Journal of Nondestructive Evaluation (2018) 37:40
https://doi.org/10.1007/s10921-018-0496-y

Welding Defect Classification from Simulated Ultrasonic Signals

Raphaella H. F. Murta[1] · Flávison de A. Vieira[1] · Victor O. Santos[1] · Elineudo P. de Moura[1]

Received: 22 January 2018 / Accepted: 16 May 2018 / Published online: 22 May 2018

Abstract
Nondestructive testing is widely used to detect and to size up discontinuities embedded in a material. Among the several ultrasonic techniques, time of flight diffraction (TOFD) combines high speed inspection, high sizing reliability and low rate of incorrect results. However, the classification of defects through ultrasound signals acquired by the TOFD technique depends heavily on the knowledge and experience of the operator and thus, this classification is still frequently questioned. Besides, this task requires long processing time due to the large amount of data to be analyzed. Nevertheless, computational tools for pattern recognition can be employed to analyze a high amount of data with large efficiency. In the present work, simulation of ultrasound propagation in two-dimensional media containing, each one, different kinds of modeled discontinuities which mimic defects in welded joints were performed. Clustering (k-means) and classification (principal component analysis and k-nearest neighbors) algorithms were employed to associate each simulated ultrasound signal with its corresponding modeled defects. The results for each method were analyzed, discussed and compared. The results are very promising.

Keywords Ultrasound · TOFD · Welding defects · K-NN · Principal component analysis · K-means

1 Introduction

Nondestructive tests are used to detect discontinuities inside an object without affecting its future usefulness. Some of them, like ultrasound testing, are even able to determine the dimension of the discontinuity. Ultrasound evaluation uses high frequency mechanical waves to detect imperfections inside the material.

Usually, echo amplitudes are related to discontinuity dimensions. The TOFD technique is not based on echoes amplitude, but uses the travel time of a diffracted wave at the upper and lower tips of a discontinuity to determine its size and depth.

Although this technique presents high detection rate and a wide application field, the large amount of data generated during an inspection requires a long time to be properly analyzed and depends heavily on the knowledge and experience of the operator. Nevertheless, a large number of signals can be quickly processed by classification and clustering algorithms in order to distinguish the main defects in welded joints detected by the TOFD technique.

Supervised and unsupervised learning algorithms can be used with this purpose [1–3]. Cluster analysis or clustering is an unsupervised algorithm that aims to determine the best way of grouping a given unknown dataset. It is expected that elements of a same group are similar and different elements of different groups are as different as possible. In supervised learning, the classifier is informed about the class to which each training signal belongs. After have been taught, these algorithms can be used to recognize patterns of unknown data.

In this study, the non-destructive test by ultrasound was modeled in a two-dimensional medium containing one of three common types of welding defects (lack of penetration, pore and crack). For each type, ultrasound signals were obtained by simulating 36 different configurations of defect size and position. A total of 108 simulated signals were generated.

A cluster analysis technique was employed with the purpose of verifying the optimum number of groups to divide simulated ultrasonic signals. Also, classifiers with supervised learning algorithm (k-Nearest Neighbors) were employed to classify these signals.

✉ Elineudo P. de Moura
elineudo@ufc.br

[1] Departament of Metallurgical and Materials Engineering, Federal University of Ceará, Bloco 729, Fortaleza 60440-554, Brazil

This work aims at evaluating the performance of classification and clustering algorithms in the identification of modeled welding defects, through the analysis of ultrasonic signals obtained by TOFD simulation, independently of its dimension and position.

2 Modeling and Simulation

A finite element software package was used to simulate the propagation of ultrasonic waves in modeled two-dimensional media, each one containing one discontinuity which mimics a welding defect.

2.1 Ultrasonic Transducers

The TOFD technique uses normal transducers for longitudinal waves mounted on acrylic wedges in order to obtain an oblique ultrasonic beam [4]. However, modeling these wedges requires increase in system size and computation cost. The working principle of ultrasonic phased array technique can be considered as alternative solution for this problem. Phased array transducers are based on constructive and destructive interference of waves generated by a number of elements excited at slightly different times.

Wooh and Shi [5] carried out a simulation study on the application of linear phased array transducer to produce a steered beam. They analyzed the influence of various transducer parameters and determined the most important variables to be adjusted in order to obtain a good beam steering. Their work emphasizes that good quality is achieved when the beam is sharply-defined and well-directed towards the desired steering direction, while suppressing or squelching the deleterious grating waves travelling in other directions.

Firstly, a better beam direction and a higher pressure in the steering direction can be achieved by increasing the number of elements, where which element is formed by a set of point sources. Secondly, the beam direction can be improved by increasing the inter-element spacing, d_{cr}. An optimal value for the inter-element spacing, d_{cr}, is given by Eq. 1 [5], and can be used to provide the best beam directivity without introducing deleterious grating lobes. Finally, the pressure in the steering direction increases with the element width, a, resulting in high signal-to-noise ratios.

$$d_{cr} = \frac{\lambda}{1 + \sin(\theta_s)_{max}} \tag{1}$$

Moreover, previous works states that better spatial resolutions are achieved for minor angle of incidence [6]. However, experimental measurements indicate that 60° transducers produce the best results [7].

