




Identification of Corrosive Substances and Types of Corrosion Through Electrochemical Noise Using Signal Processing and Machine Learning

Lorraine Marques Alves¹ · Romulo Almeida Cotta¹ · Patrick Marques Ciarelli¹  · Evandro O. T. Salles¹ · Klaus F. Côco¹ · Jorge L. A. Samatelo¹

Received: 18 June 2018 / Revised: 22 August 2018 / Accepted: 8 October 2018 / Published online: 23 October 2018
© Brazilian Society for Automatics–SBA 2018

Abstract

Several systems in industries are subject to the effects of corrosion, such as machines, structures, and a lot of equipment. As consequence, the corrosion can damage structures and equipment, causing financial losses and accidents. Among the most common types is the localized corrosion, and it is present in most industrial processes and is the most difficult to detect. Such consequences can be reduced considerably with the use of methods of detection, analysis and monitoring of corrosion in hazardous areas, which can provide useful information to maintenance planning and accident prevention. In this work, we analyze some features extracted from electrochemical noise for the classification of different types of localized corrosion. Furthermore, we use some techniques to identify corrosive substances that may cause corrosion in materials. For both tasks, we apply signal processing and machine learning techniques. Experimental results show that the features obtained using wavelet transform and recurrence quantification analysis are effective to solve both tasks: the corrosion identification and the classification of substances. Almost all evaluated machine learning techniques achieved an average accuracy above 90%.

Keywords Corrosion · Electrochemical noise · Machine learning · Wavelet transforms · Recurrence quantification analysis

1 Introduction

Several important systems in the industrial field are subject to the effects of corrosion, such as means of transportation, storage tanks, heat exchangers, reactors, so that the corrosion can cause the deterioration of structures and a lot of equipment, as well as accidents (Gentil 2003). Furthermore, corrosion can be a source of unplanned costs. The global cost of corrosion is estimated around U\$ 2.5 trillion, equivalent to 3.4% of world gross national product (GDP) (Koch et al. 2016).

Fortunately, due to the simultaneous occurrence of oxidation and reduction reactions during the corrosion process, the current and electrical potential fluctuations can be measured on the surfaces that are suffering this process. These measured signals are called electrochemical noise (ECN).

An example of application of ECN signals is the identification of corrosive substances, that can be useful to troubleshoot faults in industrial processes, assist in maintenance planning and even avoid accidents. Another application is the identification of types of localized corrosion that are common in most industrial processes, but they are hardly detected using traditional electrochemical techniques.

The identification of substances and corrosion types that are affecting the metal surface enables the planning and implementation of more effective solutions for the treatment and prevention of corrosion in the affected areas. An example is the choice of the best inhibitor material, that can provide greater protection to the metal (Barr et al. 2001).

In this work is described a methodology for the detection of corrosive substances and corrosion types, using machine learning (ML) techniques and features extracted from ECN. This work is continuation of (Alves et al. 2017) and (Alves et al. 2017), but with a better analysis. Machine learning techniques have been successfully utilized in various process control, monitoring and optimization applications in different industries. However, the authors have identified that few studies in the field of corrosion types detection used ML techniques, being this one of the motivations of this work.

✉ Patrick Marques Ciarelli
patrick.ciarelli@ufes.br

Lorraine Marques Alves
lorraine_ma@hotmail.com

¹ Federal University of Espírito Santo, Av. Fernando Ferrari, 514, Vitória-ES, Brazil

In this paper, we compare two techniques for feature extraction of the ECN signals, one based on the wavelet transform and another based on recurrence quantification analysis (RQA), and we compare with features used in other works. An important contribution of this work is the identification of types of corrosive substances in aqueous solution. To the best of our knowledge, there are not works of this type performed by other research groups using ML techniques. The results obtained in the experiments indicate that the presented approach is promising to identify some types of localized corrosion and corrosive substances and, in many situations, were achieved accuracies above 90%.

The remainder of this paper is organized as follows. In Sect. 2, we present some classical tools in data analysis of electrochemical noise. In Sects. 3 and 4 are described wavelet transform and RQA, respectively. In Sect. 5, we describe briefly basic concepts of machine learning. Subsequently, in Sect. 6 is shown the materials and methodology used to collect the data for the experiments. In Sect. 7, we present the experiments and achieved results. Our conclusions are presented in Sect. 8.

2 Electrochemical Noise Data Analysis

Corrosion is a spontaneous process characterized by chemical or electrochemical reactions (oxidation and reduction) occurring simultaneously on the surface of solid materials and involving the transfer of electrons from one substance to another (Gentil 2003). Localized corrosion is the degradation of specific points on the metal surface caused by adverse environmental conditions. Common forms of localized corrosion include pitting and crevice on the surface of the metal, or the appearance of cracks (watermark) along the grain boundaries of the structure of the material (intergranular corrosion). Under certain conditions, a passive state, called passivation, may occur during the corrosion, in which a protective film is formed and the material becomes resistant to corrosion (Gentil 2003). The fluctuations of current and potential over time between two electrodes immersed in corrosive solution, caused by oxidation and reduction reactions, are called “current noise” and “potential noise”, respectively. These signals, called electrochemical noise (ECN), are defined as the spontaneous and random fluctuations of the voltage or current from corrosive reactions.

ECN analysis can be performed so that the potential and current noise data are processed independently, using statistical measures, such as mean, standard deviation, kurtosis, and skewness for the data interpretation. The relationship between the two signals can also be analyzed using the concept of electrochemical noise resistance (R_n), defined as the standard deviation of the potential σ_E divided by the standard deviation of the current σ_I , according to Eq. 1:

$$R_n = \frac{\sigma_E}{\sigma_I}. \quad (1)$$

The R_n value is associated with the corrosion rate. The higher the resistance value, the smaller the corrosion rate of the metal. The standard deviation value of the current reflects the fluctuation magnitude of the current in the process, and it can be used to estimate the corrosion activity (Cottis 2001).

