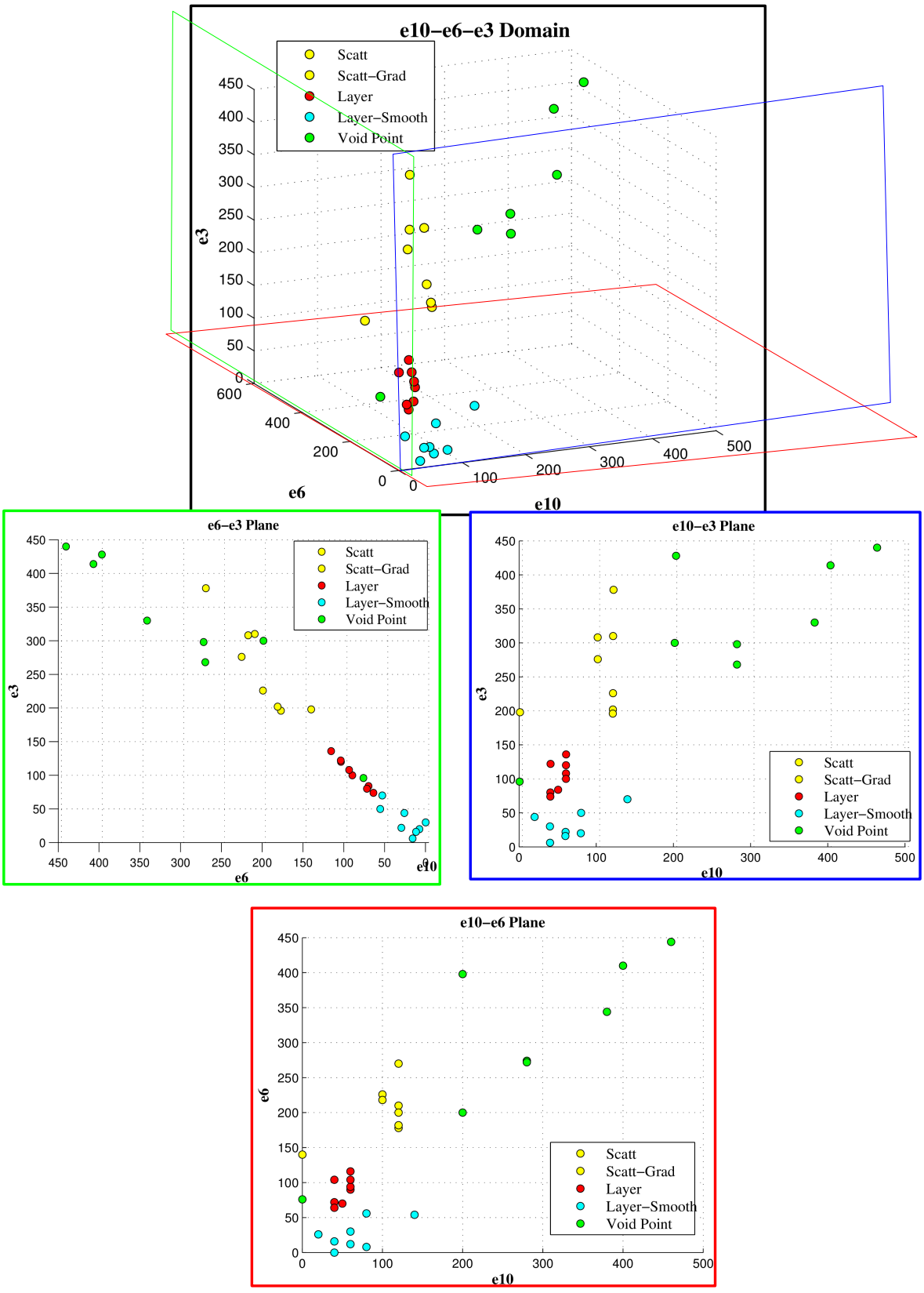
Source: From author

4 RESULTS AND DISCUSSION

The training data are depicted in figures 20 and 21. The degree of clustering is very important for our kNN algorithm because if the training vectors are not well clustered, the classification for any other event will fail. The observations for each plane is important in case we want to create a new domain to plot the correlograms of the evnts. The e6-e3 plane has a high positive covariance. Edges diffractions in blue are close to zero values. The reflections in red are located from 60 m to 130 m. Between 150 m and 260 m are the diffractions (yellow) and after 270 m we have void points (green). In the e10-e3 plane the events are fairly clustered in the lower portion and upper portion. It is clear that all the groups are well clustered with no superpositions. The best clustering plane is e10-e6, except for the fact that we have some zero points. In this plane, edge diffractions and reflections are close and diffractions (yellow) are well clustered. This last fact probably will reduce the error when classifying diffractions.