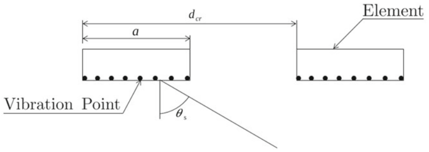

Fig. 1 Schematics representation of the simulated transducer

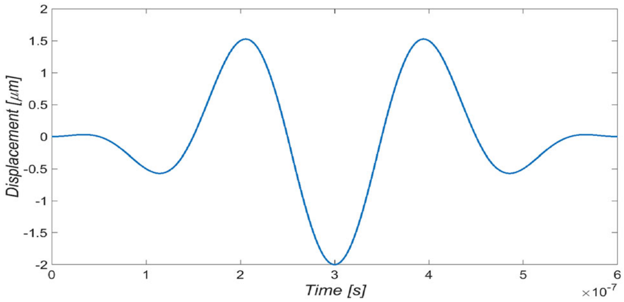

Fig. 2 Ultrasonic pulse with 5 [MHz]

With all this in mind, the desired ultrasonic angle beam inside the material wasn't obtained by using wedge modeling, but by exciting different points on the surface in an appropriate linear time delay.

Thus, an array of eight elements placed on the surface of the model was employed to produce ultrasonic excitation. The element width, a, and the inter-element spacing, d_{cr}, were defined as 0.2λ and 0.5λ, respectively. Each element is composed by eight single points sources of radiation, resulting in 64 point sources. The Fig. 1 shows a schematics representation of the simulated transducer.

The ultrasonic pulse given by Eq. 2 was applied at each element in both x and y directions with an appropriate time delay to produce an angle of incidence defined as being 60°. This pulse has been used in previous works [8] and its waveform is shown in Fig. 2. A frequency, f, of 5 MHz was used.

$$F(t) = \begin{cases} \left[1 - cos\left(\frac{2\pi f t}{3}\right)\right] cos(2\pi f t), & 0 \leq t \leq \frac{3}{f} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

2.2 TOFD Technique

The conventional configuration for the TOFD technique consists of emitter and receiver transducers aligned on either side of the weld bead, so that the region of interest is entirely within the area covered by the sonic emitter.

In a two-dimensional simulation, length and height (thickness) of the solid medium need to be defined. The relationship between the height, the incidence angle of ultrasonic beam, θ_s, and the distance between emitter and receiver transducers, d_{tr}, is given by equation:

$$d_{tr} = 2L \tan\theta_s \tag{3}$$

Table 1 Material properties

Properties	Value
Density	7800 kg/m^3
Longitudinal wave velocity	5900 m/s
Transverse wave velocity	3200 m/s

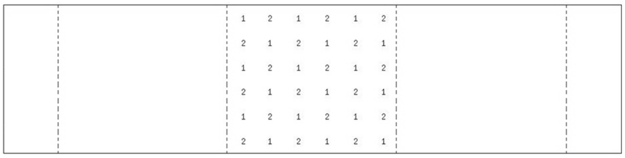

Fig. 3 Map of pores position

A height L of 19 mm and a beam incidence angle θ_s of 60° were used in Eq. 3 to yield $d_{tr} = 80$ mm.

The receiver has the same number of elements and point sources defined to emitter. Likewise, the same time delay is used to read the sound pressure in different nodes in the receiver, but in the opposite sense, that is, opposite time difference. For each simulation, the time variation of sound pressure read in these sixty-four points of the receiver are time adjusted and combined to generate one single A-scan.

2.3 Model and Material Properties

The properties of the simulated material are those of the AISI 1020 steel and shown in Table 1. In the numerical simulation the material is considered to be isotropic, so the welded-joint properties are not considered different to the base metal.

2.4 Characteristics of Discontinuities

Three kinds of discontinuities found in welded joints were modeled: pore, crack and lack of penetration. For each one of them, 36 simulations of wave propagation were performed by changing the position or dimensions of the discontinuity. Thus, a total of 108 A-scan signals were produced. Particular rules were used to define the position and size for each kind of discontinuity.

Pore-type discontinuities with 1 and 2 mm of diameter positioned as shown in Fig. 3 were used in wave propagation simulations. The number positions in the Fig. 3 indicate the exact location of the pore and its corresponding diameter.

A unique pore, as the present in the cross-section depicted by Fig. 4, is embedded in the medium at a time and before each simulation. The combination of two pore sizes and eighteen positions yielded thirty-six simulations, with one A-scan for each simulation.

Crack-type discontinuities were modeled with 6 different sizes. Crack heights from 4 to 9 mm, in 1 mm steps, were used. The width dimension was defined as being equal to one tenth of the height. As in Fig. 3, Fig 5 shows the map of cracks positions, where the numbers indicate the position and the height of each crack-type discontinuity.

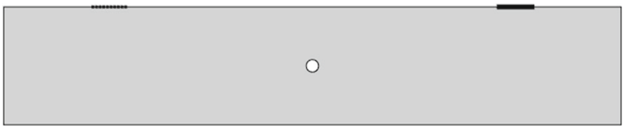

Fig. 4 Pore-type discontinuity model

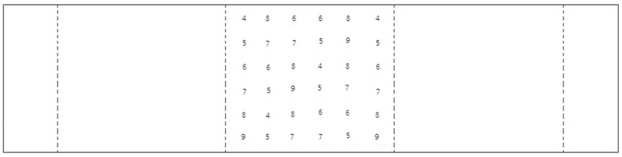

Fig. 5 Map of cracks positions

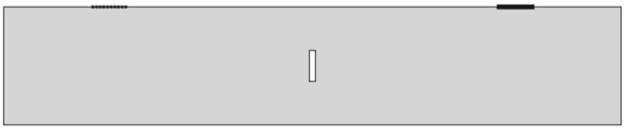

Fig. 6 Crack-type discontinuity model

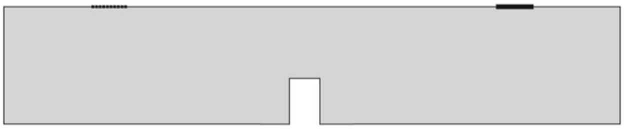

Fig. 7 Lack of penetration model

Each simulation has one crack-type discontinuity as shown in Fig. 6. Combining the dimensions and positions, a total of 36 simulated A-scan of crack-type discontinuities are generated.