A methodology for electrochemical noise analysis, called shot noise, considers that the current has the form of a series of statistically independent charge packets, and each packet has a short duration of time. The total charge passing in a certain time interval is then a sample from a binominal distribution, and if the average number of pulses is fairly large, it approximates from a normal distribution with known properties. Applying this theory to electrochemical noise signals, three parameters can be obtained: average current of corrosion (I_{corr}), average electric charge on each event (q), and frequency of related events (f_n). These parameters are related by Eq. 2 (Cottis and Turgoose 1999; Cottis 2001).

$$I_{\text{corr}} = qf_n. \quad (2)$$

These values cannot be measured directly, but they can be estimated from the potential and current noise data, according to Eqs. 3 and 4 (Cottis 2001):

$$f_n = \frac{I_{\text{corr}}}{q} = \frac{B^2}{\psi_E}, \quad (3)$$

$$q = \frac{\sqrt{\psi_E \psi_I}}{B}, \quad (4)$$

where ψ_E and ψ_I are the low frequency values of power spectral density of the potential and current noise, respectively. B is the Stern–Geary constant which can be estimated by Tafel’s extrapolation (Cottis 2001), where β_a and β_c are anodic and cathodic inclinations, respectively:

$$B = \frac{\beta_a \beta_c}{2.303(\beta_a + \beta_c)}. \quad (5)$$

With the shot-noise methodology, the electric charge q involved in each case can be estimated, as well as the frequency of occurrence f_n of these events. These two parameters provide information about the nature of the corrosion process. Thus, q gives an indication of the mass of metal lost in the event, while f_n provides information about the rate at which these events occur. Therefore, a system that suffers uniform corrosion can have both the charge and frequency elevated. For localized corrosion systems, a low frequency and a high charge are expected. In the case of passivation, the charge is low and the frequency depends on the process that is occurring in the passive film (Amaya et al. 2005).

A problem of this approach is the Stern–Geary constant B , whose value can be estimated by Tafel constant (Eq. 5). Several experimental disadvantages can be associated with Tafel plots. For example, relatively high potentials used in Tafel extrapolation can cause changes in the metal surface, disabling the electrode, requiring the use of two metal specimens to obtain the complete Tafel plot (Research 1980). In some cases, such as corrosion of steel in concrete, the value of B is not constant and can be estimated by linear polarization resistance (LPR), without Tafel plots (Poursaei 2010). Nevertheless, this technique requires the use of a specific instrumentation and does not provide enough information to detect and distinguish different types of localized corrosion (Cox 2014).

Other more recent techniques of ECN analysis include the use of mathematical techniques as Fourier transform, wavelet transform and concepts of chaos theory (Planinsic and Petek 2008). The stochastic nature of ECN imposes some limitations on the use of Fourier Transform, since it does not take into account the variation of the frequency content over time. Wavelet transform overcomes this limitation of the Fourier transform, since it enables the decomposition of the signal into different frequency components for different time intervals (Cottis et al. 2015).

3 Wavelet Transform Analysis

In conventional Fourier analysis, it is not possible to localize the time range that certain frequency band of a signal occurred, since this information is lost during the transform. A way to overcome this problem is to use the wavelet transform. Wavelet can distinguish the local characteristics of a signal on different scales and, by translations, they cover all the region in which the signal is studied. This locality property of wavelets is an advantage over the Fourier Transform in the analysis of nonstationary signals, being a more efficient tool, and applicable to the study of electrochemical noise signals (Aballe et al. 1999; Cottis et al. 2015).

The discrete wavelet transform (DWT) is commonly used for the analysis of discrete signals in order to obtain the coefficients values of different frequency bands for each time interval. These values are obtained by convolution of the sampled signal by functions that are displaced and dilated versions of a wavelet function (or mother wavelet). Thus, the original signal can be written as a sum of wavelet functions ($\phi_{J,n}(t)$ and $\psi_{J,n}(t)$) weighted by their corresponding coefficients, called detail ($d_{J,n}$) and smooth coefficients ($s_{J,n}$). These coefficients indicate the correlation between the wavelet function and the corresponding signal segment (Aballe et al. 1999; Mallat 2009), as showed by Eqs. 6, 7 and 8:

$$x(t) \approx \sum_{n=1}^N s_{J,n} * \phi_{J,n}(t) + \sum_{n=1}^N d_{J,n} * \psi_{J,n}(t) + \quad (6)$$

$$\sum_{n=1}^N d_{J-1,n} * \psi_{J-1,n}(t) + \cdots + \sum_{n=1}^N d_{1,n} * \psi_{1,n}(t),$$

$$s_{J,n} = \int x(t) \phi_{J,n}^*(t) dt, \quad (7)$$

$$d_{J,n} = \int x(t) \psi_{J,n}^*(t) dt, \quad (8)$$

where N is the length of the discrete signal and J represents the decomposition level of DWT.

The coefficients generated by DWT can be difficult to interpret for some ECN signals. A useful way to use the results of the wavelet transform in the analysis of electrochemical noise is through the concept of coefficient energy distribution. In this case, the contribution in energy of each decomposition level is calculated regarding the total energy of the signal, that can be calculated by Aballe et al. (1999), Mallat (2009):

$$E = \sum_{n=1}^N x_n^2, \quad (9)$$

where E is the total energy of the signal, x_n is the signal value in the instants $n = 1, 2, 3, \dots, N$ and N is the length of the discrete signal.

From the total energy E , the fraction of energy of each detail coefficient (E_j^d) and of smooth coefficient (E_j^s) can be calculated, respectively, according to Eqs. 10 and 11, where J are the levels used in the decomposition of the signal through DWT.

$$E_j^d = 1/E \sum_{n=1}^{N/2^j} d_{j,n}^2. \quad (10)$$

$$E_j^s = 1/E \sum_{n=1}^{N/2^j} s_{j,n}^2. \quad (11)$$

Another recently developed ECN analysis tool is based on the concept of Shannon entropy associated with wavelet transform (Moshrefi et al. 2014). While the transform coefficients indicate the transient behavior of the signal, the concept of entropy is used to measure this degree of variability. Thus, the concept of entropy based on wavelet analysis reveals the degree of disorder of ECN signals, which varies according to the conditions of the corrosion process. The entropy of a discrete random variable x with probability $p(x_i)$ can be defined by Mallat (2009):

$$H(x) = - \sum_{i=1}^n p(x_i) \log(p(x_i)), \quad (12)$$

where $p(x_i)$ is estimated using a kernel density.

