Chapter 3

Identifying Water Network Anomalies Using Multi Parameters Random Walk: Theory and Practice

Eyal Brill and Barak Brill

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.71566

Abstract

A noise pattern analysis is used to demonstrate how water quality events can be classified. The algorithm presented mimics a random walk process in order to measure the level and type of noise in the water quality data. The resulting curve is analyzed and four different cases are identified. i.e. sensor problem, water source change, operational change and contamination. For each problem, the algorithm identifies a different pattern. This pattern can be used later to reduce the level of false alarms in the monitoring system.

Keywords: water network, abnormality detection, multi parameters, clustering, unsupervised learning

1. Introduction

Anomaly detection in multivariate time series, such as water quality data, derived from water quality monitoring stations, is considered a non-trivial task. This is mainly due to the implication involved with both false positive and false negative situations.

In the case of false positive, i.e., the water is declared non-drinkable, and an alternative for customers must be found by the relevant water utility. This type of one-time event may be costly, while repetitive mistakes of this kind will eventually cause the monitoring system to be perceived as unreliable.

In the case of false negative, i.e., the monitoring system failed to detect a problem, some health hazard situations may develop and, in the long run, once again, the monitoring system may be considered unreliable.

Several methods have been suggested as a methodology for detecting abnormal events in similar cases. The basic approach for solving such problems may be based on unsupervised machine learning (USL), As described in detail by Celebi and Aydin [1]. One of the first and most fundamental methods of USL is based on clustering. *Clustering* is a methodology which groups vectors into several similar groups, where the members of each group are as similar as possible and the differences between groups are as great as possible. Clustering may be distance-based or density-based. Examples of distance-based clustering, such as the kMean algorithm, were presented by Knorr and Ng [2, 3] who compute abnormality score by counting neighbors to each point. More updated work in this field was introduced by Angiulli and Pizzuti [4] who compute the anomaly score of a data instance as the sum of its distances from its k-nearest neighbors. Ramaswamy et al. [5] extend this technique to spatial data. Their methods are also based on the kNN algorithm. Bay and Schwabacher [6] introduced the same algorithm with regard to pruning.

Another clustering philosophy is based on density of points. Examples of such cases are Breunig et al. [7, 8], in relation to relative density. Jin et al. [9] showed how some of the calculations can be skipped. Tang et al. [10] further improved clustering by adding the idea of a connectivity-based outlier factor, which refers to the number of connections between points. Jin et al. [9] also introduced improvements by adding the idea of symmetric neighborhood relationship. The main density-based algorithm is known as the EM algorithm.

In both methods - distance or density - the result is a multi-dimensional data structure which contains centroids. A *centroid* is a center of a group. It is, in the broadest sense, the group's center of gravity, i.e., the coordinates of the center of the group in each dimension. Once such a data structure exists, each new incoming record is evaluated based on its distance from the most nearest centroid. If the new incoming record is too far from any known centroid, it is declared a suspicious record, one that should be examined. After examination, the new point is classified, either as a True or False event. This classification is added to the model's learning set.

A second method, which has been used for abnormality detection, is based on a prediction methodology. According to this methodology, one of the variables in a multi-dimensional space is considered to be a dependent variable, whose value or class (in the case of discrete values) is related to the other variables. Given that, a mathematical model is then constructed, which describes the relation between the dependent variable and all other variables. This model can be based, for example, on linear regression, decision trees or neural networks. In this case, each new incoming record is used to generate a prediction. If the predicted value is too far or has a different class value than the actual value of the dependent variable, the new incoming record, is again considered abnormal and must be investigated. An example of such a methodology is given by Stefano et al. [11], Odin and Addison [12], Hawkins et al. [13] and Williams et al. [14].

A third method, which will be the focus of this chapter, is based on examining noise pattern changes, generated by the multi-dimensional data. Several methods have been suggested along this line. A fundamental method has been demonstrated by Cheng et al. [15], which used an RBF[1] function to identify abnormal patterns in a moving window.

[1]Radial basis function (see https://en.wikipedia.org/wiki/Radial_basis_function).

The methodology suggested in this chapter is based on Brill [16]. This methodology is based on detection and classification changes in noise patterns. Noise is measured based on the distance traveled by an artificial particle located at the normalized coordinates of the multi-dimensional vector. The difference between Cheng et al. [15] and Brill [16] is in the classification type of the abnormal events. While the first uses a True and False classification, the second adds the hazard and non-hazard classification. As will be demonstrated later in this chapter, patterns in this noise can be explained by different events related to the water network.

The aim of this chapter is to describe a different methodology for abnormality detection in water network. The chapter describes the basic model and presents a numerical illustration of the calculation framework. Than it illustrates four different cases which enable the identification of changes in the noise pattern and their related events. The last section concludes the chapter.

2. The model

The following section presents an overview of the mathematical model used in this chapter. It starts by examining Brownian Motion (BM), which is named after Robert Brown [17], who discovered the typical movement of flowers seeds on the water's surface. Einstein [18] used the idea of BM in order to provide precise details about the movement of atoms. This explanation was later further validated by Perrin who awarded the physics noble price for 1926.

BM was also used by Louis Bachelier [19] (1900) in his Ph.D. thesis "The Theory of Speculation", in which he presented a stochastic analysis of the stock and option markets. His work went on to inspire the novel work of Black and Scholes [20], which awarded them the Nobel Prize in economics.

Modern literature gives many examples of the usage of BM in various areas; most are related to biology, chemistry, physics and other fields of life sciences. However, it is rare that such a technique or a similar one is used for the analysis of abnormal water events - the main topic of the current chapter.

