Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 70 (2014) 688 – 697

Procedia Engineering

www.elsevier.com/locate/procedia

12th International Conference on Computing and Control for the Water Industry, CCWI2013

Feature extraction in pressure signals for leak detection in water networks

M.M. Gamboa-Medina[a]*, L.F. Ribeiro Reis[a], R. Capobianco Guido[b]

[a] *University of São Paulo, Av. Trabalhador Saocarlense 400 São Carlos-SP 13566-590, Brasil*
[b] *Paulista State University, Rua Cristóvão Colombo 2265 São José do Rio Preto-SP 15054-000, Brasil*

Abstract

Techniques based on signal analysis for leak detection in water supply systems typically use long pressure and/or flow data series of variable length. This paper presents the feature extraction from pressure signals and their application to the identification of changes related to the onset of a leak. Example signals were acquired from an experimental laboratory circuit, and features were extracted from temporal domain and from transformed signals. Statistical analysis of features values and a classification method were applied. It was verified the feasibility of using feature vectors for distinguish data acquired in the absence or presence of a leak.

Keywords: Leakage detection; feature extraction; signals analysis; water supply system.

1. Introduction

New methodologies for leakage management in water supply networks have been proposed over the last years (Puust et al. 2010), and several of them have focused on the detection of leaks based on the identification of changes in the state of the hydraulic system caused by bursts or leaks. Signal processing techniques have been used, commonly dealing with long series of the state variables of water supply systems (pressures and flows) collected by sensors installed at strategic points of the respective systems (Romano et al. 2012; Mounce et al. 2010; Ye et al. 2011). However, such series may include errors or missing data, beside their large sizes, so the

* Corresponding author. Tel.: +55-16-33738267.
E-mail address: mmgamboam@usp.br

doi:10.1016/j.proeng.2014.02.075

analysis can be complex and computationally expensive (Beale & Jackson 2010). This paper presents the feature extraction of pressure signals and their application to the identification of changes related to the onset of a leak as an alternative to improve the signal analysis.

Applying the concept of feature extraction, a few values calculated from the original signal can properly represent it for a particular aim, providing the same or even better results in the subsequent analyses. In this study the purpose is to identify whether a pressure signal acquired on the network shows the occurrence of a leak based on experimental data. An experimental circuit equipped with pressure sensors was used for the acquisition of samples of pressure signal in the absence or presence of leaks (NOLEAK or LEAK, respectively).

The following example illustrates what the feature extraction applied here means. In Figure 1 are shown two series of pressure data acquired with 240 sequential pressure values, each belonging to a different class (LEAK and NOLEAK). The difference in the magnitudes of such signals is due to the different flow rates at which they were obtained, while behavior changes are created by the beginning of a leak. Commonly, a visual identification of those behavior changes cannot be made from such series easily. After standardizing the values of the series, features were extracted and formed one vector of seven elements for each example, according to details presented in Section 3. The seven elements for each example are shown in Figure 2, where the feature values are on the y-axis. By comparing the values of the resulting features, it is possible to distinguish the two examples as leak or non-leak, more easily than looking at the behavior of the original series in Figure 1.

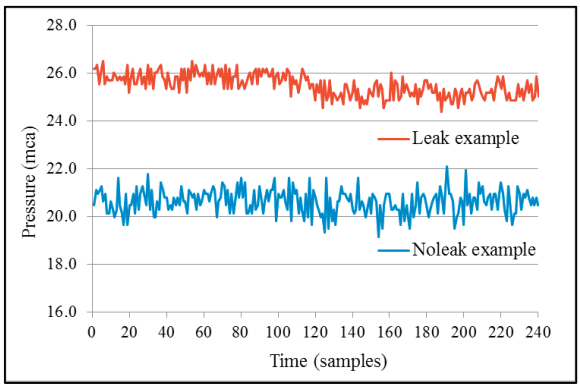

Figure 1. Example of two pressure signals acquired.