Lack of penetration is a weld defect found at the backwall of a weld bead. This kind of discontinuity was added always in the mid point between emitter and receiver transducers, as shown in Fig. 7. Height and width were changed simultaneously with height changed from 3 mm to 8 mm, in 1 mm steps, and width changed from 1 mm to 6 mm, with an increment of 1 mm, yielding 36 different combinations.

2.5 Simulation Data

Each wave propagation simulation yielded an A-scan signal corresponding to the sound pressure amplitude measured by the receptor. This way, 108 ultrasonic trials were simulated and each simulation produced an A-scan signal with 512 points. The simulated ultrasonic signals were analyzed by classification and clustering algorithms to evaluate their ability to correctly assign a particular signal to one of the 3 weld defects.

A triangular element free meshing was used for the simulation of ultrasonic wave propagation. Parameters associated with the convergence of the numerical solution were tested. Values of $\lambda_L/\Delta x_{max} = 8$ or more and time steps around $1/(100f)$ were suggested in reference [9]. After the mesh refinement study, the maximum size of triangular element was defined as $\lambda_L/\Delta x_{max} = 12$ and the time step used equal to $1/(100f)$.

3 Preprocessing

Besides the simulated dataset of 108 A-scans, two other sets of signals were produced through preprocessing techniques. One of these sets consisted in calculating an envelope for each of the 108 signals through the application of the Savitzky-Golay digital filter [10]. It has been shown that this procedure improves the average success rate of the classifier [11]. The second preprocessed set consists of the normalization of the simulated signals in order to obtain signals with an average of zero and a standard deviation of one. The bare, enveloped and normalized sets of signals were analyzed using principal component, K-means and K-nearest neighbors techniques.

4 Pattern Recognition

The classification of the three datasets was carried out by two different algorithms. The first classifier was implemented by using principal component analysis combined with the nearest-class-mean rule. The second classifier was performed by k-nearest neighbors algorithm. For the clustering approach, the k-means method was applied.

The goal is to evaluate the performance of such algorithms in the identification of the weld defects based on the analysis of the ultrasound signal obtained by simulation of wave propagation in a medium with the discontinuity embedded.

The following subsections present a review of principal component analysis, K-Nearest Neighbours and K-means.

4.1 Principal Component Analysis

Consider a set of p possibly correlated observations, each represented by an m-dimensional vector. Principal component analysis (PCA) is a mathematical tool that converts the set of observations into linearly uncorrelated data by a rotation of the original data onto a new set of orthogonal axes, diagonalizing the covariance matrix of the data.

Assume that the input data is arranged in a $m \times p$ matrix X, where p is the number of signals produced and m is the number of dimensions (independent measurements) of each signal. Each column of X represents an observation. With respect to an arbitrary coordinate system, the covariance matrix is explicitly given by

$$\Sigma = \begin{bmatrix} \hat{V}ar(x_1) & \hat{C}ov(x_1x_2) & \cdots & \hat{C}ov(x_1x_m) \\ \hat{C}ov(x_2x_1) & \hat{V}ar(x_2) & \cdots & \hat{C}ov(x_2x_m) \\ \hat{C}ov(x_3x_1) & \hat{C}ov(x_3x_2) & \cdots & \hat{C}ov(x_3x_m) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{C}ov(x_mx_1) & \hat{C}ov(x_mx_2) & \cdots & \hat{V}ar(x_m) \end{bmatrix}, \tag{4}$$

where

$$\hat{C}ov(x_ix_j) = \sum_{k=1}^{p} \frac{(x_{i,k} - \bar{x}_i)(x_{j,k} - \bar{x}_j)}{p} \tag{5}$$

and

$$\hat{V}ar(x_i) = \sum_{k=1}^{p} \frac{(x_{i,k} - \bar{x}_i)^2}{p}, \tag{6}$$

$\bar{x}_i$ being the average of the elements in the ith row of X, i.e. the average value of the ith measurement over all observations.

The covariance matrix calculated as in Eq. 4 is then diagonalized to extract its eigenvalues and eigenvectors. The eigenvalues λ of the covariance matrix Σ are the roots of [12]

$$\|\Sigma - \lambda I\| = 0, \tag{7}$$

while the corresponding eigenvectors $\mathbf{v}$ are obtained from the solutions of

$$(\Sigma - \lambda I)\mathbf{v} = 0. \tag{8}$$

The eigenvalues of Σ are arranged in decreasing order, such that $\lambda_1 > \lambda_2 > \ldots > \lambda_m$. The first eigenvector, $\mathbf{v}_1$, which is associated with the largest eigenvalue of Σ, accounts for the largest variation in the original data. The projections of the original observations onto the direction of the nth eigenvector of Σ define the nth principal component of the data.