One of the central results of BM theory is an estimation of the traveling distance of a particle, which travels using random movement in a given time interval across a multi-dimensional space. According to this theory, if ρ(x,m) is the density function of particles at location x (were x is a single dimension, e.g., one axis) at time m, then ϱ satisfies the diffusion equation:

$$\frac{\partial \rho}{\partial m} = D\frac{\partial^2 \rho}{\partial x^2} \tag{1}$$

where D is the *mass diffusivity*, a term which measures how fast particles of a given type may move in a specific material, in our case, water. The solution of Eq. (1) gives a density function with a first moment, which is seen to vanish, and a second moment given by:

$$\overline{x^2} = 2D * m \tag{2}$$

The left side of Eq. (2) expresses the distance at which a particle can be found from its origin, given the elapsed time m and the diffusivity parameter D. Assuming x is distributed normally, the maximum value a particle can travel for a given time can be calculated using (2) with a given confidence interval.

Using Eq. (2), and assuming that the left hand side of (2) is distributed normally, and its standard deviation (S) can be estimated empirically, the probability of a particle to travel a given distance from its origin within m units of time can be calculated by:

$$L = \sqrt{2D * m} + S^* t(\alpha) \tag{3}$$

where $t(\alpha)$ is the confidence interval factor, based on the level of confidence required (drawn from the student distribution). Thus, if a particle is found within m steps at a distance which is greater than L units from its origin, it is considered an abnormal event.

In the physical or chemical diffusion process, the value of D is determined based on material properties, and the value of m is measured in continuous time. In the current model, these values should be determined using another methodology as explained in the following paragraphs.

Let us denote with vector X_m the set of quality measurements of water at each moment m and, assuming X_m has K dimensions. The value of X_m can be normalized with the following process equation:

$$\hat{v}_m^k = \frac{X_m^k - X_{\min}^k}{X_{\max}^k - X_{\min}^k} \tag{4}$$

where v_m^k is the normalized k dimension of vector V_m and m is the discrete time index. The subscripts max and min refer to the maximum and minimum value of this dimension over the whole data set.

Let's also define the distance between two vectors to be denoted by DN_m^n (where DN stands for Dynamic Noise). This measurement is calculated as the normalized Euclidian distance between two values of V_m and is given by the equation

$$DN_m^n = \hat{V}_m - \hat{V}_{m-n}. \tag{5}$$

Note please that unlike in the case of BM where the distance of a particle from its origin increases with time, in the case of the DN, the particle may turn back to its origin.

An illustration of this distance in a normalized two-dimensional space is shown in **Figure 1**. In this case, V_m is a two-dimensional vector.

In terms of **Figure 1**, assuming a dataset with M records and two variables in each record (x_1 and x_2), one may look at this dataset as a description of location for a particle in each time stamp. After normalizing the dataset according to Eq. (4), the Euclidian distance between each two points is the distance this particle travels. If the distance is measured in a five-step gap, the result may be a chart as shown in **Figure 1**.

http://dx.doi.org/10.5772/intechopen.71566

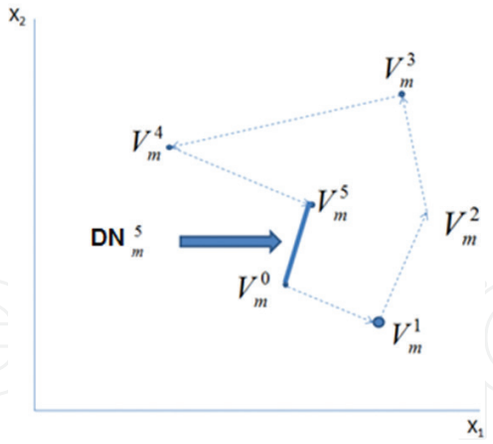

Figure 1. BM distance traveling in two dimensions.

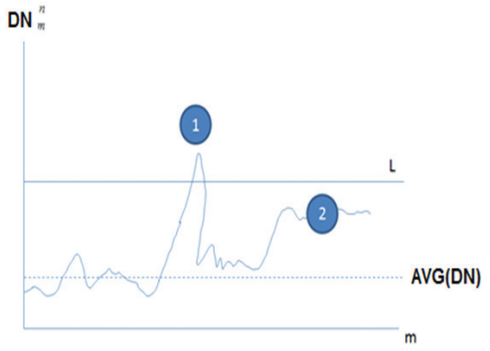

Figure 2. DN over time.

Figure 2 shows a schematic chart of DN_m^n over time without exceeding the L limit. As explained previously, under normal conditions the value of DN_m^n is not expected to go above the value of L, with a confidence level equal to $1 - \alpha$. The value of L can also be obtained. In **Figure 2**, point 1 refers to values of DN above the level of L for a constant period, while point 2 underline a significant change in the value of DN.

The first is calculated by comparing the value of DN_m^n with the result of Eq. (3), as calculated by data accumulated until the m point in time. The second is calculated by comparing the average normalized value of DN_m^n at a fixed-width moving window of DN_m^n with the average DN_m^n known prior to this window.

3. Numerical example

This section numerically describes the calculation procedure as described in the previous section. **Table 1** contains an example data set with 20 records. The measured variables are Free Chlorine (CL), Turbidity (TU), pH and Conductivity (CO). These are a common water quality indicators.