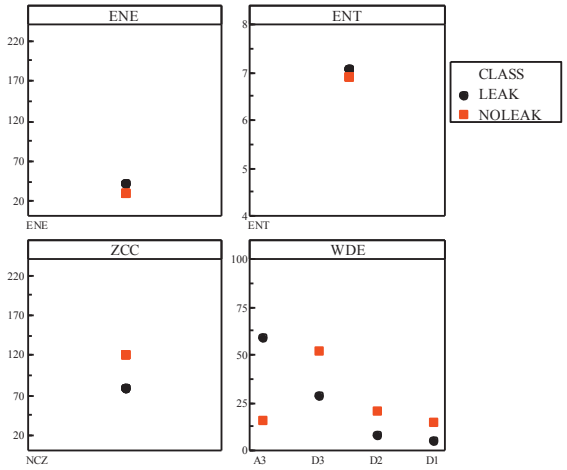

Figure 2. Features extracted from the two pressure signals.

Clearly the research is not made with only two signals, as the previous example, but with all the data acquired from the experimental circuit. In order to identify features that can meet the objectives outlined, four types of common features are analyzed and their ability is assessed so that the onset of a leak in the signal can be identified. Some of the features considered are calculated from the original (temporal) representation of the signal and others from the transformed representation using the wavelet decomposition. The potential advantages of the transformed representation are also considered. Additionally, the feature extraction is coupled with a learning machine methodology by using a binary classifier. The results after training and testing the classifiers are compared for different situations: using isolated features, combined features and the original signal as inputs.

2. Data acquisition

The data used were acquired in an experimental laboratory circuit. The circuit is integrated by PVC pipes of 53, 75 and 101 mm diameters with a total length of approximately 200 m, arranged in a grid format with nine vertical

pipes and five horizontal pipes, forming several internal loops, whose topology can be changed using the valves located in the pipes. The water used is pumped from a constant-level reservoir which also receives the water after it had passed through the circuit. The opening of small valves (hydrometers) in the system allows simulating the onset of a leak. Fifteen pressure sensors are installed at different points in the circuit (Figure 1).

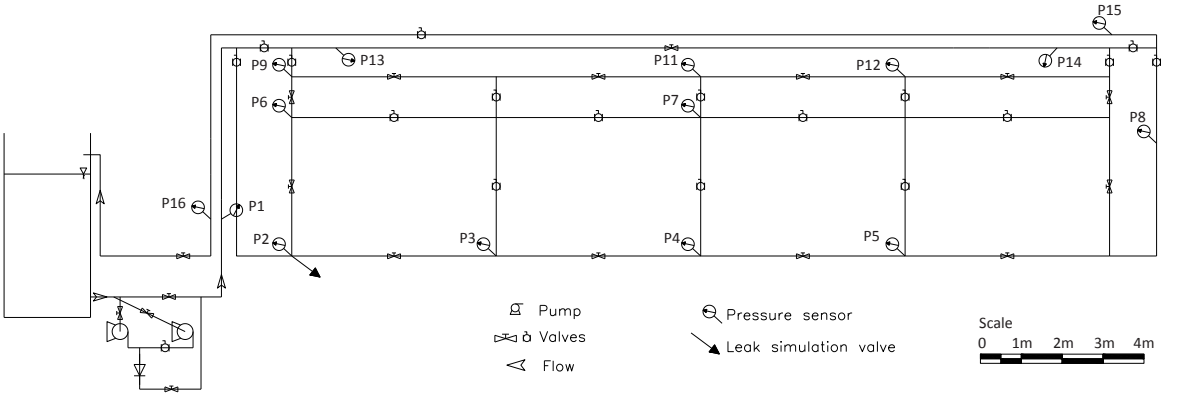

Figure 3. Experimental circuit

Experiments were performed to acquire pressure signals in two different situations: at no leak or any change in the system (NOLEAK) and at the beginning of a leak simulated by the rapid opening of a small valve (LEAK). The sampling rate used was 4 samples/sec and each 1 min segment (240 samples) was considered an example for the analyses. 310 NOLEAK examples were acquired with different flow rates in the circuit (Q) by all sensors simultaneously. 310 LEAK examples with all sensors were also acquired at different flow rates (Q) and leaks (V) in the circuit (the V/Q rates of the examples are shown in Table 1).

