IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 36, No. 6, December, 2022, pp. 969-977
Journal homepage: http://iieta.org/journals/ria

Faulty Detection System Based on SPC and Machine Learning Techniques

Mohamed Elamine Benrabah[1*], Ouahab Kadri[2], Kinza Nadia Mouss[1], Abdelghani Lakhdari[3]

[1] Automatics and Production Laboratory, Department of Industrial Engineering, Batna 2 University, Chahid Bokhlof Street, Batna 5000, Algeria
[2] Department of Computer Science, Batna 2 University, Chahid Bokhlof Street, Batna 5000, Algeria
[3] Department CPST, Higher School of Industrial Technologies, P.O. Box 218, Annaba 23000, Algeria

Corresponding Author Email: m.benrabah@univ-batna2.dz

https://doi.org/10.18280/ria.360619

Received: 20 November 2022
Accepted: 20 December 2022

Keywords:
diagnostic, manufacturing process, SPC, anomaly detection algorithms, case study

ABSTRACT

Starting from a worrying observation, that companies have difficulties controlling the anomalies of their manufacturing processes, in order to have a better control over them, we have realized a case study on the practical data of the Fertial Complex to analyze the main parameters of the ammonia neutralization by nitric acid process. This article proposes a precise diagnostic of this process to detect dysfunction problems affecting the final product. We start with a general diagnosis of the process using the SPC method, this approach is considered an excellent way to monitor and improve the product quality and provides very useful observations that allowed us to detect the parameters that suffer from problems affecting the quality. After the discovery of the parameters incapable to produce the quality required by the standards, we applies two machine learning technologies dedicated to the type of data of these parameters for detected the anomaly, the first technique called The kernel connectivity-based outlier factor (COF) algorithm consists in recording for each object the degree of being an outlier, the second technique called the Isolation Forest, its principle is to establish a forest to facilitate the calculation and description. The results obtained were compared in order to choose which is the best algorithm to monitor and detect the problems of these parameters, we find that the COF method is more efficient than the isolation forest which leads us to rely on this technology in this kind of process in order to avoid passing a bad quality to the customer in future.

1. INTRODUCTION

Since the 1920s, the theory of statistical process control (SPC) has played an effective role in improving and controlling product quality, SPC is one of the tools that many large manufacturing plants around the world have been adopting in order to remain competitive in the global marketplace. The approach focuses on reducing costs and probability that problems will be passed to customers by collecting data samples at different temporal points of the process, so that process disturbances that may affect the quality of the final product or service can be detected and corrected, thus, it is possible to predict the abnormal situation of the production process, so that the measures to eliminate the anomalies and restore the process stability can be taken in timely time, to achieve the objective of quality control and improvement. Diagnosis of the neutralization process is performed using SPC and machine learning techniques, which can advance product quality. The SPC has many theories, among the most important and practical theories are the statistical indicators. If the manufacturing process is affected by random factors and specific factors, then the statistical indicators will presente abnormal results. In other words, these indicators are set up for the continuous monitoring of the process adjustment state and in particular of the machine. They are a very good tools to show when you have a problem and when you have succeeded in correcting them. For example, the indecator skewness when the result is greater than zero, it indicates that the saturation is at the lower tolerance limit. And when the result is less than zero, it indicates that the saturation is at the upper tolerance limit. Figure 1 shows the continuous process monitoring system. This system was not available at the beginning of the SPC application, as the statistical indicators were calculated manually. After the birth of computers and smart algorithms, the system of continuous monitoring appeared. It facilitates the operation and allowed great efficiency in tracking products.