Max	1	1	9.5	600					
Min	0	0	6	300		Normalized data			
No	CL	TU	pH	CO	CL	TU	pH	CO	DN
1	0.60	0.09	7.69	536.00	0.60	0.09	0.48	0.79	0.0000
2	0.59	0.09	7.70	541.50	0.59	0.09	0.49	0.81	0.0000
3	0.60	0.08	7.70	538.00	0.60	0.08	0.49	0.79	0.0000
4	0.60	0.09	7.70	538.00	0.60	0.09	0.49	0.79	0.0000
5	0.60	0.09	7.71	538.00	0.60	0.09	0.49	0.79	0.0000
6	0.60	0.08	7.70	537.00	0.60	0.08	0.49	0.79	0.0003
7	0.60	0.09	7.70	536.00	0.60	0.09	0.49	0.79	0.0001
8	0.60	0.12	7.70	535.50	0.60	0.12	0.49	0.79	0.0008
9	0.60	0.12	7.70	536.00	0.60	0.12	0.49	0.79	0.0010
10	0.60	0.09	7.70	533.00	0.60	0.09	0.49	0.78	0.0002
11	0.60	0.09	7.70	533.00	0.60	0.09	0.49	0.78	0.0002
12	0.59	0.08	7.70	529.00	0.59	0.08	0.49	0.76	0.0016
13	0.60	0.08	7.71	529.00	0.60	0.08	0.49	0.76	0.0024
14	0.59	0.09	7.71	545.00	0.59	0.09	0.49	0.82	0.0017
15	0.59	0.09	7.71	545.00	0.59	0.09	0.49	0.82	0.0017
16	0.60	0.09	7.71	545.00	0.60	0.09	0.49	0.82	0.0029
17	0.59	0.08	7.71	544.00	0.59	0.08	0.49	0.81	0.0025
18	0.59	0.09	7.71	545.33	0.59	0.09	0.49	0.82	0.0000
19	0.59	0.08	7.71	540.00	0.59	0.08	0.49	0.80	0.0004
20	0.59	0.08	7.71	538.00	0.59	0.08	0.49	0.79	0.0006

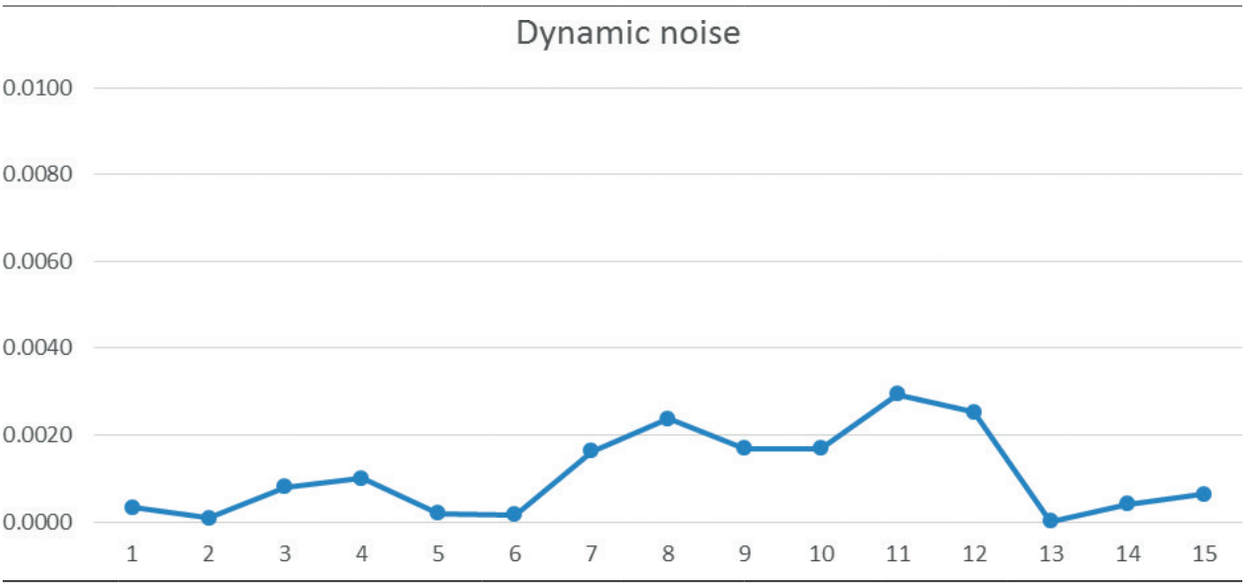

Table 1. Data for numerical example.

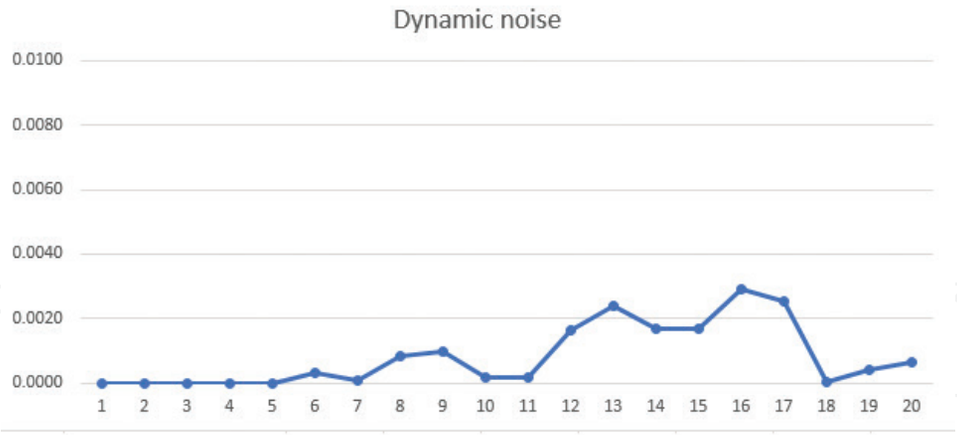

Figure 3. Dynamic noise chart.