Table 1. V/Q rate of LEAK examples

V/Q	< 3%	3% - 4%	4% - 5%	5% - 6%	6% - 7%	7% - 8%	8% - 9%	9% - 10%	10% - 11%	> 11%
Number of examples	10	33	51	39	30	36	29	35	32	15

3. Feature extraction

As the examples were simulated and acquired in the laboratory under controlled conditions, they have the same size (240 points). Under more realistic conditions the signals may present different sample rates, changing durations, missing data or wrong values that should be disregarded, so the signal or input vector commonly has a large and variable size. Additionally, the signals acquired and digitized consist in a series of instantaneous pressure values that may be affected by noise, variations in measurements and transmission, or physical phenomena like microturbulence. The analysis of the vector values of the acquired signal may be impaired by the presence of disturbances, the variability in the size of the signals, and the high computational cost of long series analyses, therefore several methodologies for signal processing use feature extraction. Feature extraction allows analyzing some features derived from the signal rather than a large set of original signal values. Features can be represented in shorter vectors that must maintain the ability to distinguish the class to which a sample belongs.

Several types of features can be extracted from the signal, and the choice of them depends on the specific problem. Since there are no previous works on extracting features from pressure signals for leak detection purpose, a search was performed to evaluate the use of four characteristics commonly employed in digital signal processing:

energy (ENE), entropy (ENT), zero crossings count (ZCC), and distribution of energy in the components of wavelet decomposition (WDE), all of them briefly presented below.

3.1. Energy (ENE)

The energy of a discrete signal is an indirect measure of the distance of the values from their mean, and is defined as (Proakis & Manolakis 2007).

$$\mathrm{ENE}=\sum_{i=1}^{N} x_i^{\,2}$$

Where: $X = \{x_1, x_2 \ldots x_N\}$ = Example signal; ENE = Energy of signal X.

A total energy value is calculated for each example signal (each vector X). For a standardized signal with dimension N and $x_i \in [-1,1]$, the energy values should be ENE < N.

3.2. Entropy (ENT)

Entropy is essentially a measure of unpredictability in a discrete probability distribution, originally presented by Claude Shannon as a fundamental part of the information theory. Signal processing (besides other areas) uses the concept of entropy as a measure of randomness, information content or dispersion of the information contained in a vector, and these last interpretation is the more interest for this research. The entropy of an n-dimensional vector is defined as (XU, Shui, 2013).

$$ENT = -\sum_{i=1}^{N} P_i log_2 P_i \qquad P_i = \frac{x_i^{\,2}}{\sum_{n=1}^{N} x_n^{\,2}}$$

Where: $X = \{x_1, x_2 \ldots x_N\}$ = Example signal; ENT = entropy of signal X, in bits; P = probability function value for each element x_i.

For each example signal (each vector X), one value of ENT is calculated. If values in X have a uniform distribution, the ENT depends only on the vector size; for the actual signals, entropy should be $ENT < \log_2(1/N)$.

3.3. Zero Crossing Count (ZCC)

The zero crossing count (ZCC), as the name implies, is merely a counting of the occurrences in which a positive element of the vector appears immediately after a negative one, or a negative element after a positive.

$$NCZ = \sum_{i=2}^{N} I(x_i \cdot x_{i-1} < 0)$$

Where: $X = \{x_1, x_2 \ldots x_N\}$ = Example signal; ZCC = zero crossing count of signal X; I(s) = 1 if s = True; I(s) = 0 if s = False.