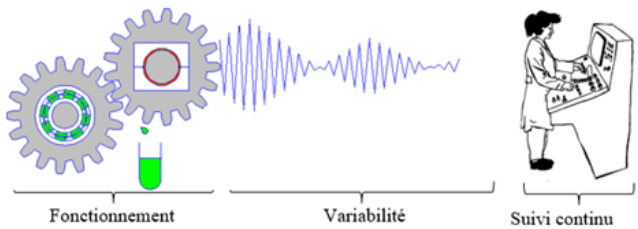

Figure 1. Continuous process monitoring

Anomaly detection is an essential element in the process monitoring domain, it models what is normal in order to discover what is not. It is an important part of data mining that researchers are more and more interested in. This method also improves data quality by removing or replacing anomalous data. In other cases, the anomalies reflect an event and provide new useful knowledge. For example, it can prevent material

damage and thus encourage predictive maintenance in the industrial domain. It has applications in many other fields such as health, cybersecurity, finance, natural disaster prediction, and many others. Data exists in many forms: static data, data streams, structured and unstructured data, etc. Each type of data is relevant to one or more domains. The multitude of data types and their different characteristics imply the existence of different methods of anomaly detection, each of them has its effectiveness in a particular domain, with a given objective. These methods generally use a decision threshold to isolate anomalies based on different techniques such as classification, clustering, regression, nearest neighbor and statistical tools.

In this paper, a statistical analysis is performed through statistical indicators, and the results obtained will be discussed and interpreted to discover the parameters that have problems affecting the final product. After the general diagnosis, it appears that the problem is in the flows of the neutralizer, so we applied two techniques of anomaly detection adopted to our type of data: Connectivity-Based Local Outlier and Isolation Forest, and based on the results of the evaluation, we chose the most effective technique to track and detect the problems of these parameters with the highest accuracy possible.

2. LITERATURE

The Anomaly detection is transversal to any domain that exploits data, and thus has many possible applications. Since application domains have their specificity depending on the data generated or exploited, not all methods of anomaly detection are suitable for all application domains. Yajie Cui made a review covering several application domains [1]. Gupta and al reviewed the temporal anomaly detection methods that are applicable in several different domains [2]. The Intrusion detection consists of the analysis of a target usually a network or host to detect anomalous behavior [3]. It is in effect fraudulent attempts to access a resource by violating the security set up for the target in question [1]. Fraud detection allows the identification of suspicious activities conducted by an individual usually under a false identity (impersonation) [4].

The emergence of new technologies has led to the generation of different types of data. There are several datasets with different characteristics that bring new challenges in anomaly detection. Different anomaly detection algorithms are applied depending on the type of dataset considered. We list in Table 1 the methods applicable for each type of dataset.

The difference with time series is that the data streams are generated continuously and infinitely at a variable speed. Given the size of the data generated, data streams cannot be exhaustively stored for future exploitation. Therefore, methods for detecting anomalies in data streams must be applied in real time often without any a priori knowledge of the data distribution [5, 6]. Online processing requires a low-complexity detection algorithm for execution faster than the data arrival speed and often also low memory consumption if the solution is implemented on resource-constrained equipment.

Existing anomaly detection techniques are based on two important properties of anomalies: they have a very different behavior than others and they are rare [7-10]. Among these techniques, we distinguish:

Table 1. Methods applicable for each type of data set

Data sets	Applicable methods
Data flow	SmartSifter, AnyOut, CluStream, LEAP, MiLOF, DCLUST, IncLOF, DenStream, Abstract-C, AMCOD, HPStream, WaveCluster, MDEF.
Time series	ARMA, ARIMA, VARMA, CUSUM, EWMA, LSA, MLP, ART, NN, AE, GAN.
Graphs	DBMM, ECOutlier, NetSpot, Parcube, Com2, NetSmile, DeltaCon.
High dimension	GLOF, HighDOD, SOF, CLIQUE, HPStream, ABOD, SOD, SPOT.
Sequential	CLUSEQ, TARAZAN.
Spatio-temporal	Outstretch, TRAOD, LSTM, CNN.
Spatial data	Moran scatterplot, DBSCAN.