The first two rows display the minimum and maximum values of each variable. These values were used to normalize the records in the left side of the table to the right side of the table. After normalization, a vector of the Euclidian Distance between each pair of records with a difference of 5 steps was calculated. This vector is the most right column in the table. The first value in this vector (located in row 6) contains the distance between record 6 to record 1. The second contains the distance between record 7 to record 2. The chart in **Figure 3** displays relevant DN chart.

In the following section, examples from a read data set are used to illustrate the analysis framework of abnormality detection and classification using this methodology.

4. Real data example

The following examples refer to a real data set recorded at a field station with measurements, as presented in **Table 2**. The table's quality measurements include Free Chlorine (Cl measured in ppm), Turbidity (TU measured in NTU), pH, Conductivity (CO measured in mS), and pressure (PRI measured in bars). For each measurement, the algorithm dynamically calculates the minimum and maximum values of the last 48 hours. (see first two rows of **Table 1**).

Symbol	Measurement	Units	Minimum	Maximum
Cl	Free Chlorine	mA	0	2
TU	Turbidity	NTU	0	2
pH	pH	pH	6	9.5
CO	Conductivity	mS	0	800
PRI	Pressure	Bar	0	15

Table 2. Measurements.

Typical minimum and maximum values are shown in **Table 2**. Given these minimum and maximum values, the raw measurements are transformed into normalized measurements, as shown by Eq. (4) of Section 2. The normalized measurements are used to calculate the "Dynamic Noise", as shown in Eq. (5), with a lag difference between the records of 10 time-stamps. **Figure 4** shows the distribution of the dynamic noise values.

As can be seen from the histogram, a value which is more than 0.25 is rare (see red arrow). Hence, the threshold for the dynamic noise was set to 0.3. In terms of Section 2 of this chapter, L = 0.3.

The first data analysis step with regard to the dynamic noise algorithm is to estimate the normal conditions, i.e., to observe how a dynamic noise curve behavie in case of a normal data flow. **Figure 5** shows a normal period of time for the four water quality measurements. Note that pH ranges between 7.70 and 7.79; Free Chlorine ranges between 0.37 and 0.48; Conductivity usually has an average of around 520–530 with short drops to 450; and Turbidity ranges between 0.09 and 0.12.

Figure 6 shows the equivalent dynamic noise for the relevant measurements. As can be seen, the values range between 0.03 and 0.25 at the most.

We will now discuss four different cases, in which the dynamic noise violation threshold is analyzed. Note please that violation of the threshold L triggers an alarm only after a delay time in which the value of the DN is above the level of L. This in order to avoid false alarms caused by short spikes.

Case 4.1: Malfunctioning of sensors

The first case shows a situation in which two of the sensors stopped functioning for a short period of time. As can be seen from **Figure 7**, the Cl and the pH dropped suddenly to zero for a short period. This may result from communication problems, which are very common with distributed I/O.

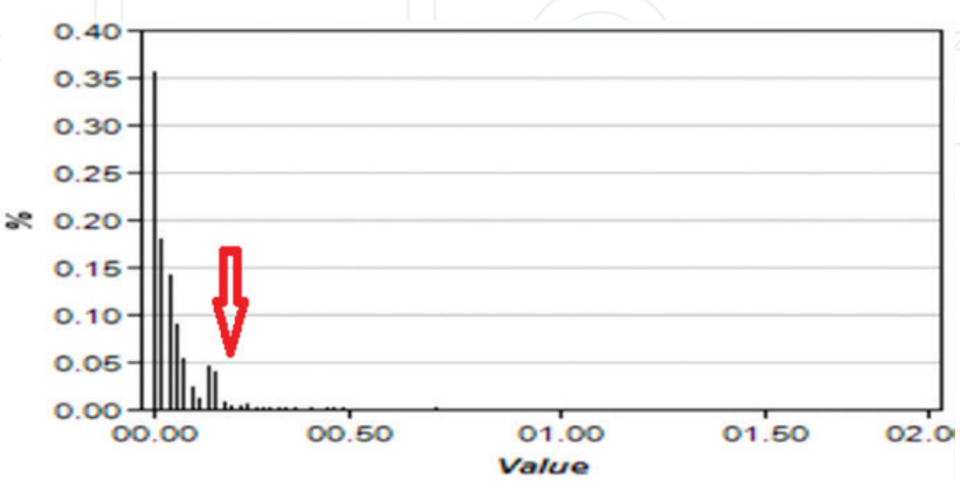

Figure 4. DN histogram.

http://dx.doi.org/10.5772/intechopen.71566

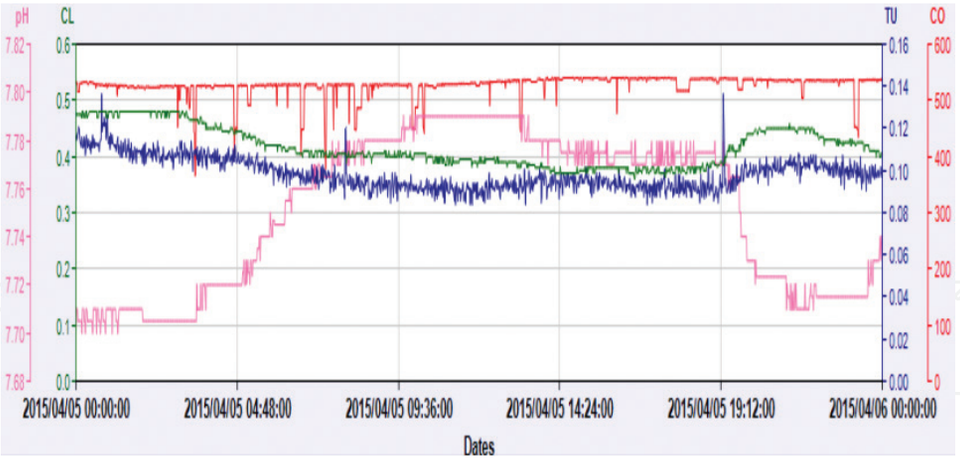

Figure 5. Normal measurements.