Principal component analysis naturally provides a scheme for truncating the m-dimensional space of the original data, keeping a number n of principal components such that the sum of the corresponding eigenvalues reaches a desired precision r defined by [13]

$$r = \frac{\sum_{i=1}^{n} \lambda_i}{\sum_{i=1}^{m} \lambda_i} \leq 1. \tag{9}$$

The classifier based on PCA was implemented following the same procedure adopted in Ref. [14]. The signals (their principal components, determined previously by PCA) were divided into a training set and a testing set. The training set is used to determine the average of the class for each class of discontinuity. The classification was performed by the nearest-class-mean rule, according to which each test signal (its first few principal components) is assigned to the class whose average signal lies closer to the test signal.

4.2 K-Means

The k-means clustering separates the data in different numbers of clusters and, using some criteria, as Silhouette [15] and Davies–Bouldin [16] indexes, it decides the optimum number of sets to group. The number of clusters tested varies from two (it doesn't make sense to create a single group with the whole dataset) to the square root of the number of input data.

The procedure is started by randomly choosing k of the input data as the initial centroids and associating each data point to the nearest centroid. After that, the positions of the k cluster's centroids are recalculated as the mean of its data resulting from the previous step. This procedure is repeated until the change in the centroids position is below a threshold, or until the predefined iteration number is reached [17]. The k-means method groups the data according to the Euclidean distance.

In this work, the number of clusters k was varied from 1 to 11 (approximately the square root of 108). For each value of k, a cost function was evaluated. The cost function was calculated based on the distance of each element to the centroid of the group to which it is associated and is given by:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{108} \left\| x_i^{(j)} - c_j \right\|^2 \tag{10}$$

where $x_i^{(j)}$ data i belongs to the group j and c_j is the centroid j. The goal of this method is to minimize the value of this function.

For the k-means approach, we used both Silhouette and Davies–Bouldin indexes to suggest the optimum number of cluster as being the minimum number of grouping proposed by them. The classification's quality is better when the Silhouette index is closer to one and the Davies–Bouldin index is closer to zero [18].

The Silhouette criterion verifies how near the elements of the same group are to the same set and how far the elements are from the nearest different group. To obtain this index, it is necessary to calculate the distance of one element i to all the elements of the same group (a_i) and the distance of one element i to all the elements of the nearest group (b_i). Equation 11 gives the index [15].

$$s = \sum_{i=1}^{108} \frac{b_i - a_i}{max(a_i, b_i)} \tag{11}$$

To obtain the Davies–Bouldin index, it is necessary to calculate all the k distances between the centroids of the cluster i to all the elements of the group i (represented by d_i) and the distance to each pair of centroids i and j (represented by $d(c_i, c_j)$). The index is given by:

$$DB = \frac{1}{k} \sum_{i=k, i \neq j}^{108} max \left(\frac{d_i - d_j}{d(c_i, c_j)} \right) \tag{12}$$

Through the relative comparison between different divisions of the set, it is possible to define the best way to group the data [19].

4.3 K-Nearest Neighbours

The k-nearest neighbours is a simple and easy-to-apply supervised algorithm. This algorithm is divided into two phases: training phase and testing phase. In the training phase, the data and its correct classification are provided to the algorithm. Then, in the testing phase, data are presented to the algorithm that returns the corresponding suggested class.

The first step to classify the test data is to calculate all the distances between the training data to a particular datum. Likewise the k-means method, the Euclidean distance was used. After that, the class to which the k nearest neighbours belongs is verified. So, the data is classified as the most frequent class among the k nearest neighbours.

To choose the optimal k, it is advised to test different values and then choose the minimum training error, given by:

$$Error = \frac{N_{inc}}{N_{tes}} \tag{13}$$

Where N_{inc} corresponds to the number of data incorrectly classified and N_{tes} to the number of tested data set.

5 Results

5.1 Simulation Analysis

Figures 8 and 9 are two different instants regarding the wave propagation for one of the cases of lack-of-penetration. The blue and red colors represent the amplitude of the ultrasonic wave. The arrow across in Fig. 8 aims at the longitudinal wave and the arrow down aims at the transversal wave. As shown, the longitudinal wave is correctly propagating towards the discontinuity.