The result for each signal is a discrete number ZCC $\in [1,N-1]$. The noise (added data without mean) commonly appears as a high-frequency oscillation, therefore the more the noise influences signal changes and behavior, the higher the ZCC values.

3.4. Distribution of the energy on the components of the wavelet decomposition (WDE)

The three aforementioned features are calculated from the original representation of the signal in the time domain, more specifically from the set of sequential values measured over time. However in many cases the temporal representation is insufficient and other forms of indexing, as domain transforms, can provide superior results (Meyer 1999). The transformed signal contains information about the frequencies if the Fourier transform has been applied or about the time and frequencies if the wavelet transform has been used.

The capability of the wavelet transform for the analysis of pressure signals has been explored over the last years to detect leaks by analyzing signals obtained during transient phenomena (Ferrante & Brunone 2003; Ferrante et al. 2007; Ferrante et al. 2009). However, this research explores the use of the wavelet transform in signals acquired with lower sampling rate and longer duration than needed for studying transient waves. The discrete wavelet decomposition with function Daubechies 2 (db2) for level 3 was used here.

In the discrete wavelet decomposition, the original signal is represented by two new signals or components, each formed by a vector of elements sorted in time. One of the components corresponds to lower frequencies (approximation or trend) and the other corresponds to high frequencies (detail). The decomposition process is performed from one level to another, dividing the approximation component of the previous level into two new components until the selected level of decomposition has been achieved. In the wavelet decomposing for level 3, the signal is represented by an approximation component (A3) and three detail components (D3, D2, D1), each in a different frequency band. To analyze the results may be searched which components concentrate more or less information, calculating the energy as defined in section 3.1. In the present study, the energy value was calculated from each component and it was expressed as a percentage of the total energy of the four components. The feature "distribution of the energy on the components of the wavelet decomposition" (WDE) corresponds to a vector of four real numbers in the 0 to 100 range.

4. Results

In order to analyse the features considered, the feature extraction process was applied to each of the 620 examples of both classes (LEAK and NOLEAK), and data from the 15 sensors were analysed independently. These features extracted were analysed in two ways: examining the distribution of their values and including the feature values in a pattern recognition system for evaluate its performance.

4.1. Histogram

Frequency histograms were built for the LEAK and NOLEAK examples for each value of the feature vector, i.e. four values for WDE and one value for each ENE, ENT and ZCC. The histograms resulting from data of two sensors are shown in Figure 4, considering the nearest sensor to the leak node (P02) and the farthest sensor downstream the leak node (P16) (see Figure 1). For each histogram an interval of feature values has been defined, such that 85% of the feature values (extracted from the examples signals) are within the interval, other 7.5% are lower and 7.5% are higher (Table 2).