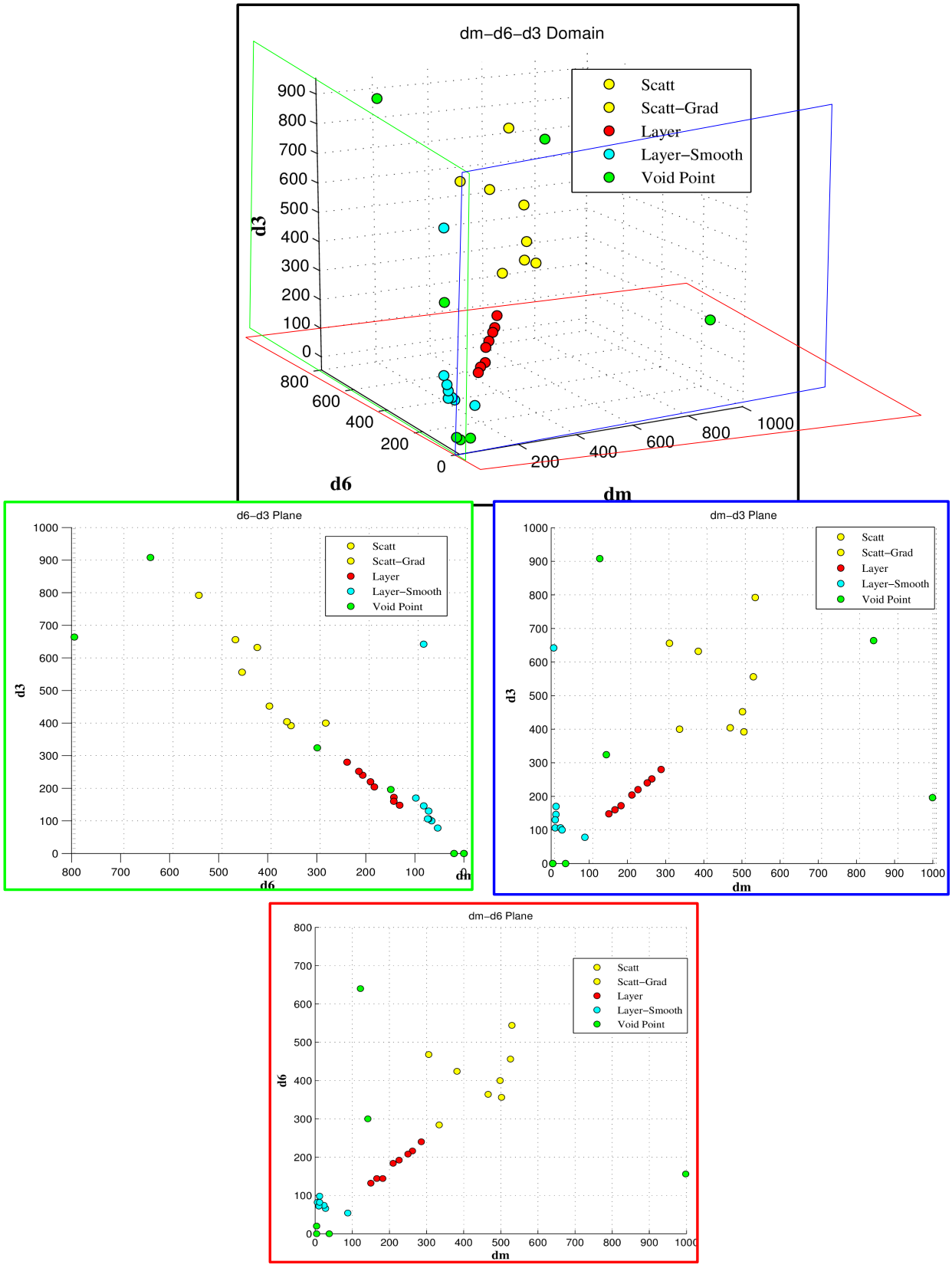
In the d6-d3 we have some reflections points close to edge diffractions, this is a problem because the kNN algorithm is very sensitive to this proximity. On the other hand, plane dm-d3 improved this clustering since d6 is very close to zero for edges and for reflections (they are around 120 to 300 m). It is possible to see the trend in reflection events which will be detected by the Mahalanobis distance. The diffraction points are clustered, but not as in plane e10-e6. The best plane was dm-d6 since edge diffractions, reflections and diffractions were well clustered. However, the algorithm can present a problem with void points because they are not clustered. This problem is due to the normalization of the data which enhances small features. Both domains showed to be good for clustering the four events. The training data set showed in figures 20 and 21 came from the selected points for the four models. The test data set is provided on table 3.