Like energy, the entropy of the decomposition levels of the transform wavelet provides information for analysis of the ECN signals that cannot be obtained through temporal analysis, making it a powerful tool for corrosion detection and diagnosis (Moshrefi et al. 2014).

4 Recurrence Quantification Analysis

RQA is a developed approach for the analysis of dynamic systems and is based on the recurrence plots (RP) study. Recurrence matrix is the starting point for the discussion of the RQA theory. The formal concept of recurrence was introduced by Henri Poincaré in 1890 and, in a simplistic way, it states that an initial state or configuration of a mechanical system, subjected to conservative forces, will reoccur again in the course of the time evolution of the system (Bergelson 2000). The recurrence plot (RP) method was developed for the visualization of dynamic's trajectories in the phase space of dynamic systems. A recurrence plot is a graphical representation of a $N \times N$ matrix, whose elements are given by Eq. 13:

$$RM_{i,j} = H(\varepsilon - \|x_i - x_j\|), i, j = 1, 2, \dots, N, \quad (13)$$

where N is the number of states in phase space, ε is a predefined threshold radius, x_i and x_j are the points in phase space occurring at time i and j , $\|\cdot\|$ denotes the Euclidean norm of the vectors, and H represents the Heaviside function. If the distance between x_i and x_j falls within the threshold radius, then $RM_{i,j} = 1$, otherwise, $RM_{i,j} = 0$ (Hou et al. 2016). In this paper, the matrix will be obtained on the time series of electrochemical noise data, similarly at Hou et al. (2016).

The threshold value ε must be chosen correctly, since this value influences directly in the recurrence analysis. If ε is too large, almost all points will be identified as a recurrence point. On the other hand, if ε is too small, there may be too few recurrence points impairing the disclosure of recurrence structure (Marwan et al. 2007). In this paper, the ε value was fixed as 20% of the standard deviation of the data segment, like used in Hou et al. (2016).

Variables derived from the recurrence matrices, such as the recurrence rate (R), determinism (D), entropy (E) and average diagonal line length (L) are used to represent quantitatively recurrence plot (Marwan et al. 2007).

Given a $N \times N$ recurrence matrix $RM_{i,j}(\varepsilon)$, $i, j = 1, 2, \dots, N$, then recurrence rate R (Eq. 14) is the measure concerning the density of recurrence points and corresponds to the correlation definition for cases where the number of

points is very large.

$$R = \frac{1}{N^2} \sum_{i,j=1}^N RM_{i,j}(\varepsilon). \quad (14)$$

Determinism D is a measure of system predictability. According to Eq. 15, $P(l)$ is the number of diagonals with length l in RP, and l_{\min} is the smallest size for a row to be considered a diagonal (usually $l_{\min} = 2$). In other words, the value of D is the reason between the number of points belonging to diagonals and the number of recurrence points.

$$D = \frac{\sum_{l=l_{\min}}^N lP(l)}{\sum_{i,j=1}^N RM_{i,j}(\varepsilon)}. \quad (15)$$

The maximum length of the diagonal L_{\max} is the largest diagonal of the RP, excluding the main diagonal line. Another important measure is divergence (DIV), defined as the inverse of L_{\max} .

$$L_{\max} = \max(l). \quad (16)$$

The average length L of the diagonal lines is the number of points belonging to diagonals divided by the number of diagonals in RP, and it can be computed from Eq. 17.

$$L = \frac{\sum_{l=l_{\min}}^N lP(l)}{\sum_{l=l_{\min}}^N P(l)}. \quad (17)$$

The value of v_{\max} is the maximum vertical (horizontal) line length, and it represents the maximum time of “imprisonment” in a given state.

$$v_{\max} = \max(v). \quad (18)$$

Laminarity L_{am} is the analogue of determinism D for vertical (horizontal) lines, and it represents the intensity of occurrence of laminar states, that is, states that do not vary or vary very slowly. Similarly to D , the equation can be understood as the ratio between the number of points belonging to the vertical lines and the total number of recurrent points.

$$L_{\text{am}} = \frac{\sum_{v=v_{\min}}^N vP(v)}{\sum_{v=1}^N vP(v)} \quad (19)$$

Trend T is a measure of non-stationarity in the process, especially if a drift of the parameters is present.

$$T = \frac{\sum_{k=1}^{\tilde{N}} (k - \tilde{N}/2)(RR_k - \langle RR_k \rangle)}{\sum_{k=1}^{\tilde{N}} (k - \tilde{N}/2)^2}, \quad (20)$$

where $RR_k = \frac{1}{N-k} \sum_{j=1+k}^N R_{i,j}$ is the density of recurrence points in the diagonal region distant of k . \tilde{N} is the number of points belonging to RR_k .

Imprisonment Time TT estimates the average time interval in which the system remained “imprisoned” in a given state, without varying it considerably.

$$TT = \frac{\sum_{v=v_{\min}}^N vP(v)}{\sum_{v=v_{\min}}^N P(v)} \quad (21)$$

Finally, E (Eq. 22) measures the Shannon entropy of the probability $p(l) = P(l)/N_l$ to find a diagonal line with length l and it reflects the complexity of the recurrence matrix with respect to diagonal lines.

$$E = - \sum_{l=l_{\min}}^N p(l) \ln p(l). \quad (22)$$

In previous studies, authors suggest that corrosive events can be distinguished by the values of these features. For example, uniform corrosion can be associated with high values of R and low values of D , whereas localized corrosion is associated with low value of R and high value of D (Montalban et al. 2007; Garcia-Ochoa and Corvo 2015).

5 Machine Learning

The work developed in Jian et al. (2013) is one of the few studies known by the authors that used machine learning techniques to identify types of corrosion through ECN. In that study, features extracted from ECN signals were used to train an ANN type multilayer perceptron (MLP) and support vector machine (SVM) to perform automatic classification of types of corrosion that occur on the surface of stainless steel: pitting and general corrosion, as well as the identification of passivation (Jian et al. 2013). In Hou et al. (2016) was also a neural network to identify types of corrosion through ECN. But, in that case, the authors used a neural network trained with features extracted using RQA to identify pitting, general corrosion and passivation.