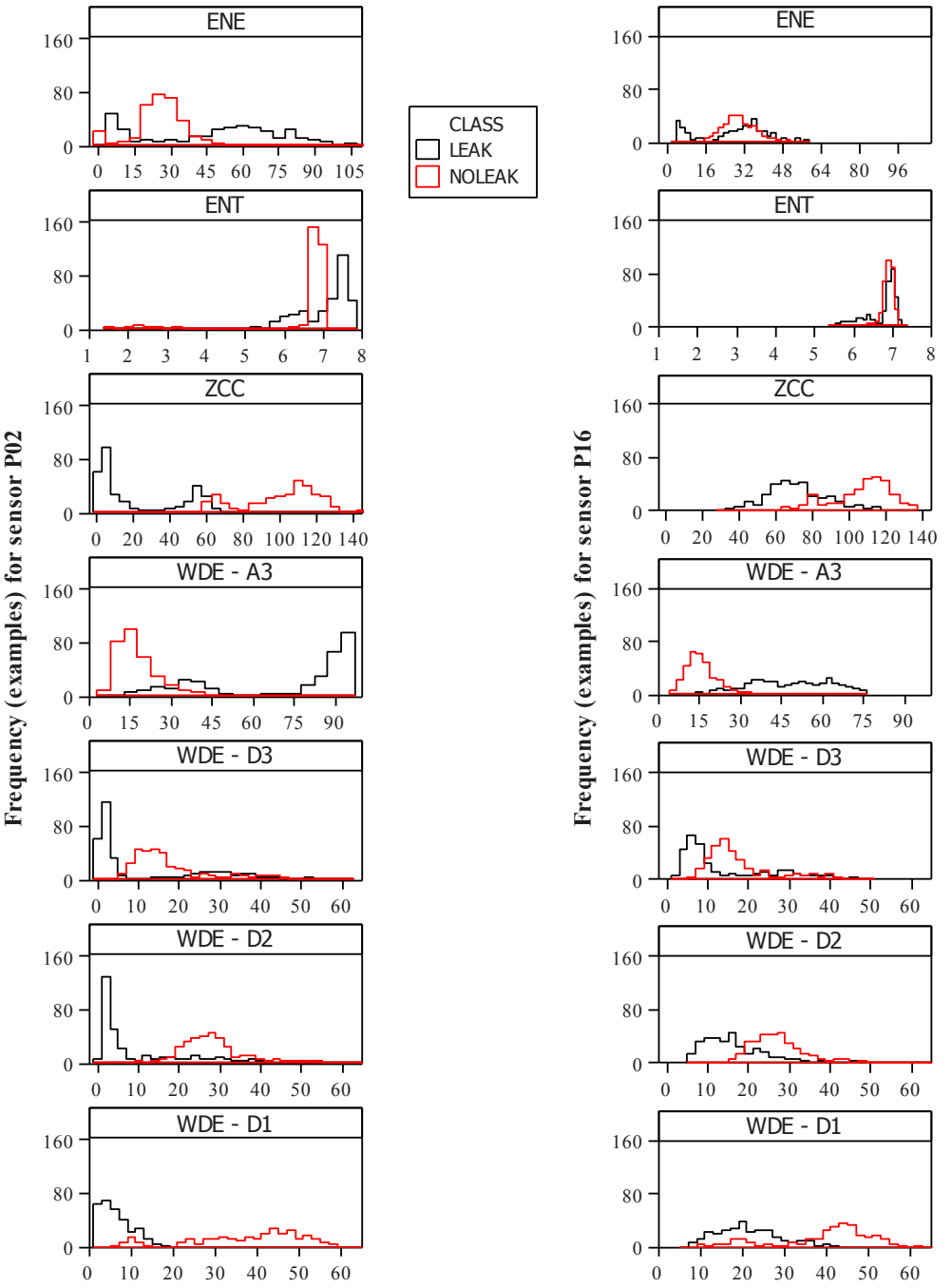

Figure 4. Histograms of feature vector values for LEAK and NONLEAK examples.

Table 2. Interval limits for 85% more frequent values of features extracted.

FEATURE	P02 NOLEAK (feature value)	P02 LEAK (feature value)	P16 NOLEAK (feature value)	P16 LEAK (feature value)
ENE	10.4 - 34.9	2.8 - 76.5	17.3 - 37.8	2.8 - 42.4
ENT	6.2 - 6.8	6 - 7.6	6.7 - 7	6 - 7
ZCC	58.5 - 121.9	1 - 57.5	76.1 - 123.9	44.9 - 92.5
WDE-A3	5.7 - 30.2	23.6 - 92.2	8 - 23.6	25.7 - 65.4
WDE-D3	4.7 - 33.6	0.3 - 34	7.7 - 34.3	1.7 - 30.5
WDE-D2	18.6 - 37.4	0.8 - 29.4	18.7 - 35.4	7.7 - 25.9
WDE-D1	9.3 - 50	1.3 - 12.6	15.6 - 50.4	9 - 33.8

The intervals of more frequent values for the LEAK examples are similar to those of the NOLEAK examples considering some features, while for other features the intervals are more clearly differentiated between classes. Following, the principal results for each feature are commented:

- For the ENE feature, for both closest (P02) and farthest (P16) sensors, the interval of common values of the NOLEAK examples is within the interval of the LEAK examples. Thus, it would be very difficult to determine the class to which an example belongs on the basis of its ENE value.
- For the ENT feature, the intervals for LEAK and NONLEAK are similar in the P02 sensor and almost identical in P16. As for ENE, the discrimination between LEAK or NONLEAK examples is not clear considering only their ENT values.
- The ZCC more frequent values are lower for the LEAK examples than for the NOLEAK examples (in both sensors); in fact for P02 the two intervals are disjointed. This trend is reasonable because the occurrence of a leak induces a change in the signal, therefor the values deviate from the mean (0), showing fewer zero-crossings in comparison to instances where there occurs no leak-caused change.
- For WDE - A3, the more frequent values of the NOLEAK examples are lower than those of the LEAK examples for both sensors. This means that after the wavelet decomposition of NOLEAK signals, a smaller part of the energy corresponds to the approximation (low frequency[†]) component, and the components of detail (high frequency) are more important. Differently, in the LEAK examples most of the energy is concentrated on the approximation component, due to the change caused by the leak. By analyzing the histograms of WDE detail components, the differentiation between the intervals of LEAK and NOLEAK examples is clearer for the higher frequency components (WDE – D1) than for lower frequency (WDE – D3) ones, and in a general sense the differentiation is less clear for sensor P16 (the farthest sensor) than for P02 (the closest sensor).

The histograms for the signals of other sensors, whose graphs are not shown here, exhibit a behavior for each feature similar to that previously described. Additionally, the more distant the sensor from the leak node, the more difficult the differentiation between the features values from both classes. WDE-A1 is the feature (component) that best maintains the distinction between the most common values of the two classes for different sensors.

4.2. Application to pattern recognition system (classification)

Binary classifiers were used to evaluate the efficiency of features to represent original signals for the recognition of the class to which each sample belongs. Thus, the feature vectors extracted are used to train a

[†] It is important to consider that in the "the more frequent values" or " frequency histograms" phrases, the word "frequency" refers how often one value or range occurs, expressed in units of quantity of occurrences. On the other hand, the "high/low frequency" phrases about de wavelet decomposition components refers the signal frequency, related with the number of cycles per time in periodic signals, and has the dimension of inverse time.

classifier, i.e. a methodology of supervised learning is applied to two classes: LEAK and NOLEAK. The methodology chosen is the decision trees, according to which the instance space is divided into subspaces adjusted by different models. Algorithm C4.5 (Quinlan 1993), which generates a set of sorted rules to be recursively applied to determine the class of each input vector, was implemented. Independent classifiers were constructed for each of the fifteen sensors by using six alternative types of input vectors:

- ALL, all four features, a vector with seven elements;
- ENE, energy, a vector with one element;
- ENT, entropy, a vector with one element;
- ZCC, zero crossing count, a vector with one element;
- WDE, distribution of the energy on the components of the wavelet decomposition, a vector with four elements;
- SGN, signal before features extraction, a vector with 240 elements. The use of original signals is possible here because all the examples were generated with a fixed size and avoiding erroneous data

For the training and evaluation processes of the decision tree, the 10-fold cross validation (Duda et al. 2001) was applied for each alternative analyzed. To quantify and compare the alternatives the area under ROC curve, a common indicator of performance for binary classifiers, was applied. The ROC curve (Receiver Operating Characteristic) considers both true and false responses in the classification tests, and the area under this curve is 1 for a perfect classifier and 0.5 for a random one (Faceli et al. 2011). Therefore, the larger the area (≤ 1) under the ROC curve (AU ROC), the better the performance. The AU ROC values for the six alternatives on each sensor, its media and its deviation are shown in Table 3. In Figure 5 each sensor is shown as a dot, with the AU ROC in y-axis and alternative input in the x-axis.