Figure 20 – Domain 1. The best clustering plane is the e10-e6 because e3 can incorporate small features after normalization.



Source: From author

Figure 21 – Domain 2. The best plane for clustering was dm-d6. Again, the void points need to be eliminated because the events are normalized and it can enhance undesirable features.



Source: From author

Table 3 – This is the data set that will be classified. Each line of this matrix is the input to our kNN algorithm so it measures the two distances (De and Dm) from each line of the training data set, arrange in ascending order and look the k nearest ones according to k. If you choose k=1 we look only the first neighbor in this new organized matrix. If k=3, we look the three closest ones and the one with more frequency is chosen and the same principle is applied for k=5. The results of our classification algorithm is the next table for both domains and for the two distances.

	Domain 1			**Domain 2**		
Test	**e10**	**e6**	**e3**	**dm**	**d6**	**d3**
D4	120	304	426	150	620	872
D6	100	298	428	122	640	908
DG4	100	168	188	670	344	384
DG6	120	182	202	470	368	408
RL2	20	82	98	166	168	196
RL3	80	116	130	270	232	262
RR2	80	118	136	284	240	272
RR3	50	100	122	254	204	248
DCL4	60	0	78	12	118	272
DCL6	140	102	80	14	76	436
DCR4	60	16	2	378	104	338
DCR6	20	14	34	114	72	106
N9	460	288	230	210	182	244
N10	370	500	403	570	596	616
N11	120	96	84	12	64	86
N12	680	692	666	32	450	32

Table 4 – Final comparison of our kNN algorithm for both domains and two distances. The Mahalanobis distance in the 10-e6-e3 domain showed to be the best combination of domain and distance measurement. D stands for diffractions, R for reflections, C for edge diffractions and N for void points

	e10-e6-e3 Domain						**dm-d6-d3 Domain**					
	Euclidean			Mahalanobis			Euclidean			Mahalanobis		
Test	K=1	K=3	K=5	K=1	K=3	K=5	K=1	K=3	K=5	K=1	K=3	K=5
D4	D	D	D	D	D	D	N	D	D	N	D	D
D6	D	D	D	D	D	D	N	D	D	N	D	D
DG4	D	D	D	D	D	D	D	D	D	D	D	D
DG6	D	D	D	D	D	D	D	D	D	D	D	D
RL2	N	R	R	R	R	R	R	R	R	R	R	R
RL3	R	R	R	R	R	R	R	R	R	R	R	R
RR2	R	R	R	R	R	R	R	R	R	R	R	R
RR3	R	R	R	R	R	R	R	R	R	R	R	R
DCL4	C	C	C	N	C	C	C	C	C	N	C	C
DCL6	C	C	R	C	C	C	C	C	C	C	D	C
DCR4	C	C	C	C	C	C	R	R	R	D	D	D
DCR6	C	C	C	C	C	C	C	C	C	C	R	R
N9	N	N	N	N	N	N	R	R	R	R	R	R
N10	N	N	N	N	N	N	D	D	D	D	D	D
N11	C	R	R	C	C	C	C	C	C	C	C	C
N12	N	N	N	N	N	N	N	R	R	N	N	N