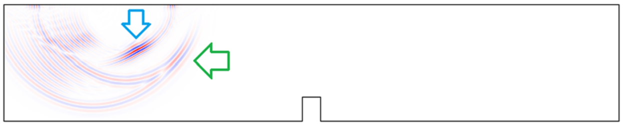

Fig. 8 Wave propagation a short time after the source emission

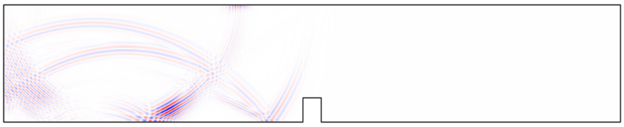

Fig. 9 Longitudinal wave impinging on the discontinuity

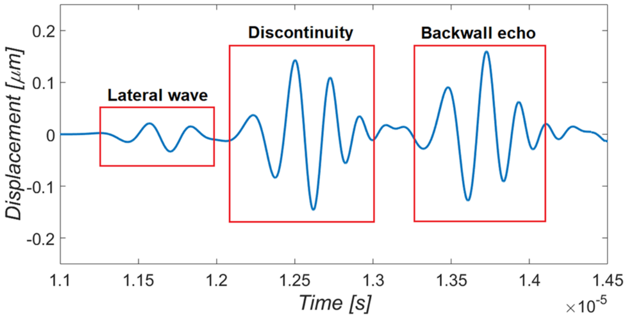

Fig. 10 Lack of penetration discontinuity A-scan example

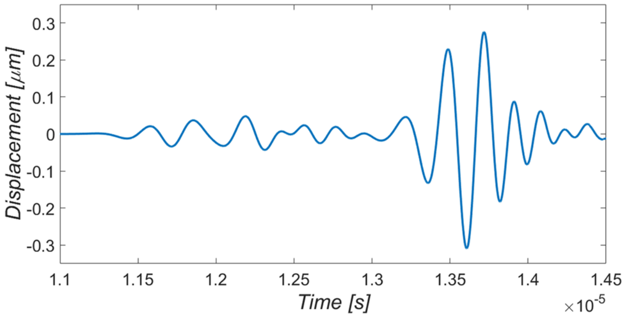

Fig. 11 Pore-type discontinuity A-scan example

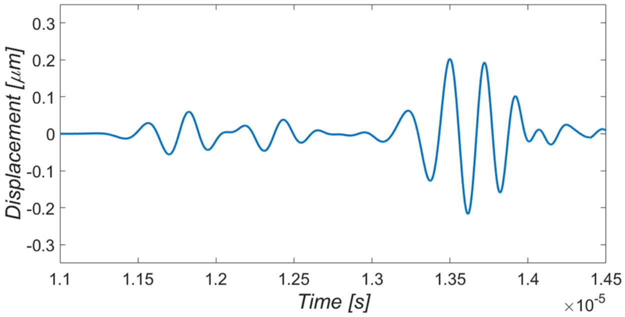

Fig. 12 Crack-type discontinuity A-scan example

In Fig. 9, the longitudinal wave front is near the discontinuity. Clearly, a part of the longitudinal wave has already reflected due the zero displacement boundary at the bottom of the material, the same will occur with the transverse wave.

As the longitudinal wave passes the points that correspond to the receiver, the pressure wave is recorded and, from that, the A-scan is generated. The Fig. 10 shows the A-scan from one of the 36 lack-of-penetration simulations. It is clear that three pulses are present. From left to right, they represent the lateral wave, the discontinuity and the backwall echo.

Figure 11 is the result from a pore-type simulation. Here, the three pulses are not clear as in Fig. 10. The backwall echo is the most evident pulse with, approximately, the same amplitude and the same time-of-flight as in the lack-of-penetration case. The lateral and discontinuity waves are not clear probably due an interference between them.

For the crack-type case, the Fig. 12 shows a similar case to Fig. 11. The backwall echo is well-defined and as well there is a small interference between the lateral and discontinuity waves.

5.2 Principal Component Analysis

The PCA of the three data sets discussed above were performed. A classifier was implemented by applying the nearest-class-mean rule to the first few principal components of the data. The average percentage of discontinuities which were correctly classified and the kept information rate for the bare, normalized and enveloped ultrasonic signals are shown, respectively, in Figs. 13, 14, and 15, as function of the number of principal components used in the analysis. A maximum of 20 principal components was used.

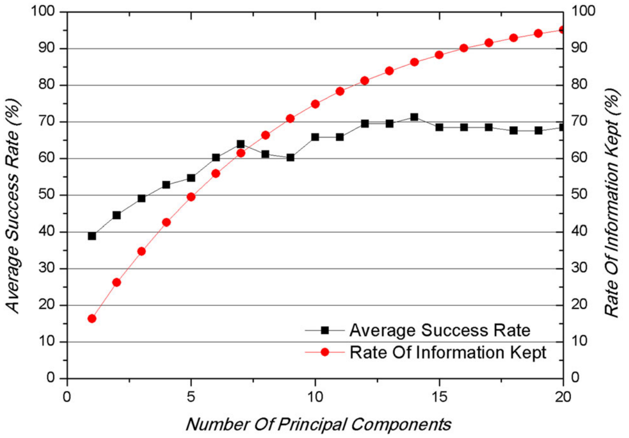

Fig. 13 Success rate and kept information rate for the PCA of bare ultrasonic signals

In this context, kept information rate is defined as the ratio of the sum over the PCA eigenvalues used to the sum over all PCA eigenvalues. The minimum amount of information kept in order to properly classify data is not a settled matter [3].

While 70–90% of information kept could be enough, it can be argued that it should vary according to the problem

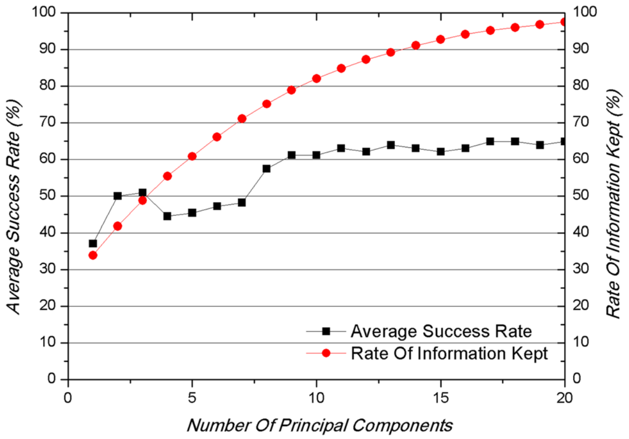

Fig. 14 Success rate and kept information rate for the PCA of normalized ultrasonic signals

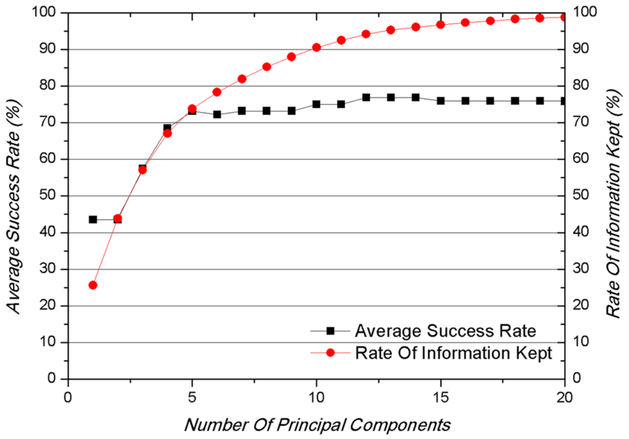

Fig. 15 Success rate and kept information rate for the PCA of the enveloped ultrasonic signals

under study. A suggestion is to take the point in the eigenvalues spectrum where there is a significant drop before the spectrum stabilizes at small values [3].