Statistical techniques can be parametric or non-parametric. Unlike the non-parametric approach, the parametric approach assumes an a priori knowledge of the data distribution [11]. statistical methods build a model with a confidence interval from the existing data. New data that do not fit this model will be considered abnormal [1].

Proximity-based techniques, which group together those based on nearest neighbors and those based on clustering. The nearest neighbor techniques determine for a given techniques determine for an observation on O its k nearest neighbors through the calculation of the distance between all observations in the dataset. There are two approaches of methods based on the nearzst neighbor: the distance-based approach [12, 13], and the density-based approach [14]. The main objective of clustering techniques is to divide the dataset into clusters containing data that have similar behaviors. There are two approaches in these techniques: the distance-based approach, where the farthest cluster represents an anomaly and the density-based approach defines the anomaly by the cluster that contains the least amount of data [15, 16].

Deep learning based techniques which represent a class of supervised or unsupervised machine learning algorithms based on the use of multiple layers of non-linear processing units. Among these methods are auto-encoders (AE) and One-Class Neural Networks (OCNN) [17].

- Other techniques exist such as those based on support vector machines, neural networks, methods adapted to high dimensions by subspace construction or dimension reduction [18-20].

In Figure 2, we present a summary of this classification with examples of algorithms belonging to each of these categories.

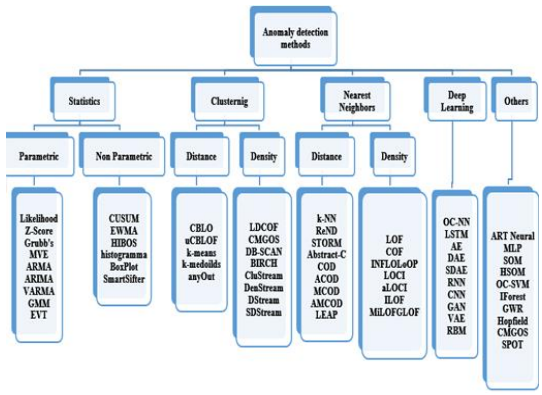

Figure 2. Classification of different anomaly detection techniques

3. FAULT DETECTION SYSTEM

3.1 Statistical indicators

3.1.1 Process capability

The concept of capability is certainly the most widely used in production workshops. Capability is measured by the ratio between the required performance and the actual performance of a process and is expressed by a number.

It measures the ability of a process to achieve a characteristic within the tolerance interval set by the specifications. The fact of using a number to characterize capability is fundamental. A number is an objective, it is not subject to interpretation.

We will dissociate two types of capability indicators:

1. Short-term indicators reflect dispersion over a very short time. We will then talk about process capability (Cp).
2. Long-term indicators that reflect the reality of the products delivered. We will then talk about the process potential (Pp).

This indicator is calculated as follows:

$$\mathrm{Cp} = \frac{(USL - LSL)}{6\sigma} \tag{1}$$

$$Pp = \frac{(USL - LSL)}{6s} \tag{2}$$

where:

- Cp: Process capability.
- Pp: process potential.
- USL: Upper Specification limit.
- LSL: Lower Specification limit.
- σ: population standard deviations.
- s: the mean of the standard deviations of the samples.

The relationship between process variability and tolerances can be formalized by considering the standard deviation σ of the process.

In a first approach, a process will be said to be capable if: the distance between the upper tolerance (USL) and the lower tolerance (LSL) is greater than or equal to 6σ, i.e. The Pp or the Cp must be greater than or equal to 1.33 (8σ / 6σ).

The relationship between 2T ($USL - LSL$) and 6σ results in three levels of process accuracy (Figure 3 shows the three levels of process accuracy):

1. 2T >> 6σ: High relative accuracy, where the tolerance band is found to be much higher than 6σ.
2. 2T > 6: Medium relative accuracy, where the tolerance band is found to be just above 6σ.
3. 2T < 6σ: Low relative accuracy, where the tolerance band is less than 6σ.