All the squares that are not white were misclassified. Below we make some observations from this table:

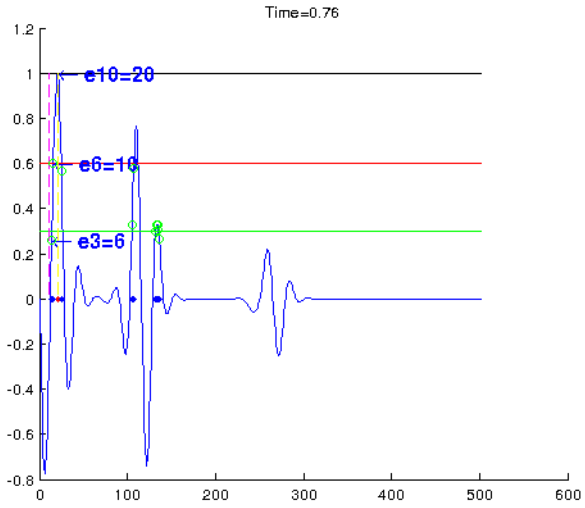
The accuracy when classifying diffractions in constant velocity background and linear gradient need to be emphasized (see D4, D6, DG4 and DG6) in the e10-e6-e3 domain. Only four misclassifications (see pink squares in Diffractions line). This is the result of having good training points presented in figure 20.

In all domains and for all distances, choosing $k = 1$ can be a problem if the data is not well clustered (look at the lighter pink squares where $k = 1$). The only misclassification in reflection events is due to $k = 1$. This is an isolated problem compared with the whole classification using both domains and distances.

The dm-d6-d3 domain showed to be a poor domain (see all different tones of pink in the right side) for clustering and classification. This problem can be attributed to instability when calculating d3 since small features (noise) are normalized which can enhance undesirable picks

(see in Fig. 22).

Figure 22 – Instability due to enhancement of small features.



Source: From Author

The reflection events where 100% classified in the dm-d6-d3 domain. This is probably because reflections events have concentrated energy so that errors on the edge of the events do not influence in dm, d6 and d3 calculations. That was the main difference from reflections to the rest of events: picks with high concentrated amplitude. In the dm-d6-d3 domain using both distances, DCR4 was totally misclassified. This is present because this distance removes the covariance from the reflections events and let they close to edge diffractions. This problem was alerted when we looked at d6-d3 plane where reflection points where close to edge diffractions (see Fig. 21 - d6-d3 plane). The N11 event was totally misclassified in all domains using both distances. However, here we come with a possible human influence: if someone chooses a wrong location or not a representative point to be the reference (training vector), the classification will be very difficult to the algorithm. Probably, this void point (N11) was selected really close to an edge diffraction and because of that the algorithm classified as C, majority.

If you consider that $k = 1$ is not a good nearest neighbor rule because some isolated points can be close to another different group, and if you consider that N11 was a human mistake made when choosing the training data, the best plane for classification was the e10-e6-e3 domain. More specifically, in this domain $k = 3$ using Euclidean distance, $k = 3$ and $k = 5$ in Mahalanobis distance are the best combination of domain, distance and k's. In general, for all domains, distances and k's, Mahalanobis is the best distance measurement. However, this result may change if you have a larger data set. In overall the e10-e6-e3 was the best domain for clustering. Moreover, the Mahalanobis distance showed to be better than Euclidean distance.

5 CONCLUSION

In this work we present a step further in the separation of seismic diffractions from other events. We improved the classes of *k* Nearest Neighbors from diffractions and not diffractions (FIGUEIREDO et al., 2013) to diffractions, reflections, edges diffractions and void points. The development of two new domains based on geometric parameters of diffraction operator showed that the e10-e6-e3 domain using Mahalanobis distance is the best clustering combination to improve the pattern recognition of these four classes.