In this work, five classifiers will be employed, including MLP and SVM, to identify three types of localized corrosion (pitting, crevice corrosion and watermark), as well as the occurrence of passivation on carbon steel AISI/SAE 1040. Furthermore, the same classifiers will be used to identify corrosive reagents. The classifiers used in this paper are: two types of neural networks (MLP Haykin 1998 and probabilistic neural network (PNN) Duda et al. 2000), k nearest neighbor (k NN) (Duda et al. 2000), decision tree (Quinlan 1988) and SVM (Theodoridis and Koutroumbas 2008).

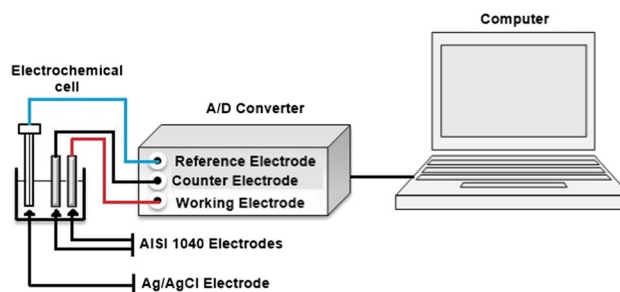


Fig. 1 Experimental apparatus configuration

6 Materials and Method to Collect the Data

Corrosion analysis, through signal processing, consists in the mounting of an experimental apparatus, called electrochemical cell, and an analog/digital (A/D) converter is used for the measurements of electrochemical noise. In this work, potential signals were measured and stored and, indirectly, current signals. Electrochemical cell is an experimental apparatus consisting of an inert metal immersed in an aqueous solution containing ions in different oxidation states.

The cell used in this study to identify types of corrosion consists of two steel electrodes AISI 1040 used as working electrodes and counter electrodes. These electrodes are nominally identical and coated with thermo-contractile, and they have exposed area to solution equal to 499, 512 mm².

The cell used to identify corrosive solutions consists of two steel electrodes AISI 1020 used as working electrodes and counter electrodes. These electrodes are nominally identical and coated with termocontract, and they have exposed area to solution equal to 18 mm².

The reference electrode used in both cases to collect data was silver/silver chloride (Ag/AgCl). The electrochemical cell is connected to a computer through the A/D converter interface, so that the ECN signals data can be stored and processed. Figure 1 shows a diagram with the instruments used for data collection.

In order to collect the data to identify types of corrosion, steel electrodes were immersed for 24 h in passivation solution Na₃PO₄ with concentration of 0.02 M (mol/L), and sampling frequency of 1 Hz. After that time, the NaCl solution was added with a concentration of 0.34 M to start the experiment in aggressive solution for a period of approximately 120 h and sampling frequency of 1 Hz. The data collected in this data set are the signals of current and potential.

To collect the data to identify corrosive solutions, we have chosen substances which are common in industrial environments, as indicated in Table 1.

The acquisition of the signals was obtained using a potentiostat. This instrument is equipped with three connections: working electrode, counter electrode and reference electrode.

Table 1 Used reagents and their concentrations (values in mol/L) in aqueous solutions

| Substance | Concentration | Application |
|--------------------------------|------------------|------------------------------|
| KCl | 0.2, 0.4 and 0.6 | Fertilizer production |
| NaOH | 0.1, 0.2 and 0.3 | Boilers |
| KOH | 0.1, 0.2 and 0.3 | Petrochemical industry |
| NaCl | 0.2, 0.6 and 1.0 | Cooling water |
| FeCl ₃ | 0.1, 0.2 and 0.3 | Coagulant in water treatment |
| H ₂ SO ₄ | 0.2, 0.3 and 0.4 | Fertilizer production |

To analyze the ECN measurements, for each reagent was performed three different concentrations, totalizing 18 measurements with 60 min each. The sampling frequency used was 4 Hz, such that each measurement has 14,400 points.

All measurements were obtained at room temperature. The formation and growth of passive film and the occurrence of corrosion were analyzed using the potential and current of the noise according to immersion time. The electrochemical noise signals of the current and potential were recorded by an A/D converter.

7 Results and Discussion

The experiments were divided into two blocks. In the first one, we extract features from ECN signals to identify the types of corrosion, and in the second block, we apply a method to identify the types of corrosive substances. In both experiments, we used machine learning techniques for classification and the accuracy metric was used as quality measure. The value of the accuracy is calculated by the ratio of the number of samples correctly classified by the total number of samples, multiplied by 100%. The higher the value of accuracy, the better. The best value is 100%. The datasets¹ used in this section are the same described in Sect. 6.

7.1 Feature Selection for Corrosion Type Identification

The first step performed for identification of corrosion type was feature selection. For the selection of the most significant features, we used sequential backward feature selection (SBS) algorithm, which is a search algorithm that starts with a complete set of features and for each iteration removes the feature with the least impact on the established criterion function (in this paper we used accuracy). Multidimensional features have been analyzed as a group, not individually.

The SBS algorithm was applied along with a MLP with 20 neurons in the hidden layer, learning rate equal to 0.001 and 1000 training epochs using Levenberg–Marquardt algorithm (Marquardt 1963) for training. A training set of 132 samples and a test set with 68 samples were used for this experiment, totalizing 200 examples selected of different parts of the sampled signal, while ensuring there is the same number of samples for all four classes. The features were normalized by the mean and standard deviation obtained from the training set, so that the distribution of each feature has zero mean and standard deviation equal to one.

In the first moment, the following features (and respective dimensions) were evaluated: detail and smooth coefficients (10 elements each), energy (8 elements), entropy (9 elements), ratio between standard deviation and mean (1 element), kurtosis (1 element), ratio between the derivative and the mean (1 element), and resistance to electrochemical noise (1 element). Each characteristic was extracted for both voltage and current signals, except resistance to electrochemical noise. Thus, the full feature vector to be analyzed by SBS has 81 elements.

The wavelet transform of Daubechies (db4) with decomposition at 8 levels was used to compute energy and entropy, as described in Sect. 3, and detail and smooth coefficients. The features were obtained from non-overlapping data packages composed of 1024 points of ECN signals of potential and current.

The attributes selected by the SBS were: entropy, energy, and resistance to electrochemical noise, resulting in a feature vector of 35 elements (8 entropy and 9 energy for each voltage and current signal, and one resistance of the electrochemical noise).