From the data shown in Figs. 13, 14, and 15, it can be seen that the minimum number of principal components required to reach hit ratio convergence, for bare, normalized and enveloped data sets, is 12 (with hit ratio peak of about 71% at 14 principal components), 11 (with three peaks of about 65% at 17, 18 and 20), and 5 (with three peaks of about 77% at 12, 13 and 14), respectively. Also, the information kept rate at the onset of convergence is about 81, 85, and 74%, respectively.

Table 2 shows confusion matrices obtained from the principal component analysis of the three data sets. In a confusion matrix, the cells in the main diagonal show the percentage of correct classification while the off-diagonal cells show the percentage of misclassifications. The data shows that the PCA classification of the bare ultrasonic signals mistakes lack of penetration with cracks, but not with pores. Correct classification of pores and cracks (61.11 and 66.67%) is less effective than that of lack of penetration (86.11%). Pores and cracks also misclassified as both other deffects.

Table 2 Confusion matrices obtained by applying the nearest-class-mean rule to the results of PCA

	LP (%)	PO (%)	CR (%)
Bare ultrasonics signals			
LP	86.11	16.67	19.44
PO	0.00	61.11	13.89
CR	13.89	22.22	66.67
Normalized ultrasonic signals			
LP	80.56	22.22	13.89
PO	5.55	47.22	19.44
CR	13.89	30.56	66.67
Enveloped ultrasonic signals			
LP	100.00	8.33	2.78
PO	0.00	66.67	33.33
CR	0.00	25.00	63.89

LP lack of penetration, *PO* pore, *CR* crack

PCA classification of the normalized ultrasonic signals is worse for all defects. Pore was the most mistaken discontinuity for this data set. PCA analysis of the enveloped signals, on the other hand, provided the best overall classification rates. Lack of penetration was classified with 100% of success while the success rate of pore classification was the highest among the three data sets. The classification rate of cracks was just a little lower than the ones obtained from the other two data sets.

The overall percentage found was 71.30, 64.82 and 76.85% for the bare, normalized and enveloped ultrasonic signals, respectively. The data and analysis presented leaves no doubt that the best approach for classification of discontinuities is by using PCA with enveloped ultrasonic signal. It showed faster convergence and better hit ratios with only five principal components.

5.3 K-Means

To determine the optimal number k of groups, the algorithm tested all values from 2 to 11. For each value, the algorithm was repeated 100 times to avoid any local minimum given by the random choice of k centroids. For each of the three sets of signals the number of groups suggested by Silhouette and Davies–Bouldin indexes was tested to find the optimum index.

For the grouping of bare ultrasonic signals, the optimum index shows that two groups is the best way to separate the data, as shown in Fig. 16. Although we simulated three different groups of discontinuity, the algorithm discriminated the data in just 2 groups with no distinguished pattern. The

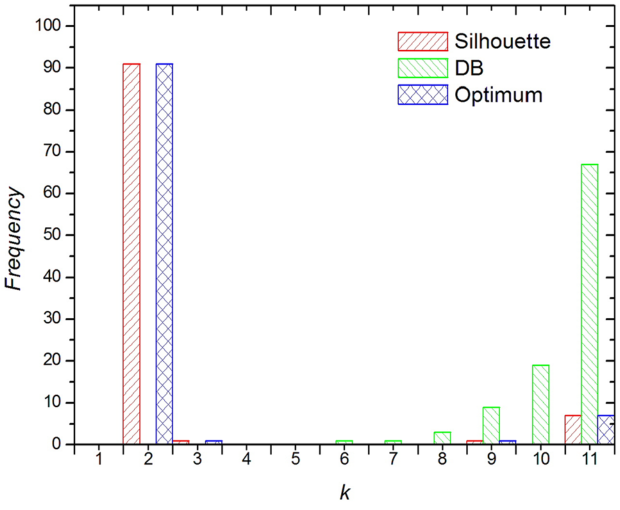

Fig. 16 Frequencies of values suggested as the optimal amount of clusters to group the bare ultrasonic signals

Table 3 Grouping of ultrasonic signals suggested by k-means algorithm

	LP	PO	CR
Group 1	25	6	12
Group 2	11	30	24

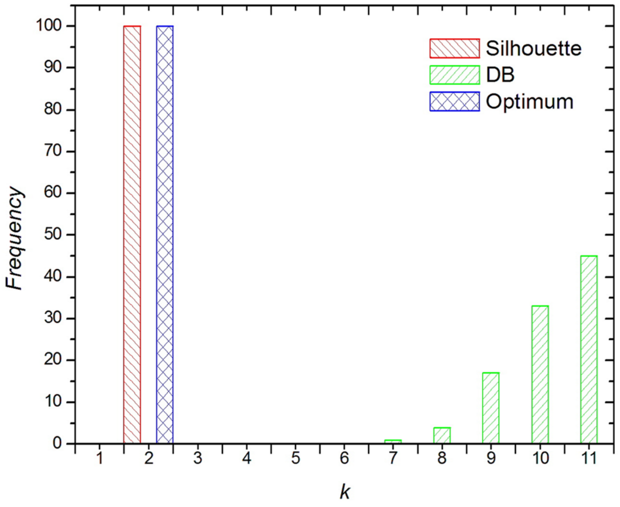

Fig. 17 Frequencies of values suggested by our test as the optimal amount of clusters to group the normalized ultrasonic signals

algorithm classified 70% of lack of penetration signals in group 1 and 75% of pore and crack signals in group 2. Table 3 shows the grouping suggested by k-means.