The next steps would be the modeling of new geological settings using elastic wave equation modeling and the creation of a new domain (e10-e6-dm-d6) since the combination of these parameters were good. Since the modeling was not contaminated with noise, one could test how robust to noise the chosen classification would be. However, the algorithm can present a problem with the void points if they are not well clustered. This problem can be overcame if a new study imposes a minimum amplitude rule to be classified in the kNN algorithm. The random points in the time section will probably have low amplitudes compared with diffractions and reflections. If one just eliminates the e3 and d3, this problem can be overcame because those were the ones with more problems.

REFERENCES

CHOPRA, S.; HADSELL, R.; LECUN, Y. Learning a similarity metric discriminatively, with application to face verification. In: IEEE. **Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on**. [S.l.], 2005. v. 1, p. 539–546.

CLAERBOUT, J. F.; GUBBINS, D. Earth sounding analysis: processing versus inversion. **Nature**, [London: Macmillan Journals], 1869-, v. 370, n. 6486, p. 190, 1994.

COIMBRA, T. A. et al. Migration velocity analysis using residual diffraction moveout in the poststack depth domain. **Geophysics**, v. 78, n. 3, p. 125–135, 2013.

FIGUEIREDO, J. de et al. Automatic detection and imaging of diffraction points using pattern recognition. **Geophysical Prospecting**, Wiley Online Library, v. 61, n. s1, p. 368–379, 2013.

FOMEL, S. Applications of plane-wave destruction filters. **Geophysics**, v. 67, p. 1946–1960, 2009.

FOMEL, S.; LANDA, E.; TANER, M. T. Post-stack velocity analysis by separation and imaging of seismic diffraction. **Geophysics**, v. 72, n. 6, p. 89, 2006.

GONZALEZ, J. A. C. **Joining diffraction filter and residual diffraction moveout to construct a velocity model in the depth and time domains: application to a Viking Graben data set**. Pará. [S.l.]: Universidade Federal do Pará, 2014. (Belém).

KLOKOV, A.; FOMEL, S. Separation and imaging of seismic diffractions using migrated dip-angle gathers. **Geophysics**, v. 77, p. 131–143, 2013.

LUIZ, J. G.; SILVA, L. M. da Costa e. **Geofísica de prospecção**. 1. ed. Belém: Editora CEJUP, 1995. 5 p.

MAHALANOBIS, P. C. Anthropological observations on the anglo-indians of calcutta: part i - analysis of male stature. **Record of the Indian Museum. 23**, p. 1–96, 1922.

NOVAIS, A.; COSTA, J.; SCHLEICHER, J. Gpr velocity determination by image-wave remigration. **Journal of Applied Geophysics**, v. 65, p. 65–72, 2008.

SAVA, P.; BIONDI, B.; ETGEN, J. Wave-equation migration velocity analysis by focusing diffractions and reflections. **Geophysics**, v. 74, n. 4, p. 25–33, 2005.

SCHLEICHER, J. et al. On the estimation of local slopes. **Geophysics**, v. 74, p. 25–33, 2009.

SCHLEICHER, J.; TYGEL, M.; HUBRAL, P. **Seismic true-amplitude imaging**. 1. ed. Campinas-SP: Society of Exploration Geophysicists, 2007. 36-37 p.

TABTI, H.; GELIUS, L.-J.; HELLMANN, T. Fresnel aperture prestack depth migration. **First Break**, v. 22, n. 3, p. 39–46, 2004.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern recognition**. 4. ed. [S.l.]: Elsevier, 2009. 5 p.

YILMAZ, Ö. **Engineering seismology with applications to geotechnical engineering**. [S.l.]: Society of Exploration Geophysicists, 2015. 65 p.