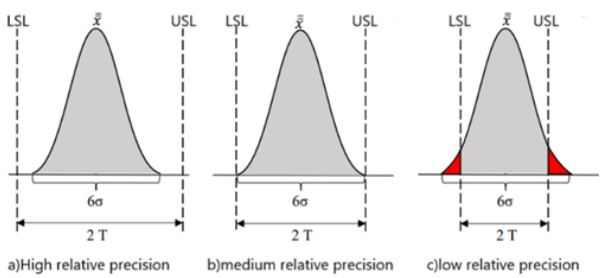

Figure 3. Process decision levels.

Process capability can be interpreted in a more general way:

- Pp ≈ Cp: The potential of the process is fully exploited.
- Pp > Cp: The potential of the process is not fully exploited [21-23].

In the case where there is only one tolerance limit in the specifications or standards, we consider Cpl and Cpu (Figure 4) as an indicator of process capability compared to the limit, these refer to the difference between the process mean and at the upper and lower specification limits (respectively), at 3σ (half of the total process variation):

$$Cpu = \frac{USL - \bar{x}}{3\sigma} \tag{3}$$

$$Cpl = \frac{\bar{x} - LSL}{3\sigma} \tag{4}$$

With:

- Cpu: process capability compared to the upper limit.
- Cpl: process capability compared to the lower limit.
- $\bar{x}$: the average.

The relationship between T ($(USL - \bar{x}) or (\bar{x} - LSL)$) and 3σ results in three levels of process accuracy:

1. T >> 3σ: High relative accuracy, where the tolerance band is found to be much higher than 3σ.
2. T > 3σ: Medium relative accuracy, where the tolerance band is found to be just above 3σ.
3. T < 3σ: Low relative accuracy, where the tolerance band is less than 3σ [21-23].

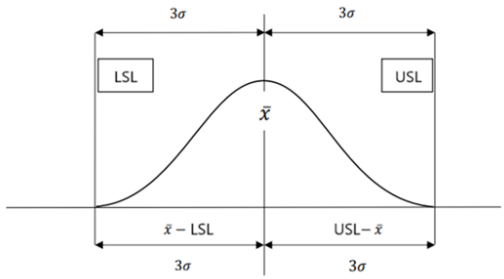

Figure 4. Cpu and Cpl indices

3.1.2 Machine setting status

It is possible to consider a relatively wide tolerance band with small process variation, but in which a significant proportion of the process is outside the tolerance band (Figure 5). This suggests the need for an index that accounts for both process variation and centering. Such an index is Cpk, which is widely accepted as a means of communicating the tuning status of the process [22, 23].

Cpk is an indicator of process tuning those measures decentering from the mean.

The coefficient Cpk has the following formula:

$$Cpk = Cp(1 - k) \tag{5}$$

Such as:

$$k = \frac{2|M - \bar{x}|}{USL - LSL} \tag{6}$$

$$M = \frac{USL + LSL}{2} \quad (7)$$

With:

- Cpk: Index of the state of adjustment of the process.
- k: Deviation coefficient.
- M: the middle of the tolerance limits.

Following the industrial engineering and computing office, the Cpk provides us with the following indications:

- If Cpk ≈ Cp: Well-tuned process.
- If Cpk < Cp: Process badly regulated.

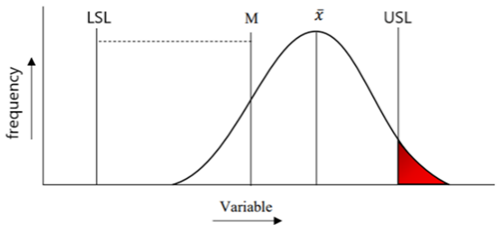

Figure 5. The capable process not centred

3.1.3 Kurtosis

The wear evolution of the machine can be measured by the kurtosis coefficient. Kurtosis (from the Greek kurtos meaning curve or rounding) is a descriptive statistic (centred moment of order 4) measuring the flatness of the distribution or what is still called its degree of curvature or sometimes its "kurtosis". Figure 6 shows examples of distributions with three different degrees of curvature.