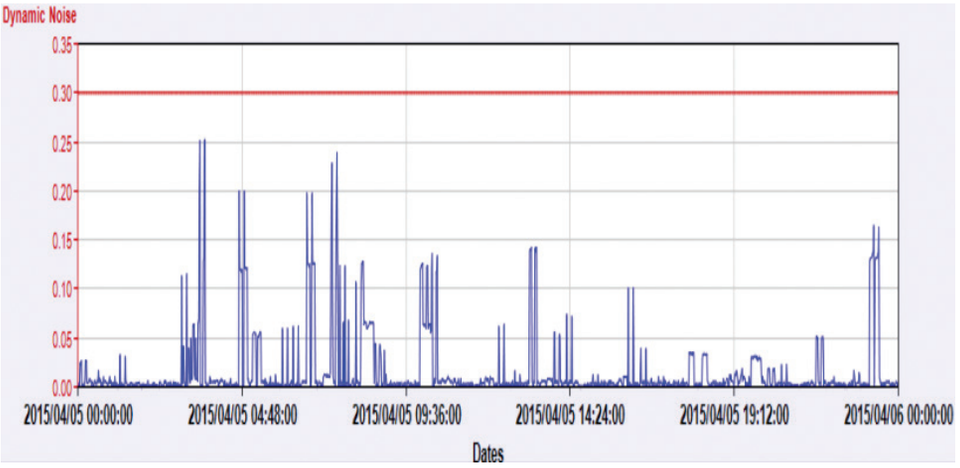

Figure 6. DN for normal measurements.

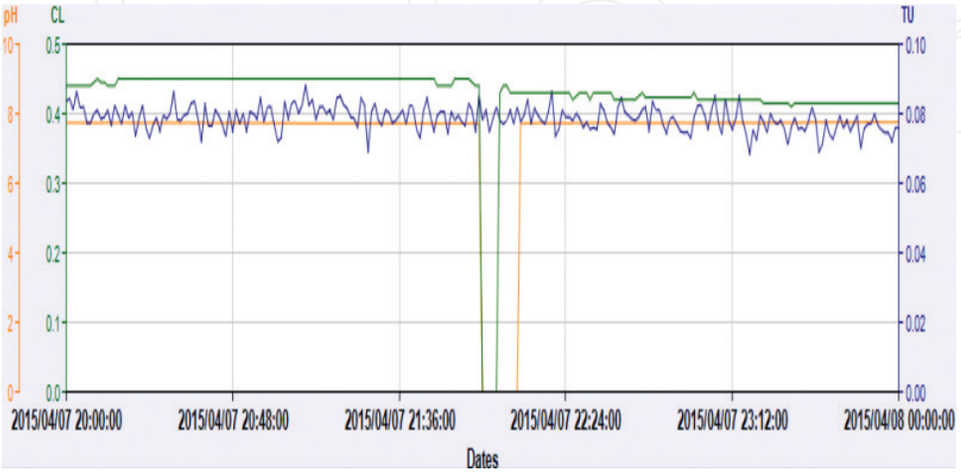

Figure 7. Malfunctioning of sensors.

Figure 8 shows the resulting dynamic noise curve of the sensor's malfunctioning. As can be seen, the drop in the values of Cl and pH causes a sharp increase in the value of the dynamic noise. After a short period, when the sensors resume functioning, the value of the dynamic noise drops back to a level below the red threshold line (0.3).

Note also that if the sensors remain non-functional for a long period of time, and the algorithm stops using the values of these sensors as part of Eq. (5), the level of the dynamic noise curve will be lower during steady state, since less sensors are transmitting data.

The gray box around the area of the event depicts the shape of the dynamic noise curve as a rectangle. This is due to the sharp change in the values of certain sensors. This sharp change can only occur during sensor failure . A chemical change in water quality cannot occur within 1 minute.

Case 4.2: Operational change

The second typical change is an *operational change*. This is defined as a situation in which one of the variables controlled by operators has been changed. Some examples of such variables may be pressure or flow. An operational change may influence the variables' quality. One such example of this type of situation is shown in **Figure 9**. The black line shows a change in the pressure (PRI) value. Shortly after this change, a peak in the Turbidity value is recorded (see red line in **Figure 9**).

Figure 10 shows the corresponding changes in the dynamic noise curve. The chart indicates that the operational change also results in the violation of the dynamic noise's "red line". However, this can be explained by the change in the operational variable.

Since operational changes are also sudden changes, the peak in the dynamic noise curve is immediate. However, the return to normal happens gradually. This is the reason why the gray area has a triangular shape in **Figure 10**.

Case 4.3: Water source change

The third case illustrates what happens when a change is made to the water source. In this case, if the attributes of the water from the new source are different, the footsteps of the water source change can also be seen in the dynamic change curve.