APPENDICES

APPENDIX A

Table 5 – Training data set that will be the reference to our kNN algorithm using both euclidean distance and mahalanobis distance.

	e10	e6	e3	dm	d6	d3
D1	0	140	198	334	284	400
D2	120	210	310	382	424	632
D3	120	270	378	530	544	792
D5	100	218	308	306	468	656
DG1	100	226	276	526	456	556
DG2	120	200	226	498	400	452
DG3	120	182	202	466	364	404
DG5	120	178	196	502	356	392
R1	40	64	74	150	132	148
R2	60	90	100	210	184	204
R3	60	94	108	226	192	220
R4	60	104	120	250	208	240
R5	40	104	122	262	216	252
R6	60	116	136	286	240	280
RL1	40	72	80	166	144	160
RL2	50	70	84	182	144	172
DCL1	40	0	30	10	72	130
DCL2	60	12	16	12	82	146
DCL3	80	8	20	12	98	170
DCL5	140	54	70	10	72	106
DCR1	80	56	50	88	54	78
DCR2	60	30	22	28	66	100
DCR3	40	16	6	6	82	642
DCR5	20	26	44	24	74	106
N1	200	200	300	142	300	324
N2	200	398	428	122	640	908
N3	380	344	330	838	800	664
N4	0	76	96	998	156	196

Table 6 – Continuation of Table 5.

	e10	e6	e3	dm	d6	d3
N5	400	410	414	4	0	0
N6	280	272	268	4	20	0
N7	460	444	440	38	0	0
N8	280	274	298	4	20	0

The calculation of the Euclidean Distance and Mahalanobis distance is done as follows: Let De and Dm be:

$$De(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T (\vec{x}-\vec{y})} \tag{.0.1}$$

$$Dm(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1} (\vec{x}-\vec{y})} \tag{.0.2}$$

where $\vec{x_i} = (x_1, x_2, x_3, ..., x_p)^T$, $\vec{y_i} = (y_1, y_2, y_3, ..., y_p)^T$ and S the covarianve matrix defined by:

$$s = \begin{bmatrix} var(x_1) & cov(x_1,x_2) & ... & cov(x_1,x_n) \\ cov(x_2,x_1) & var(x_2) & ... & cov(x_2,x_3) \\ cov(x_3,x_1) & cov(x_3,x_2) & ... & cov(x_3,x_n) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ cov(x_n,x_1) & cov(x_n,x_2) & ... & cov(x_n,x_n) \end{bmatrix} \tag{.0.3}$$

where cov means covariance and defined by:

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) \quad \text{and} \quad var(X1) = cov(X1,X1) \tag{.0.4}$$

The covariance matrix for each distance (De and Dm) in table becomes.

$$Se = \begin{bmatrix} var(e10) & cov(e10,e6) & cov(e10,e6) \\ cov(e6,e10) & var(e6) & cov(e6,e3) \\ cov(e3,e10) & cov(e3,e6) & var(e3) \end{bmatrix} \tag{.0.5}$$

$$Sd = \begin{bmatrix} var(dm) & cov(dm,d6) & cov(dm,d6) \\ cov(d6,dm) & var(d6) & cov(d6,d3) \\ cov(d3,dm) & cov(d3,d6) & var(d3) \end{bmatrix} \tag{.0.6}$$

APPENDIX B

Once the modeling is done and the data is loaded to finder. It generates all the diffractions panels for each source position. You are asked if you want to see all the panels. In figure 23 we show only one panel (the 26^{th} panel). Its output will be the chosen panel (diffraction operator).