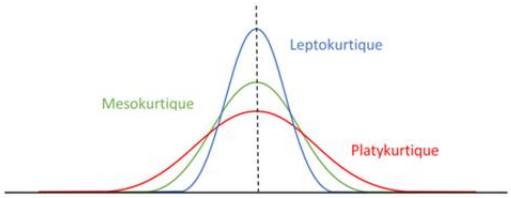

Figure 6. Examples of distributions with three different degrees of curvature

In practice, a corrected coefficient Krt (normalized kurtosis) is most often used. The value of this coefficient is then zero for a normal distribution (curve then called mesokurtic). A negative kurtosis coefficient indicates a rather flattened distribution (platykurtic) and a positive kurtosis coefficient, a "pointed" distribution (leptokurtic).

The mathematical formula of this kurtosis coefficient is given by:

$$Krt = \frac{M_4}{(M_2)^2} - 3 \quad (8)$$

$$M_2 = \frac{\sum(x_i - \bar{x})^2}{n} \quad (9)$$

$$M_4 = \frac{\sum(x_i - \bar{x})^4}{n} \quad (10)$$

With:

- Krt: Coefficient of flattening "Kurtosis".
- M_2: Moment of order 2.
- M_4: Moment of order 4.
- x_i: element of the population.
- n: population size.

According to the Bureau of Engineering Industrial and Informatics, the flattening coefficient (Kurtosis) is interpreted as follows:

- Krt<0: wear mark of the machine.
- Krt>0: no wear of the machine [22, 23].

3.1.4 Skewness

Skewness is a coefficient that allows us to evaluate the asymmetry of a distribution that reflects the saturation of the machine. A statistical distribution is symmetrical if the observations identified by their frequencies are equally dispersed on either side of a central value. The skewness coefficient corresponds to the third-order moment of the reduced centred variable. Figure 7 shows examples of asymmetric curves.

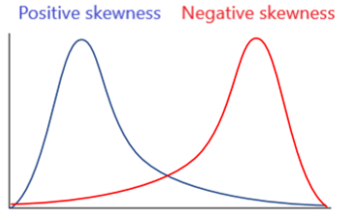

Figure 7. Examples of asymmetric curves

The mathematical formula for this asymmetry coefficient is given by:

$$Skw = \frac{M_3}{(M_2)^{1.5}} \quad (11)$$

$$M_3 = \frac{\sum(x_i - \bar{x})^3}{n} \quad (12)$$

where:

- Skw: Skewness asymmetry coefficient.
- M_3: Moment of order 3.

In general, the value of this coefficient is 0, for the scores observed in a cognitive test, a positive coefficient of asymmetry is related to a floor effect (a difficult task) and a negative coefficient of asymmetry is related to a ceiling effect (task too easy).

- Skw < 0: saturation in the upper part of the average.
- Skw > 0: saturation in the lower part of the average.
- Skw ≈ 0: no saturation [22, 23].

3.2 Anomaly detection algorithms

3.2.1 Kernel connectivity-based outlier factor algorithm

The kernel connectivity-based outlier factor (COF) algorithm consists of recording each object the degree of being an outlier, which is named the connectivity-based outlier factor. In order to formulate the kernel COF algorithm, we have to follow this step:

Using the kernel function, reestablish the initial data in to new features, and then apply this step to it.

Identifier the k nearest neighbors for each object x, $N_k(x)$ will be the name given to the Set that represent point x and those neighbors.

Establish a data set based on the nearest trail (SBN) from

data point x, in a way that all $1 \le i \le k-1$, x_{i+1} is the nearest neighbor point of set set $\{x_1, \ldots, x_i\}$ in set $\{x_{i+1}, \ldots, x_k\}$.