In Fig. 2 is shown in 2D graphics how good the selected features are to separate the samples of the different types of corrosion. Figure 2a shows the separation using all evaluated features and Fig. 2b using only the selected features. As we can see, the separation using the selected features is more evident than using all features. A dimensionality reduction technique was necessary to visualize this graphics, whose mapping hold neighborhood relations among samples. For this purpose, we used *t*-distributed stochastic neighbor embedding technique (*t*-SNE), which produces one of the best mappings in terms of preserving neighborhoods (Fadel et al. 2015).

Similarly, using the features extracted from RQA described in Sect. 4, the attributes selected by SBS were: recurrence rate (*R*) and determinism (*D*). In Fig. 3 are shown in 2D graphics the efficiency of the features of RQA for the separation of the samples of the different types of corrosion. In Fig. 3a is shown the graph for visualization of all attributes of RQA described in Sect. 4 and in Fig. 3b only the attributes selected by SBS. As can be observed, the separation between the different classes is not as clear as in the case of the wavelet.

¹ Available on <https://github.com/lorrainemarques/Electrochemical-Noise-Data>

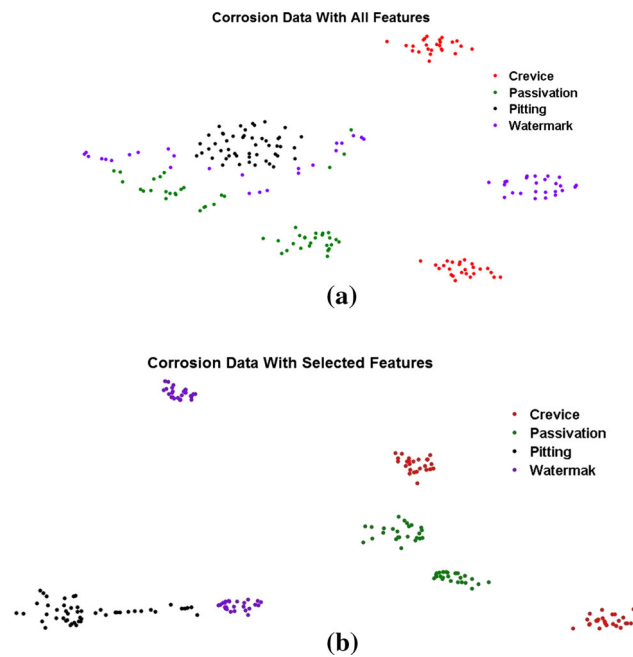


Fig. 2 Visualization of features derived from wavelet in the task of corrosion type identification using t -SNE. Visualizing of corrosion data with **a** all features and **b** selected features

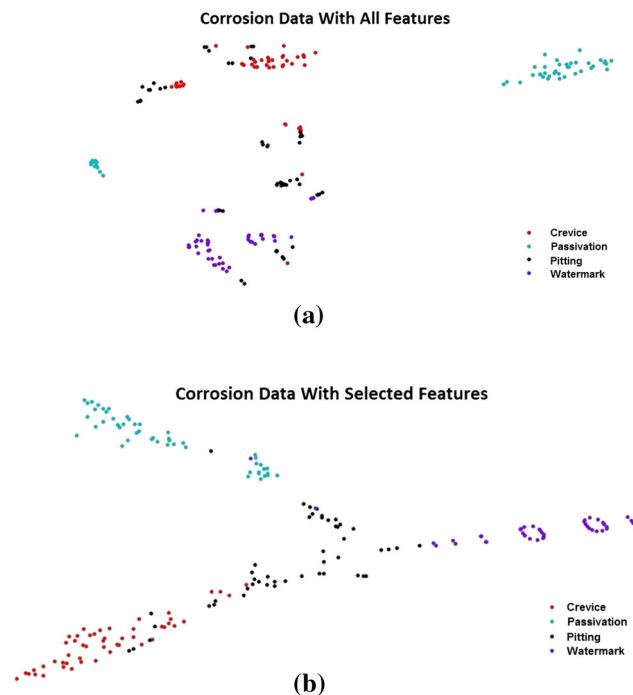


Fig. 3 Visualization of features derived from RQA in the task of corrosion type identification using t -SNE. Visualizing of corrosion data with **a** all features and **b** selected features

Table 2 Classification of corrosion types using different feature set (accuracy values are in percent)

| Features | MLP | PNN | k NN | Tree | SVM |
|--------------------|--------------|--------------|--------------|--------------|--------------|
| Wavelet | 96.97 | 96.88 | 95.83 | 91.67 | 97.42 |
| RQA | 96.04 | 94.01 | 91.67 | 86.98 | 92.18 |
| (Jian et al. 2013) | 83.59 | 84.11 | 75.52 | 70.31 | 70.31 |
| (Hou et al. 2016) | 95.37 | 95.57 | 91.67 | 86.98 | 93.75 |

The results in bold are the best results for each classifier

7.2 Corrosion Type Determination

To verify if the selected features have a good ability to distinguish the different types of corrosion, the next step is the training of the classifiers presented in Sect. 5.

Initially, 48 samples for each one of the four types of corrosion were obtained, each one composed of the 35 features previously selected in Sect. 7.3. The features were extracted from non-overlapping data packages of 1024 points each, from potential and current signals. These packages were selected from different parts of the sampled signal (which were different to those used in Sect. 7.1).

After this step, the samples were stratified into threefold of data, each one with 64 samples. Then, for each classifier were obtained 3 results from 3 tests, and each result was achieved using twofold for training/validation and onefold for testing. For each result was used a different fold for testing. The features were normalized by the mean and standard deviation obtained from the training set, as described in Sect. 7.3.

The following configurations were tested for each technique in order to maximize the accuracy on the training set and, possibly, also on the test set. The MLP was trained using Levenberg–Marquardt algorithm, learning rate of 0.001, 1000 epochs and evaluated different numbers of neurons in the hidden layer (1–50 neurons). Different values of standard deviation were tested for PNN (0.1 to 1.0, with steps of 0.1). For k NN, we employed Euclidean distance and we varied the value of k (1 to 10). The SVM used in the experiments had linear kernel, and different values of cost c were evaluated (0 to 10, with steps of 0.5). For the decision tree was used the standard MATLAB implementation, which does not have parameters to tune.