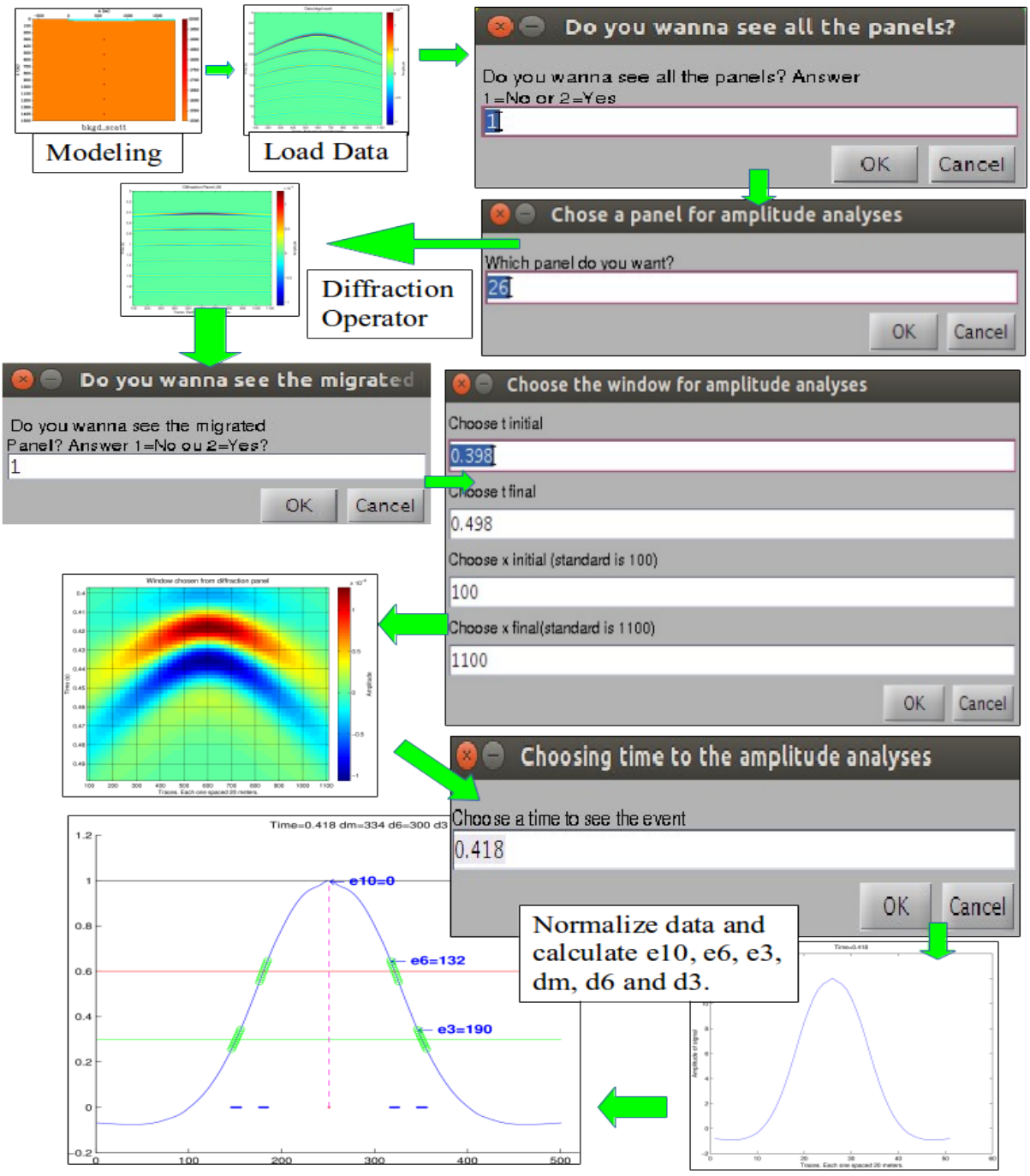
Finder asks you to show the migrated data. We have chosen '*No*' because we do not need this step. In the diffraction panel we look at a specific window, for instance, the window chosen in the dialog box "Choose the window for amplitude analyses". In this specific window we choose a specific time (0.418 s) to plot the event and calculates its 6 parameters.

It is important to emphasize that *finder* attributes e10, e6, e3, dm, d6 and d3 according to a range of 1% to 5% error because some events are steeper than others and since the algorithm interpolates (using spline interpolation) the points between the 51 traces, sometimes we get an error on locating these parameters.

The idea is well understood if you imagine that the e6 or e3 line crosses in a position that has no points, so we must interpolate values between those two points. However, it is not possible to know if the lines will cross the diffractor operator, so we established a range of 1 to 5%. This means that if we do not have the value 60% we look from plus 5% and minus 5%. In practice we look values from 57 to 63% around the line which is represented by the green dots in the last part in figure 23.

Once the two pairs of three parameters are calculated for each training data set, the data test has its attributes calculated. Each line of the data set is used to construct a matrix of the same size of the training matrix. Then, $finder$ calculates the Euclidean and Mahalanobis distance from this two matrices. The algorithm puts in ascending order the calculated distances and chooses the first one if k=1. If you choose the 3 k nearest neighbors, $finder$ will select the 3 lines and it will assign the one with more frequency. The same approach for k=5.

Figure 23 – Finder Steps.



Source: From author