Let $e = \{e_1, \ldots, e_k\}$, be a sequence of edge points related to the *SBN* path, that composes the consecutive nearest neighbours from point x in set $N_k(x)$. Each e_i is an edge point and $dist(e_i)$ means the distance between sets comprising an edge.

Compute the average chaining distance from x to $N_k(x) - \{x\}$, represented by $dist_N k\ (x)$ and defined as:

$$dist(x) = \sum_{i=1}^{k} \frac{2(k+1-i)}{k(k+1)} dist(ei) \tag{13}$$

With:

- $dist(x)$: the average chaining distance from x to $N_k(x) - \{x\}$ with $N_k(x)$ the Set that represent point x and those neighbors.
- k: the nearest neighbors for each object x.
- e: the sequence of edge points related to the *SBN* path.

$dist_N k\ (x)$ can be seen as the biggest distance in the cost description for the SBN path from point x.

Calculate the connectivity-based outlier factor (COF) at the data point x by its k–th neighbor with this equation:

$$COF(x) = \frac{dist(ei)}{\frac{1}{k}\sum_{o \in Nk(x)} dist(o)} \tag{14}$$

With:

- COF(x): the proportion of the average distance from x to $N_k(x)$ and the average distance of its neighbor records.

Which means the more the COF increases the more likely that the chance of object being is an outlier will increase [24].

3.2.2 Isolation Forest algorithm

The Isolation Forest algorithm designee is based on two the anomaly data features: a. The exception data represent a smaller ratio of the global size of the data set. b. There is an important difference between the attribute value of the normal data and the abnormal data. In a training set containing just numeric types, the data is divided recursively until iTree can tell the difference between data. Since they are highly sensitive to segregation, the normal data is found away from the root node of the tree, and the anomaly data is found nearer the root node, so with a small number of characteristic conditions, the anomaly data can be detected.

The main of the Isolation Forest algorithm is to establish a forest (iForest) composed of iTree. To make the calculation and the description easy, the Isolation Forest algorithm introduces a definition of the length of the isolation tree.

Definition: Isolation Tree. T is a node of an isolation tree. T can be either an external-node with no child or it can be an internal-node with one test and exactly two daughter nodes (Tl, Tr). Split the data points into Tl and Tr in function of the size of the split values and the property values. To build iTree, chose at random an attribute A and A split value P from the data set $D = \{d_1, d_2, \ldots, d_n\}$ and then divide each data object di by the value of its attribute A (called $d_i(A)$). If $d_i(A) < p$, then leave it in the left subtree and vice versa. In this manner, the right and left subtrees are built iteratively until one of the conditions is satisfied: a. there is only one data or several identical data in D; b. the tree reaches its maximum height.

Definition: Path Length h(d) of a point d is measured by the number of edges d traverses an iTree from the root node until the traversal is ended at an external node. Given the data set D, [12] gives the path length of the failed query in the binary search tree:

$$C(n) = 2H(n-1) - (2(n-1)/n) \tag{15}$$

where:

- C(n) is the average of h(d) given n, we use it to normalize h(d).
- H(i) is the harmonic number.
- n is the number of leaves.

H(i) can be calculated by ln(i) + 0.5772156649 (Euler's constant). The anomaly score s of an instance d is defined as:

$$S(d,n) = 2^{\frac{-E(h(d))}{c(n)}} \tag{16}$$

where:

- E(h(d)) is the average of h(d) from a collection of isolation trees.

The Isolation Forest algorithm generates a precise number of iTree and forms iForest. More precisely, the subsets of D are randomly sampled to construct each iTree in order to guarantee the diversity of the iTree. Process the data d by traversing the iTree collection in the iForest to identify the leaf node of d. Then, based on the length of its path, the abnormal fraction of d is computed and the abnormal evaluation of d is performed.

When $E(h(d)) \to C(n)$, $S \to 0.5$, that is, when all the data is returned to S of 0.5, there is no obvious abnormal value in all samples; When $E(h(d)) \to n-1$, $S \to 0$, that is, when the S of the data is much less than 0.5, then they have a big potential to be evaluated as normal. When $E(h(d)) \to 0$, $S \to 1$, that is, when the S of the data return is very close to 1, then they are outliers [25-27].