Table 3. Area under ROC curve for different types of input vectors.

Sensor	ALL	ENE	ENT	ZCC	WDE	SGN
P01	0.958	0.653	0.605	0.858	0.951	0.871
P02	0.987	0.906	0.932	0.973	0.973	0.915
P03	0.976	0.928	0.913	0.965	0.968	0.915
P04	0.989	0.910	0.918	0.966	0.966	0.932
P05	0.948	0.746	0.656	0.868	0.961	0.849
P06	0.975	0.907	0.928	0.943	0.977	0.926
P07	0.982	0.829	0.860	0.944	0.970	0.894
P08	0.967	0.766	0.715	0.901	0.961	0.885
P09	0.981	0.831	0.821	0.932	0.975	0.905
P11	0.966	0.822	0.856	0.940	0.973	0.920
P12	0.988	0.931	0.922	0.955	0.963	0.918
P13	0.968	0.826	0.838	0.898	0.957	0.904
P14	0.950	0.736	0.740	0.875	0.953	0.888
P15	0.977	0.741	0.757	0.877	0.967	0.900
P16	0.953	0.715	0.675	0.863	0.966	0.862
Average	0.971	0.816	0.809	0.917	0.965	0.899
Deviation	0.014	0.088	0.110	0.042	0.008	0.024

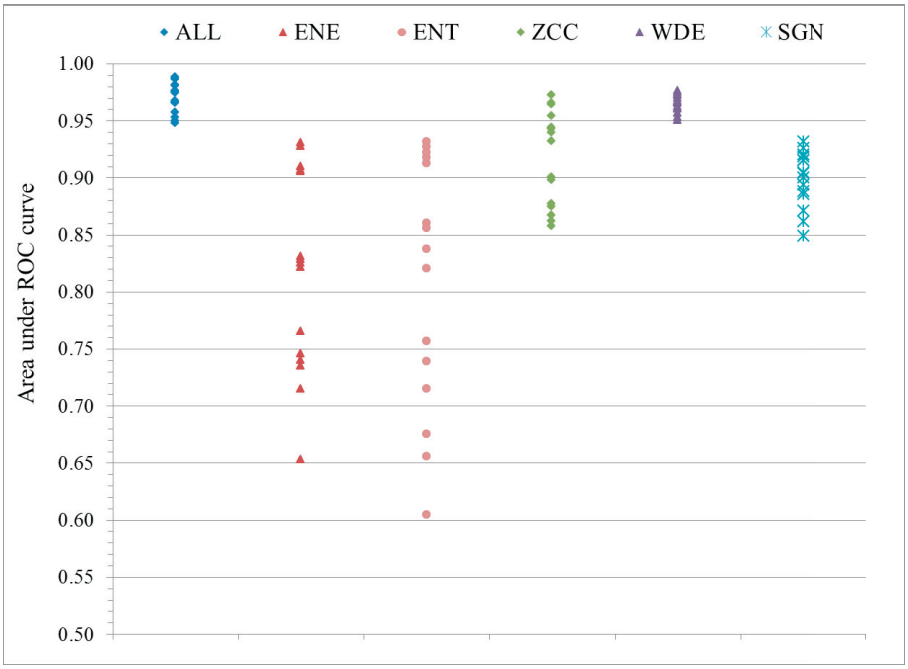

Figure 5. Area under ROC curve for different types of input vectors.

The results in Table 3 show that the best classifier performance (highest AU_ROC value) was obtained with vector ALL for 10 sensors, whereas vector WDE showed the best performance for the remaining 5 sensors and the second best performance for the other sensors. The worst performances were obtained by ENE (smallest AU_ROC value for 8 sensors), ENT (smallest for 6 sensors) and SGN (smallest for 1 sensor).