For each test, the parameters of each technique were adjusted in order to maximize the accuracy on the training folds, and these parameters were used to classify the samples of the test fold. Table 2 shows the values of average accuracies obtained by each technique when used the features extracted by wavelet and RQA. In addition, the features used in Jian et al. (2013) and Hou et al. (2016) were also evaluated. The best result for each classifier was highlighted.

Comparing the results of the four feature sets, there is a slight advantage of the features obtained by wavelet in the

task of identifying the type of corrosion. As can be seen, all algorithms achieved an average accuracy above 95% when using features extracted by wavelet, with the exception of the decision tree. The SVM achieved the best mean accuracy, but, in general, MLP and PNN achieved more stable results. PNN has a advantage over SVM and MLP because it is a more simple and flexible technique, and it is useful for cases where simplicity is more valuable than performance.

Decision tree did not achieve performance similar to the other techniques. Some reasons may have been the construction of a tree with little capability of generalization and the fact that the process of feature normalization, through mean and standard deviation, have no effect on the decision trees. However, the high accuracy values in other classifiers indicate that methodology using wavelet was effective to identify the types of corrosion analyzed.

As we can observe in Table 2, the results obtained with RQA were slightly inferior in comparison with the results obtained with wavelet. However, the attributes selected by SBS reduce significantly the feature vector length, when compared with the number of features used with wavelet approach. Therefore, in applications whose computational cost is a determining factor, the system could be analyzed using the attributes selected from RQA.

The results of these two approaches were also compared with the features used in Jian et al. (2013) and Hou et al. (2016). These were some of the few studies known by the authors that uses features extracted from ECN signals to train a machine learning technique. The features used in Jian et al. (2013) were the electrochemical noise resistance (R_n), the frequency of events (f_n), the charge (q) and the energy of the detail coefficients extracted by wavelet transform. The features used in Hou et al. (2016) were derived from RQA: recurrence rate, determinism, entropy, and the average length of the diagonal lines.

The main difference between the approach using wavelet, presented in this paper, and the features used in Jian et al. (2013) is the use of entropy information. In Jian et al. (2013) is used the Stern–Geary constant, but it can be a great source of error if this value is not precisely estimated (Ahmad et al. 2014). Moreover, this value is not constant in some cases (Poursaei 2010). On the other hand, the main difference between the approach using RQA, presented in this paper, and the features used in Hou et al. (2016) is that in the former are used only recurrence rate and determinism.

The results shown in Table 2 indicate that the features used in Jian et al. (2013) are not as good as the other evaluated approaches. Nevertheless, the results obtained using the features employed in Hou et al. (2016) are very similar to that obtained by our approach using RQA. Hence, it is possible that the features entropy and average length of the diagonal lines, both presents only in Hou et al. (2016), do not contribute significantly in the results.

Table 3 Classification of corrosion types using different features (accuracy values are in percent)

| Signal | MLP | PNN | kNN | Tree | SVM |
|-----------|--------------|--------------|--------------|--------------|--------------|
| Potential | 97.66 | 90.36 | 97.92 | 94.79 | 94.72 |
| Current | 87.76 | 95.38 | 85.42 | 93.23 | 87.50 |

The results in bold are the best results for each classifier

The same experiments were performed using separately the selected wavelet features extracted from potential and current signals. The average accuracies are shown in Table 3. As can be observed, the signal of potential was, in general, more efficient to perform the task. In addition, the use of potential was as effective as both types of signals to identify the type of corrosion. This indicates that it is possible to use a simpler data collection system while maintaining similar classification performance.

7.3 Sample Size and Selection of Features for Identification of Corrosive Substances

The features used to identify the type of substance are extracted of several data segments, however, there is a limitation on the amount of data collected. Therefore, it is important to define its size, because few points per segment supply little information, but, if is used many points per segment, it will be impossible to evaluate properly the proposed method. Thus, we evaluated non-overlapping data segments in the range of 144–2880 points, with increments of 144 points. For this experiment was used a total of 14,400 points equally distributed in all six classes, where 70% of the points were used to train and the others 30% to test.

Figure 4 shows the accuracy values obtained for different numbers of points, when using a MLP with 20 neurons in the hidden layer, learning rate equal to 0.001 and 1000 training epochs, using the algorithm of Levenberg–Marquardt for training and the features of entropy, energy, detail and smooth coefficients, as described in Sect. 3, resulting in a feature vector with 27 elements. After computing the features, they were normalized by the mean and standard deviation. The first step of wavelet analysis method is to remove the mean of each time series and to define the corresponding wavelet family (father and mother) (Aballe et al. 1999). The features used in this experiment were computed from signal of potential with wavelet transform of Daubechies (db4) with decomposition at 8 levels.

It is observed in Fig. 4 that the greater the number of points per segment, the better is the accuracy. Ideally, a good choice is a large number of points per segment (sample). However, the higher number of points is, the smaller is the number of samples available for training and testing. A compromise between the two quantities is 960 points per segment, as indi-

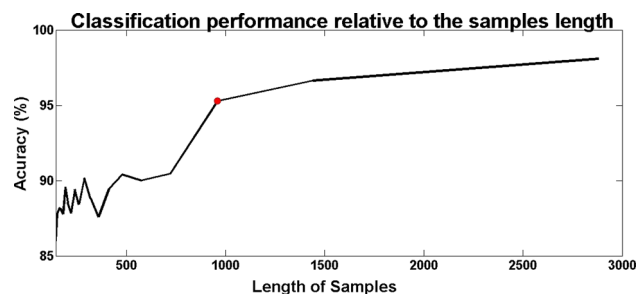


Fig. 4 Accuracy \times sample size

cated in Fig. 4. Therefore, we can have a reasonable number of samples per class.

Once the size of the segment is defined, the next step is the selection of the most significant attributes. In this step the SBS algorithm was used, using the same configuration described in Sect. 7.1. For the selection task, a training set with 180 examples and a test set with 90 examples were used, totaling 270 examples, but ensuring that all six classes had the same amount of samples. The attributes were standardized by the mean and standard deviation obtained from the training set, so that the distribution of each has mean zero and standard deviation one.