The optimum index obtained for the grouping of normalized ultrasonic signals was the same as the one obtained for the bare signals, with two groups being the recommended way to separate the data, as shown in Fig. 17.

The k-means algorithm grouped the normalized signals in a similar way as the bare signals. Only one lack of penetration signal had its classification changed from group 1 to group 2 when using the normalized signals. Table 4 shows the grouping suggested by k-means.

Table 4 Grouping of normalised ultrasonic signals suggested by k-means algorithm

	LP	PO	CR
Group 1	24	6	12
Group 2	12	30	24

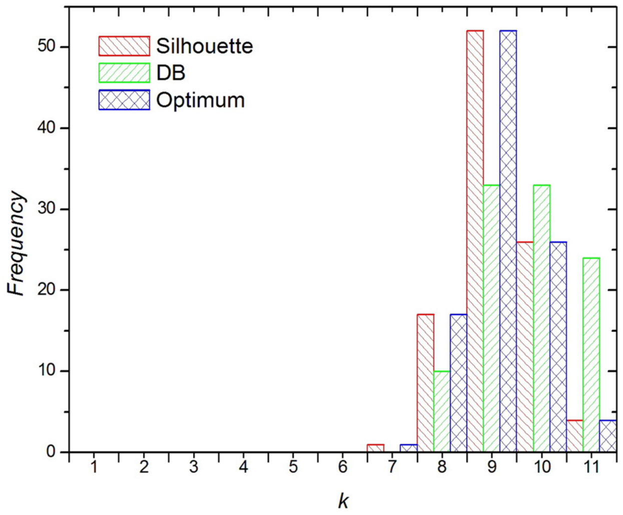

Fig. 18 Frequencies of values suggested by our test as the optimal amount of clusters to group the enveloped ultrasonic signals

For the grouping of enveloped ultrasonic signals the optimum index suggests 9 groups as the best way to separate the data, as shown in Fig. 18.

The k-means analysis reveals the existence of a pattern corresponding to {1;1;2;2;3;3;1;1;2;2;3;3;1;1;2;2;3;3; 1;1;2;2;3;3;1;1;2;2;3;3;1;1;2;2;3;3} in the enveloped ultrasonic signals corresponding to lack of penetration. Therefore, the signals were subdivided in three subsets according to the depth of the discontinuity (shallow, medium and deep) and independently of its width. Table 5 shows the suggested classification for the enveloped ultrasonic signals using k-means.

Group 1 is comprised of all the enveloped signals of lack of penetration with 2 mm and 3 mm of depth, while those with 4 mm and 5 mm of depth were joined in group 2. All the enveloped signals of lack of penetration with 6 and 7 mm were associated with group 3. Groups 4, 5, 6, 7 and 8 joined the enveloped signals of pore and crack. Only 5 enveloped signals of cracks belong to the group 9.

5.4 K-Nearest Neighbor

To apply the k-nearest neighbor (k-NN) classifier, a study about the influence of the number of neighbors *k* in the success rate was carried out for each dataset.

Each dataset was randomly divided into a training set and a testing set. The training set contained 80% of the signals of

Table 5 Grouping of envelopes ultrasonic signals suggested by k-means algorithm

	LP	PO	CR
Group 1	12	0	0
Group 2	12	0	0
Group 3	12	0	0
Group 4	0	3	5
Group 5	0	11	12
Group 6	0	11	4
Group 7	0	5	6
Group 8	0	6	4
Group 9	0	0	5

Table 6 Confusion matrix obtained by k-NN for bare ultrasonic signals

	LP (%)	PO (%)	CR (%)
LP	98.00	8.00	5.14
PO	1.29	57.86	48.57
CR	0.71	34.14	46.29

each class, while the testing set contains the remaining 20%. Confusion matrices, described in Sect. 5.2, were calculated for an average taken over 100 randomly chosen sets of events. This allows to compare the k-NN results with those obtained by principal component analysis.

The number k was varied from 1 to 4 and the study showed that the better classifications were obtained with $k = 1$, for all datasets, showing that that a datum normally belong to the same class of the nearest datum.

Table 6 shows the confusion matrix obtained for the bare ultrasonic signal. From the results portrayed in Table 1, it is possible to observe that the lack of penetration defect (LP) had the best success rate (98%), indicating that this is the most easily separable class. The approach used can satisfactorily separate lack of penetration discontinuities from pores and cracks.

However, the same doesn't happen for pores and cracks. Only 57.86% of the pores were correctly classified as pores, whereas 34.14% were classified as cracks and 8% were as lack of penetration. Finally, 46.29% of the cracks were correctly classified as cracks, whereas 48.57% were classified as pores and 5.14% classified as lack of penetration. The overall percentage of bare signals which were correctly classified (calculated over the principal diagonal) achieved 67.38%.

Table 7 shows the results using the normalized ultrasonic signals. Here, the lack of penetration discontinuity is not as well classified as in Table 6. On the other hand, the classification success rate was slightly better for pores and cracks, achieving 61.57 and 56.29% of success, respectively. However, the overall performance was 70.71%. Therefore, normalizing the ultrasonic signal has not shown significant impact in the classification.