The average results for all sensors allow conclusion compatible with noted before about alternative input vectors: based on the mean AU_ROC values the best performance is obtained by using ALL, followed by WDE. Both standard deviation values and spread of the dots in Figure 5 show that the results for ENE and ENT have greater dispersion, while the results for WDE are the less scattered. The use of the original signal (SGN) provided better results in comparison with the two worst features, but the computational time was longer as much larger vectors (240 elements) were involved. The superiority of the mean performance with vector ALL over WDE is small and the dispersion is greater for WDE than for ALL. The use of additional features (further WDE) to complete the ALL vector requires a higher computational cost. Consequently it can be assessed that the use of ALL vector instead of WDE vector doesn't bring additional benefits enough to justify its use.

5. Conclusions

It has been verified the feasibility of represent a vector of large and variable size as a pressure signal in the water supply system by a vector of constant and much smaller size as a features vector. The feature vector preserves the information needed to identify whether the signal came from a system in which a leak there occurred.

Both analysis of histograms and performance indicators have led to the same conclusions about the choice among the four features analyzed, indicating that the best results were provided by WDE and ZCC. Regarding the use of original signals (SGN) in the classifier, it was identified that depends on the uniform size of all the examples, and the computational time are greater. In addition, by using SGN can be obtained good performances but smaller than whose of the best features.

The results of this exploratory study suggest further research on the use of extracting features and wavelet analysis with machine learning tools for the detection of leaks in water distribution networks based solely on

pressure signals acquired in the system. Therefore, the knowledge obtained is expected to improve the potential of leakage detection for real water supply systems during the next steps of investigation.

Acknowledgements

The authors would like to thank CAPES (Coordination for the Improvement of Higher Education Personnel) and CNPq (National Counsel of Technological and Scientific Development) for the scholarships and financial support.

References

Beale, R. & Jackson, T., 2010. Pattern Recognition. In *Neural Computing - An introduction*. New York: Taylor&Francis Group, pp. 15–37.

Duda, R., Hart, P. & Stork, D., 2001. *Pattern Classification* 2nd ed., New York: Willey-Interscience.

Faceli, K. et al., 2011. *Inteligência artificial. Uma abordagem de aprendizado de máquina.*, Rio de Janeiro: LTC.

Ferrante, M. & Brunone, B., 2003. Pipe system diagnosis and leak detection by unsteady-state tests. 2. Wavelet analysis. *Advances in Water Resources*, 26(1), pp.107–116.

Ferrante, M., Brunone, B. & Meniconi, S., 2009. Leak detection in branched pipe systems coupling wavelet analysis and a Lagrangian model. *Journal of Water Supply: Research and Technology—AQUA*, 58(2), p.95.

Ferrante, M., Brunone, B. & Meniconi, S., 2007. Wavelets for the Analysis of Transient Pressure Signals for Leak Detection. *Journal of hydraulic engineering*, 133(11), pp.1274–1282.

Meyer, Y., 1999. Time-Frequency / Time-Scale Analysis. In *Wavelet Analysis and its applications*. London: Academic Press.

Mounce, S.R., Boxall, J.B. & Machell, J., 2010. Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows. *Journal of Water Resources Planning and Management*, 136(3), pp.309–318.

Proakis, J.G. & Manolakis, D.G., 2007. *Digital Signal Processing* 4th ed., UpperSaddleRiver: Pearson Education.

Puust, R. et al., 2010. A review of methods for leakage management in pipe networks. *Urban Water Journal*, 7(1), pp.25–45.

Quinlan, J.R., 1993. *C4.5: programs for machine learning*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Romano, M., Kapelan, Zoran & Savic, D., 2012. Automated Detection of pipe burst and other events in water distribution systems. *Journal of Water Resources Planning and Management*.

Ye, G., Ph, D. & Fenner, R.A., 2011. Kalman Filtering of Hydraulic Measurements for Burst Detection in Water Distribution Systems. *Journal of pipeline systems engineering and practice*, 2(1), pp.14–22.