Figure 5 shows the samples before and after the SBS application. Note that prior to the feature selection phase the classes of acids KCl, NaCl, and H₂SO₄ overlap at some points in the graph, which may indicate classification errors. Observing the resulting graph after the selection of features, it is noticed that the classes are more delimited, showing that the attributes selected by SBS are more discriminating.

Similarly, using the features extracted from RQA, described in Sect. 4, the attributes selected by SBS were: recurrence rate (R), determinism (D), entropy (E), laminarity (L), v_{\max} , and trap time (TT). In Fig. 6 is shown in 2D plots the efficiency of the selected features for the separation of the samples of the different types of corrosive agents. Comparing the two graphs we can observe that there were not many changes in the class separation with the SBS application.

7.4 Identification of Corrosive Substances

To train each classifier using the features extracted from the ECN signals of different reagents, the following procedure was performed. Initially, each one of the 18 ECN signals was divided into 15 examples of 960 points each, and formed threefold of data, each one containing 5 examples of each concentration of each reagent. It is important to observe that each one of the 6 reagents is composed of 45 examples.

For the methods based on wavelet and RQA were extracted the features selected in Sect. 7.3 totaling into 27 and 6 features, respectively. In this experiment, the features used in Hou et al. (2016) were also evaluated. All features were

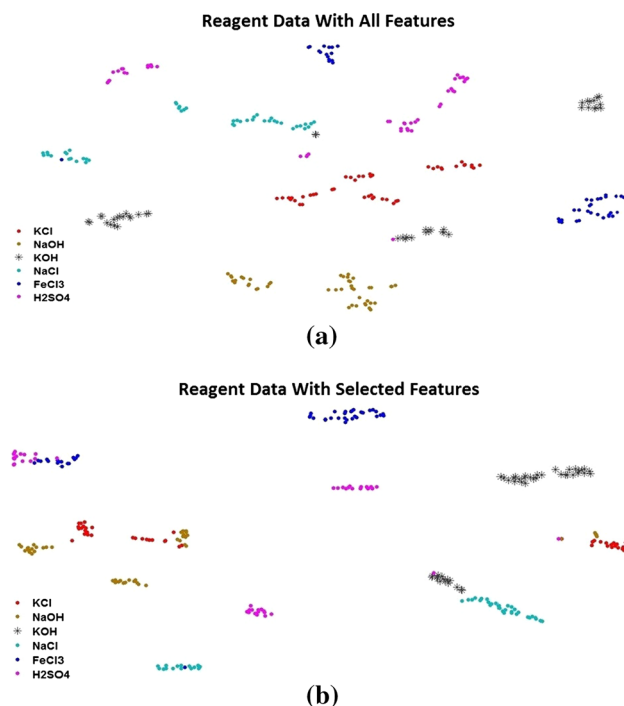


Fig. 5 Visualization of attributes derived of wavelet in the reagent type identification task using t -SNE. Visualizing of reagent data with **a** all features and **b** selected features

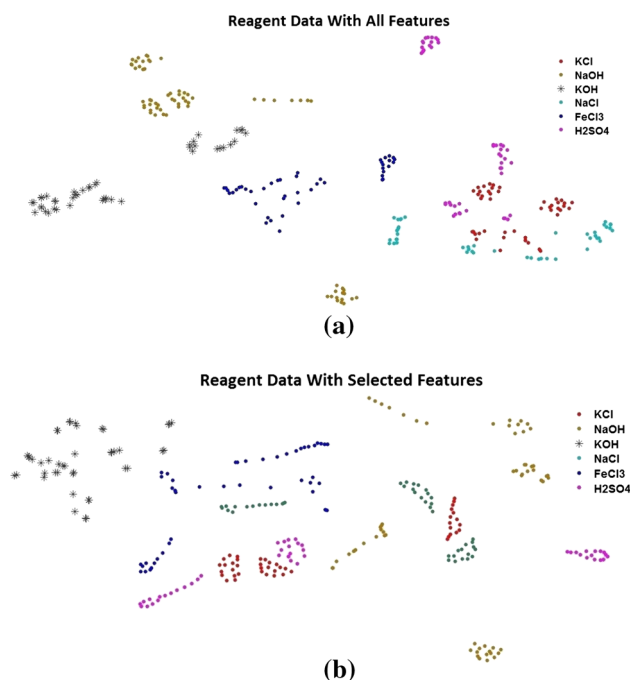


Fig. 6 Visualization of attributes derived of RQA in the reagent type identification task using t -SNE. Visualizing of reagent data with **a** all features and **b** selected features

Table 4 Classification of corrosive substances using different feature set (accuracy values are in percent)

| Features | MLP | PNN | kNN | Tree | SVM |
|-------------------|--------------|--------------|--------------|--------------|--------------|
| Wavelet | 96.29 | 71.97 | 90.48 | 85.40 | 91.12 |
| RQA | 91.11 | 90.37 | 93.70 | 92.59 | 90.37 |
| (Hou et al. 2016) | 87.41 | 90.00 | 91.11 | 92.59 | 84.44 |

The results in bold are the best results for each classifier

standardized by mean and standard deviation. Then, 3 tests were performed for each classifier, and for each test was used twofold for training and onefold for testing. For this dataset only ECN signals of potential were collected and the electrochemical tests of the Tafel curves were not performed, so that it was not possible to use the features used in Jian et al. (2013) from the shot-noise theory.

Table 4 shows the average accuracy of each classifier for each feature set, and the best result for each classifier is highlighted. In this experiment, the results obtained with the features extracted from RQA are more consistent (have a smaller variance) than the results obtained using wavelet, so that all results were above 90%. Nevertheless, the best result was obtained when combining wavelet with MLP, although kNN has presented more consistent results. The variability of results obtained by each classifier indicates that some classifiers are more appropriate for certain types of features than others. We can also see in Table 4 that there was a slight improvement in the results when using the selected features of RQA in relation to the features used in Hou et al. (2016).

8 Conclusions

This paper presented an approach to identify some types of corrosion and reagents on metal surface through electrochemical noise signals.