Table 7 Confusion matrix obtained by k-NN for normalised ultrasonic signals

	LP (%)	PO (%)	CR (%)
LP	94.29	8.86	3.86
PO	0.86	61.57	39.86
CR	4.86	29.57	56.29

Table 8 Confusion matrix obtained by k-NN for enveloped ultrasonic signals

	LP (%)	PO (%)	CR (%)
LP	100.00	3.00	1.29
PO	0.00	57.43	36.71
CR	0.00	39.57	62.00

Results of the k-NN classification for enveloped ultrasonic signals are shown in Table 8. All lack of penetration were correctly classified, but some pores and cracks were misclassified as lack of penetration. The classification success rate achieved was 57.43 and 62% to pores and cracks, respectively. Most classification errors correspond to pores classified as cracks (39.57%), and vice-versa (36.71%). A good average success rate classification (73.14%) has been achieved by k-NN classifier for the enveloped ultrasonic signals.

6 Conclusion

The numerical simulation of two-dimensional medium was successfully realized, considering the characteristics inherent the ultrasound test like density and the wave propagation velocity inside the material. Three types of well-defined welding defects (lack of penetration, pore and crack) were modeled and 36 wave propagation simulations were performed to produce a set of typical A-scan signals for each kind of discontinuity studied.

These simulated ultrasonic signals were analyzed by pattern recognition techniques and clustering algorithms.

The classifier based on principal component analysis presented performance similar to the one implemented using the k-nearest neighbors algorithm. The PCA with enveloped signals resulted in an overall success rate of 76.85%. For k-nearest neighbors, the success rate obtained was 73.14%.

In the unsupervised case, although the optimum number of groups suggested by k-means to divide the datasets differs from the number of classes regarded in modeling, the clustering obtained reveals some pattern. Enveloped ultrasonic signals were grouped in 9 groups and the lack of penetration discontinuity was equally divided in 3 of these groups in accordance to depth. The division was not so good for pores and cracks.

Moreover, the best average success rate classification was achieved for the enveloped ultrasonic signals in all classification schemes.

The lack of penetration discontinuity was the most easily separable class for all classification schemes used, most likely due the following reasons: (1) the backwall echo position is slightly altered by this type of defect and; (2) dimensions, shape and position are markedly different for lack of penetration defects than for pores and cracks. Cracks and pores are similar and thus harder to be distinguished.

References

1. Kaufman, Leonard, Rousseeuw, Peter J.: Finding Groups in Data: An Introduction to Cluster Analysis, vol. 344. Wiley, New York (2009)
2. Haykin, S.S.: Neural Networks and Learning Machines, vol. 3. Pearson, Upper Saddle River, NJ (2009)
3. Webb, Andrew R.: Statistical Pattern Recognition. Wiley, New York (2003)
4. Silk, M.G., Lidington, B.H.: Defect sizing using an ultrasonic time-delay approach. Br. J. Non-Destr. Test. **17**(2), 33–36 (1975)
5. Wooh, Shi-Chang, Shi, Yijun: A simulation study of the beam steering characteristics for linear phased arrays. Non-Destr. Eval. **18**(2), 39–57 (1999)
6. Ogilvy, J.A., Temple, J.A.G.: Diffraction of elastic waves by cracks: application to time-of-flight inspection. Ultrasonics **21**(6), 259–269 (1983)
7. Temple, J.A.G.: Predicted ultrasonic responses for pulse-echo inspections. Br. J. Non-Destr. Test. **28**(3), 145–154 (1986)
8. Baskaran, G., Lakshmana Rao, C., Balasubramaniam, K.: Simulation of the tofd technique using the finite element method. Insight-Non-Destr. Test. Cond. Monit. **49**(11), 641–646 (2007)
9. Ghose, B., Balasubramaniam, K., Krishnamurthy, C.V.,Rao, A.S.: Two dimensional fem simulation of ultrasonic wave propagation in isotropic solid media using comsol. In COMSOL Conference (2010)
10. Orfanidis, S.J.: Introduction to Signal Processing. Prentice-Hall, Inc., Upper Saddle River (1995)
11. de Moura, E.P., Siqueira, M.H.S., da Silva, R.R., Rebello, J.M.A.: Welding defect pattern recognition in tofd signals part 2. Non-linear classifiers. Insight-Non-Destr. Test. Cond. Monit. **47**(12), 783–787 (2005)
12. Jolliffe, Ian: Principal Component Analysis. Wiley Online Library, New York (2002)
13. Varella, C.A.A.: Análise de componentes principais. Universidade Federal Rural do Rio de Janeiro, Rio de Janeiro (2008)
14. de Moura, E.P., Vieira, A.P., Gonçalves, L.L.: Fluctuation analyses for pattern classification in nondestructive materials inspection. EURASIP J. Adv. Signal Process. **1**, 262869 (2010)
15. Rousseeuw, P.J.: Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
16. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **2**, 224–227 (1979)
17. Anil, K.: Jain. Data clustering: 50 years beyond k-means. Pattern recognition letters **31**(8), 651–666 (2010)
18. Petrovic, S.: A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In Proceedings of the 11th Nordic Workshop of Secure IT Systems, pp 53–64 (2006)
19. Dimitriadou, E., Dolničar, S., Weingessel, A.: An examination of indexes for determining the number of clusters in binary data sets. Psychometrika **67**(1), 137–159 (2002)