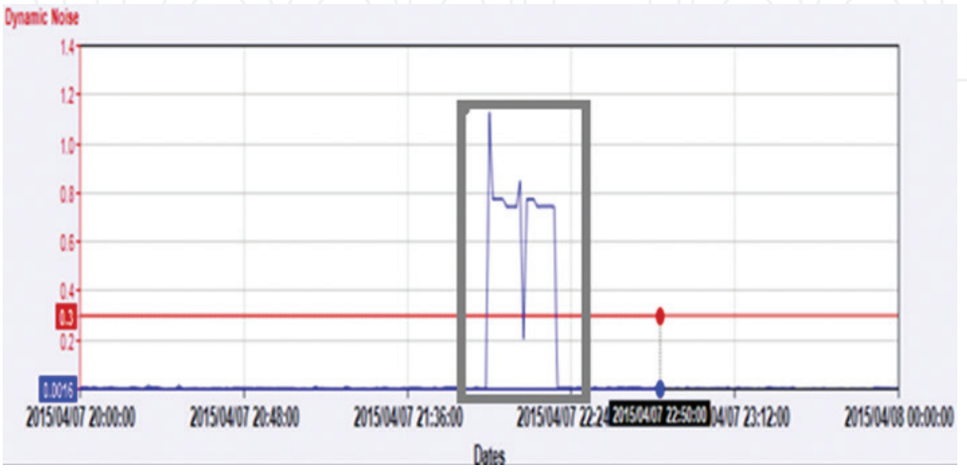

Figure 8. Noise curve for sensor malfunctioning.

http://dx.doi.org/10.5772/intechopen.71566

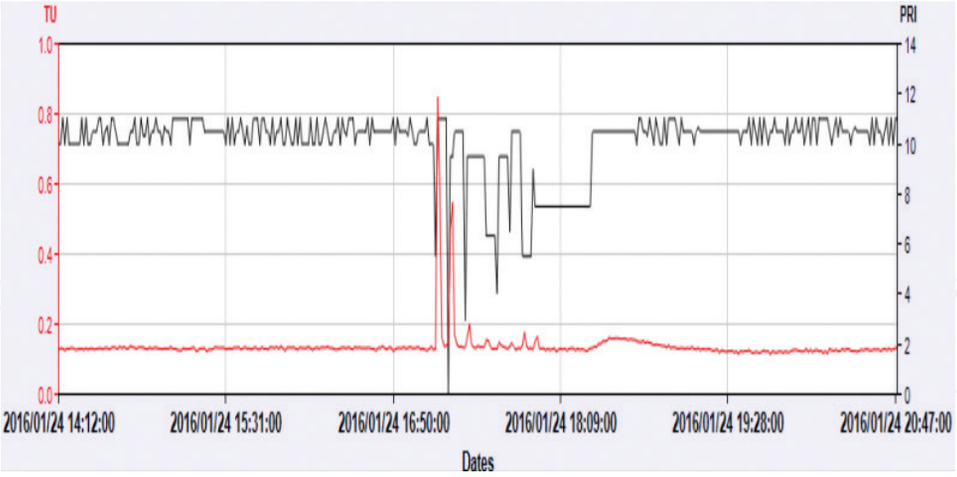

Figure 9. Operational change - raw data.

Figure 11 shows a change in the water source. Conductivity rose from a level of 345 mS to a level of 385 mS (within 6 hours, from 10:00 am to 4:30 pm). Together with this, the pH level dropped from 8.20 to 8.07.

The effect on the dynamic noise curve can be seen in **Figure 12**. The average value of the dynamic noise curve has changed. This is due to the change in the noise level of the water measurements from the new source.

A change in water source, which can take between 30 and 60 minutes and up to several hours, will result in a change in the noise of the dynamic noise curve. Sometimes, it will also cause a change to occur in the average value of the dynamic noise.

Case 4.4: Contamination event

Finally, we have the case of contamination. **Figure 13** shows raw data from a typical contamination event. The event starts with a drop in the Free Chlorine (see green line in **Figure 13**). This is due to chlorine consumption by the contaminator. Shortly after, the Turbidity level starts to rise (see blue line in **Figure 13**).

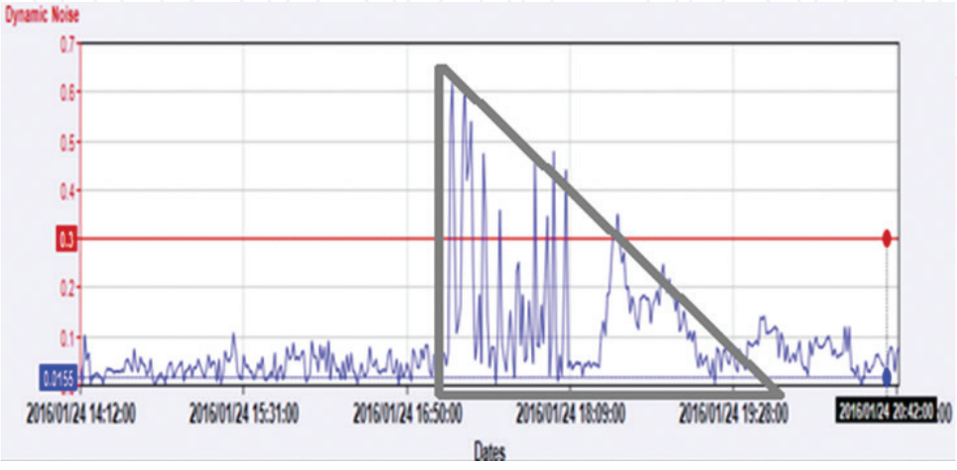

Figure 10. Operational change - dynamic noise curve.