In experimental results was observed that resistance to electrochemical noise and features extracted from wavelet transform, entropy and energy were the most discriminative to identify some types of corrosion and that RQA achieved more stable results to identify some types of reagents, although wavelet transform obtained similar results.

After analyzing the results, we noted that all five evaluated classifiers achieved an average accuracy above 90% to perform the task of identifying types of corrosion, and SVM achieved an average accuracy slightly better than those obtained by the other classifiers. For the task of identifying types of reagents, we noted that MLP achieved an average accuracy above 95%. In both tasks, the feature vector obtained from RQA is smaller than that used for wavelet, and results indicated that potential signal seems to be more effective to identify corrosion types than current signal.

The results of this study highlight the importance of using wavelet transform and RQA for electrochemical noise analysis. In future work, we intend to analyze other machine learning techniques, as deep learning, and other features in order to improve the results using only a type of signal, simplifying the instrumentation necessary to collect the data. Other possibility is to analyze the identification of the mixture of different types of corrosion and corrosive solutions.

Acknowledgements Evandro O. T. Salles would like to thank FAPES - Fundação de Amparo à Pesquisa e Inovação do Espírito Santo by the partial support under grant 244/2016. Patrick Marques Ciarelli thanks the partial funding of his research work provided by CNPq (Grant 312032/2015-3).

References

- Aballe, A., Bethencourt, M., Botana, F., & Marcos, M. (1999). Using wavelets transform in the analysis of electrochemical noise data. *Electrochimica Acta*, 44(26), 4805–4816.
- Ahmad, S., Jibrán, M. A. A., Azad, A. K., & Maslehuddin, M. (2014). A simple and reliable setup for monitoring corrosion rate of steel rebars in concrete. *The Scientific World Journal*, 2014, 10.
- Alves, L.M., Cotta, R.A., & Ciarelli, P.M. (2017). Identification of types of corrosion through electrochemical noise using machine learning techniques. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, (pp. 332–340).
- Alves, L.M., Cotta, R.A., Prado, A.R., & Ciarelli, P.M. (2017). Identification of corrosive substances through electrochemical noise using wavelet and recurrence quantification analysis. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, (pp. 718–723).
- Amaya, J., Cottis, R. A., & Botana, F. (2005). Shot noise and statistical parameters for the estimation of corrosion mechanisms. *Corrosion Science*, 47(12), 3280–3299.
- Barr, E., Pierrerd, L., & Greenfield, A. (2001). Application of electrochemical noise monitoring to inhibitor evaluation and optimization in the field: Results from the kaybob south sour gas field. In *Corrosion*. NACE International. <https://www.onepetro.org/conference-paper/NACE-01288>.
- Bergelson, V. (2000). The multifarious poincaré recurrence theorem. *Descriptive Set Theory and Dynamical Systems*, 1, 31–57.
- Cottis, R. (2001). Interpretation of electrochemical noise data. *Corrosion*, 57(3), 265–285.
- Cottis, R., & Turgoose, S. (1999). *Corrosion testing made easy: Impedance and noise analysis* (1st ed.). Houston: NACE International.
- Cottis, R. A., Homborg, A., & Mol, J. (2015). The relationship between spectral and wavelet techniques for noise analysis. *Electrochimica Acta*, 202, 277–287.
- Cox, W. M. (2014). A strategic approach to corrosion monitoring and corrosion management. *Procedia Engineering*, 86, 567–575.
- Duda, R., Hart, D. G., & Stork, P. (2000). *Pattern recognition* (1st ed.). Hoboken: Wiley Interscience.
- Fadel, S. G., Fatore, M., Duarte, S., & Paulovich, V. (2015). Loch: A neighborhood-based multidimensional projection technique for high-dimensional sparse spaces. *Neurocomputing*, 150, 546–556.
- Gentil, V. (2003). *Corrosão*. 4a ed. Rio de Janeiro: Livros Técnicos e Científicos.

- Garcia-Ochoa, E., & Corvo, F. (2015). Using recurrence plot to study the dynamics of reinforcement steel corrosion. *Protection of Metals and Physical Chemistry of Surfaces*, 51(4), 716–724.
- Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). Upper Saddle River: Prentice Hall.
- Hou, Y., Aldrich, C., Lepkova, K., Suarez, L., & Kinsella, B. (2016). Monitoring of carbon steel corrosion by use of electrochemical noise and recurrence quantification analysis. *Corrosion Science*, 112, 63–72.
- Jian, L., Weikang, K., Jiangbo, S., Ke, W., Weikui, W., Weipu, Z., et al. (2013). Determination of corrosion types from electrochemical noise by artificial neural networks. *International Journal of Electrochemical Science*, 8, 2365–2377.
- Koch, G., Varney, J., Thompson, N., Moghissi, O., Gould, M., & Payer, J. (2016). International measures of prevention, application and economics of corrosion technologies study (impact). Technical Report, NACE International.
- Mallat, S. (2009). *A wavelet tour of signal processing: The sparse way* (3rd ed.). Burlington: Academic Press Elsevier.
- Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2), 431–441.
- Marwan, N., Romano, M., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, 438, 237–329.
- Montalban, L., Henttu, P., & Piche, R. (2007). Recurrence quantification analysis of electrochemical noise data during pit development. *International Journal of Bifurcation and Chaos*, 17, 3725–3728.
- Moshrefi, R., Mahjani, M., & Jafarian, M. (2014). Application of wavelet entropy in analysis of electrochemical noise for corrosion type identification. *Electrochemistry Communications*, 48, 49–51.
- Planinsic, P., & Petek, A. (2008). Characterization of corrosion process by current noise-based fractal and correlation analysis. *Electrochimica Acta*, 53(16), 5206–5214.
- Poursaei, A. (2010). Potentiostatic transient technique, a simple approach to estimate the corrosion current density and Stern–Geary constant of reinforcing steel in concrete. *Cement and Concrete Research*, 40(9), 1451–1458.
- Quinlan, J. (1988). *Decision trees and multivalued attributes* (11th ed.). Oxford: Oxford University Press.
- Research, E.P.A. (1980). Basics of corrosion measurements—application note corr-1. Technical Report, Princeton.
- Theodoridis, S., & Koutroumbas, K. (2008). *Pattern recognition* (4th ed.). Amsterdam: Elsevier.