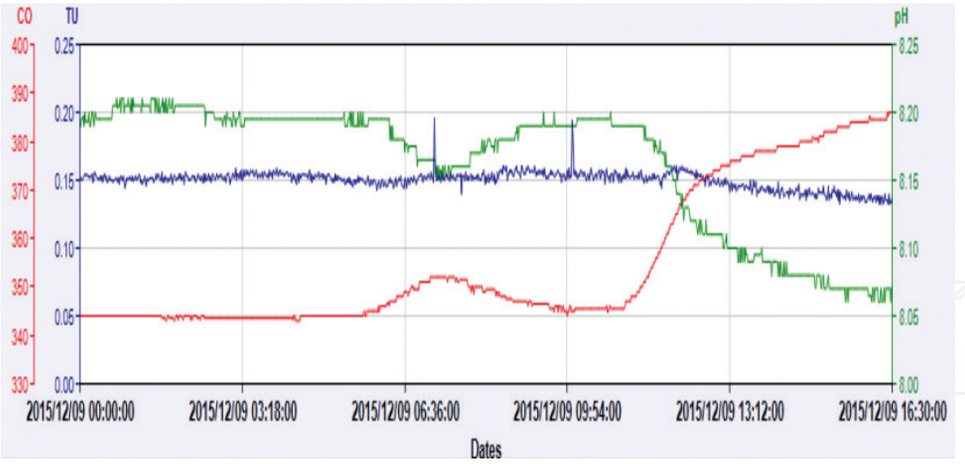

Figure 11. Water source change - raw data.

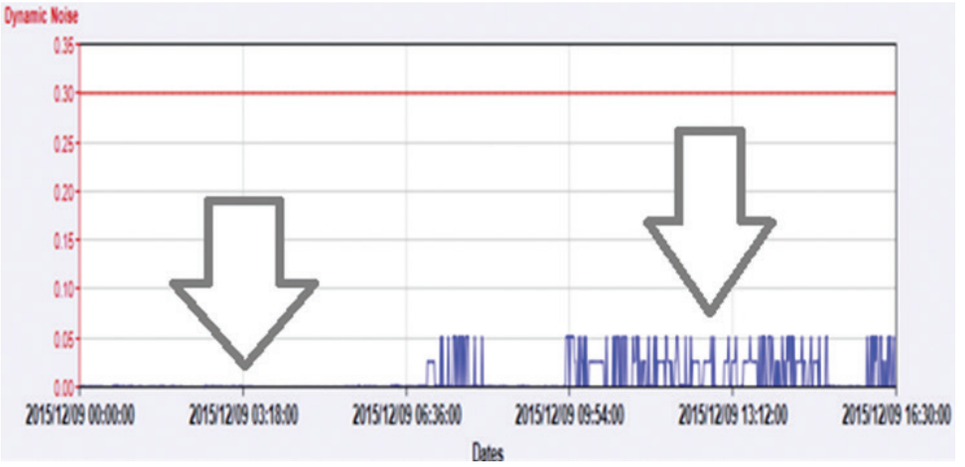

Figure 12. Water source change - dynamic noise.

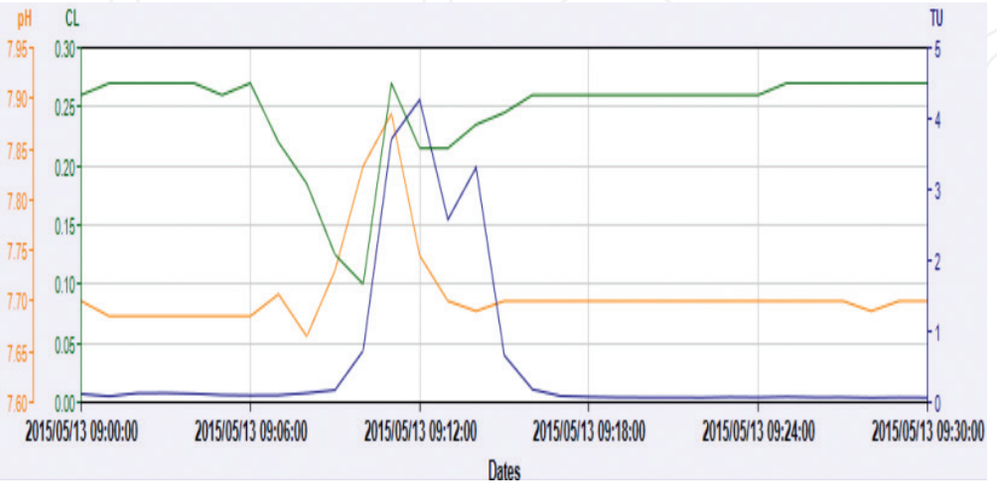

Figure 13. Contamination event - raw data.

http://dx.doi.org/10.5772/intechopen.71566

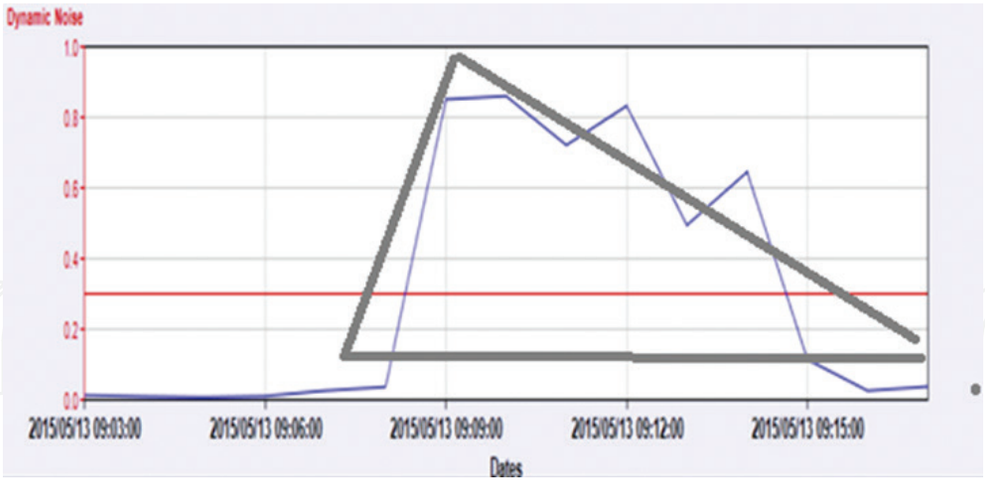

Figure 14. Contamination event - dynamic noise.

At the same time, the chlorination dosing system reacts to the situation and the level of Free Chlorine once again rises to a normal level; and due to the diffusion factor, the Turbidity level gradually drops. The result of this event can be seen in the dynamic noise line shown in **Figure 14**.

As seen in **Figure 14**, when the event starts, the dynamic noise curve rapidly rises above the red line of the threshold. It is fast, but not steep, as in the case of sensor malfunctioning or water source change. Once the maximum level of contamination has been obtained, the level starts to gradually drop, due to the diffusion effect, which causes the contamination to be diluted with the incoming water. This is why the right side of the curve in **Figure 14** is not symmetric to the left side of the blue curve. The overall situation creates the shape of non-symmetric triangle, as is illustrated by the gray area in **Figure 14**. Note that the farther the contamination's penetration to the system is from the measuring point, the less non-symmetric and the shorter the triangle will be. This is due to the dilution effect.

5. Concluding remarks

The current chapter demonstrates how the simple pattern recognition of a curve created by a noise capture process, similar to a random walk, can be used to classify different types of abnormal events. The presented algorithm uses the imaginary center of gravity of the water quality measurements in order to measure the noise of the process captured as traveling distance. It has been shown that the created curve has a maximum value, due to the nature of the process. This threshold is violated when abnormal events occur.

Four different types of abnormal events were examined: malfunctioning of sensors, operational change, water source change and contamination events. Numerical examples based on real data show that each of the events has a different "signature", which enables the identification of the event's nature.

The current chapter shows how water analytics can be used as part of the information system which helps operators protect the water system. The above framework can also assist control systems in regard to the automatic classification process of observed events, in order to reduce the level of false alarms in water monitoring systems. For example, this may be achieved by eliminating alarms like the first three types analyzed in this study and notifying operators only in the case of contamination event alarms.

Author details

Eyal Brill[1]* and Barak Brill[2]

*Address all correspondence to: eyalb@hit.ac.il

1 Faculty of Technology Management, Holon Institute of Technology, Israel

2 Department of Statistics and Operations Research Tel Aviv University, Israel

References

[1] Emre Celebi M, Aydin K. Unsupervised Learning Algorithms. Springer International Publishing, Switecrland; 2016. ISBN: 978-3-319-24211-8

[2] Knorr EM, Ng RT. A unified approach for mining outliers. In: Proc. Conf. of the Centre for Advanced Studies on Collaborative Research (CASCON), Toronto, Canada; 1997

[3] Knorr EM, Ng RT. Algorithms for mining distance-based outliers in large datasets. In: Proc. Int. Conf. on Very Large Data Bases (VLDB), New York, NY; 1998

[4] Angiulli F, Pizzuti C. Fast outlier detection in high dimensional spaces. In: Proc. European Conf. on Principles of Knowledge Discovery and Data Mining, Helsinki, Finland; 2002

[5] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. In: Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX; 2000

[6] Bay SD, Schwabacher M. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In: Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD), Washington, DC; 2003

[7] Breunig MM, Kriegel H-P, Ng RT, Sander J. OPTICS-OF: Identifying local outliers. In: Proc. European Conf. on Principles of Data Mining and Knowledge Discovery (PKDD), Prague, Czech Republic; 1999

[8] Breunig MM, Kriegel H-P, Ng RT, Sander J. LOF: Identifying density-based local outliers. In: Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), Dallas, TX; 2000

[9] Jin W, Tung A, Han J. Mining top-n local outliers in large databases. In: Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), San Francisco, CA. 2001

[10] Tang J, Li H, Cao Y, Tang Z. Email data cleaning. In: KDD'05, August 21-24, 2005, Chicago, Illinois, USA. Copyright 2005 ACM 1-59593-135-X/05/0008; 2005

[11] Stefano C, Sansone C, Vento M. To reject or not to reject: That is the question: An answer in the case of neural classifiers. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2000;**30**(1):84-94

[12] Odin T, Addison D. Novelty detection using neural network technology. In: Proceedings of the COMADEN Conference; 2000

[13] Hawkins S, He H, Williams GJ, Baxter RA. Outlier detection using replicator neural networks. In: Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery. Aix-en-Provence, France, September 4-6, 2002. pp. 170-180

[14] Williams G, Baxter R, He H, Hawkins S, Gu L. A comparative study of RNN for outlier detection in data mining. In: Proceedings of the IEEE International Conference on Data Mining. IEEE Computer Society, Maebashi City, Japan, December 2002; p. 709

[15] Cheng H, Tan P-N, Potter C, Klooster S. Detection and characterization of anomalies in multivariate time series Haibin Cheng, Pang Ning Tan, Christopher Potter, Steven Klooster. In: Proceedings of the 2009 SIAM International Conference on Data Mining; 2009

[16] Brill E. Dynamic Brownian motion with Density superposition for abnormality detection. PCT. Application number: 14601862; 2015

[17] Brown R. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. The Philosophical Magazine. 2009;**4**:161-173. DOI: 10.1080/14786442808674769

[18] Einstein A. Investigations on the theory of Brownian movement. Annalen der Physik. 1905;**322**(8):549-560

[19] Davis M, Etheridge A. Louis Bachelier's Theory of Speculation: The Origins of Modern Finance. Princeton; Oxford: Princeton University Press, JSTOR; 2006

[20] Black F, Scholes M. The pricing of options and corporate liabilities. Journal of Political Economy. 1973;**81**(3):637-654

IntechOpen