4. DESCRIPTION OF THE PROCESS

Neutralization is affected in a neutralization reactor, which is composed of a U-tube and a separator. Nitric acid (57% by weight) with a little sulfuric acid and ammonia gas are introduced into an ammonium nitrate solution circulating in the U-tube. The ammonia gas is fed by a distributor placed at the bottom of the mixing tube, the nitric acid is fed just under the mixing tube. Recycled ammonia from the ammonia recovery system and ammonium nitrate solution from the dissolving tank are injected into the return tube. The reaction of ammonia and nitric acid takes place in the mixing tube, causing the ammonium nitrate solution to flow. The neutralization reactor operates at a temperature of about 180°C and a pressure of 3 bar effective. The heat released by the reaction in the reactor is sufficient to increase the concentration of the formed ammonium nitrate solution to about 77%, by evaporation of the water.

After the reaction, the solution reaches the top of the mixing tube and flows into the separator, where the (superheated) steam and the ammonium nitrate solution are separated.

The emphasis in this study was on the first block which is the reactor area and what preceded it, and this was due to the

problems that arose in the parameters included in this block which are (Figure 8 and 9 explain the link detected in this process):

- the NH3 Neutralizer flow rate.
- the HNO3 Neutralizer flow rate.
- the Ammonia temperature.
- the Neutralizer temperature.
- the Ammonia Pressure.
- the neutralizer pressure.

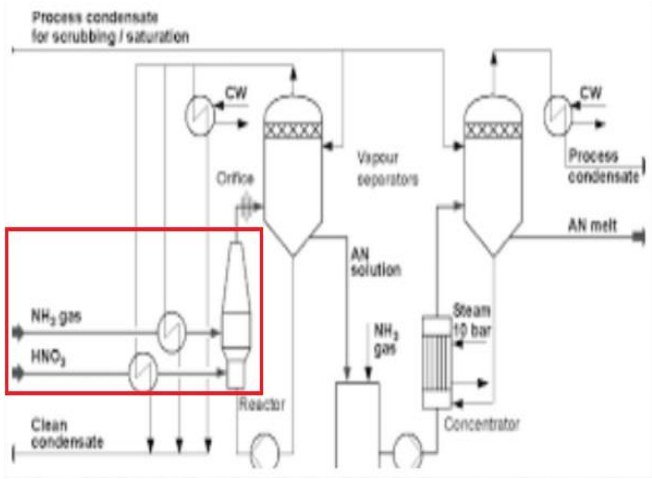

Figure 8. Neutralization scheme

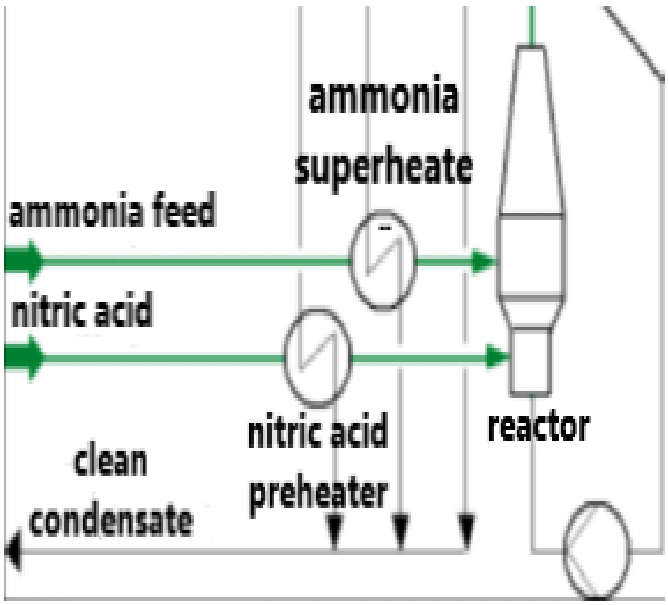

Figure 9. The first block of neutralisation process

5. RESULTS AND DISCUSSION

5.1 General diagnosis

We recovered the data from the daily inspection sheets of the company, and these inspection data were performed by the controllers when the production was measured for 6 months, and from the statistical parameters used in the company, we built a structure file of 1320 raw data. Then, we analyzed this file using our model. The general diagnosis of the process gave us the following results:

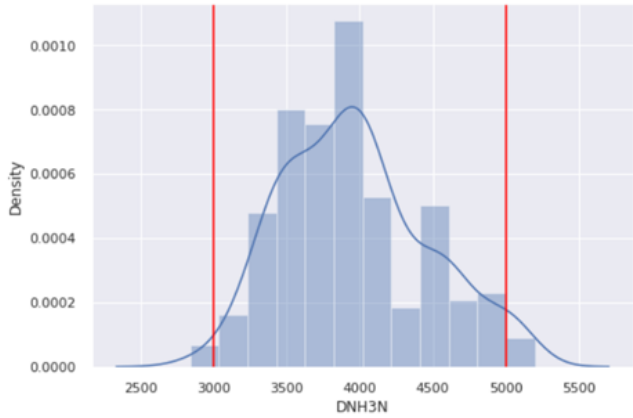

Figure 10. Diagnosis of the NH3 Neutralizer flow rate

Table 2. Statistical indicators <<NH3 Neutralizer flow rate>>

CP	Pp	Cpk	Krt	Skw
0.67	1.04	0.65	0.4	0.41

In the diagnosis of the NH3 Neutralizer flow rate, it is observed that the histogram and the curve of gausse are almost totally within the tolerance range (Figure 10). Although the process is not capable to produce a quality required by the standards, since the Cp is equal to 0.67, which is less than 1, but it is well regulated because the Cpk that shows a value of 0.65 is almost equal to Cp. In addition, it has a potential of 1.04, so we can make the process capable to produce the quality required without investment (Table 2).

Regarding the state of the process, there is no form of wear, but there is a very slight saturation at the lower tolerance limit.

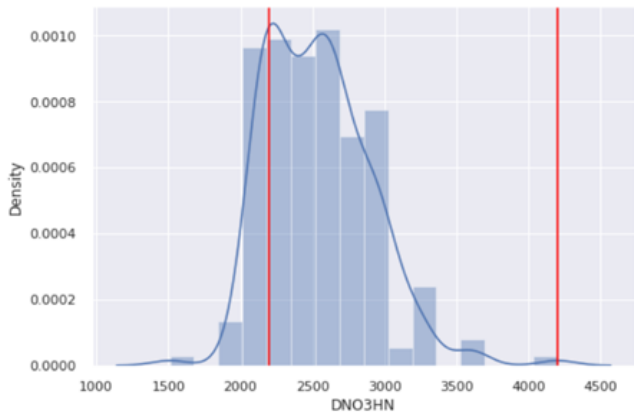

Figure 11. Diagnosis of the HNO3 Neutralizer flow rate

Table 3. Statistical indicators <<HNO3 Neutralizer flow rate>>

CP	Pp	Cpk	Krt	Skw
0.92	1.74	0.88	-1.53	0.81

In the diagnosis of the flow of the HNO3 Neutralizer flow rate, it is observed that the majority of the histogram and the curve of gausse are in the tolerance range (Figure 11). Although the process is not capable to produce a quality required by the standards, since the Cp is equal to 0.92, which is less than 1, but it is well regulated because the Cpk that shows a value of 0.88 is almost equal to Cp. In addition, En

has for this process, a potential of 1.74, so we can make the process capable to produce the quality required without investment (Table 3).

Regarding the state of the process, there are very slight forms of wear and saturation at the lower tolerance limit.

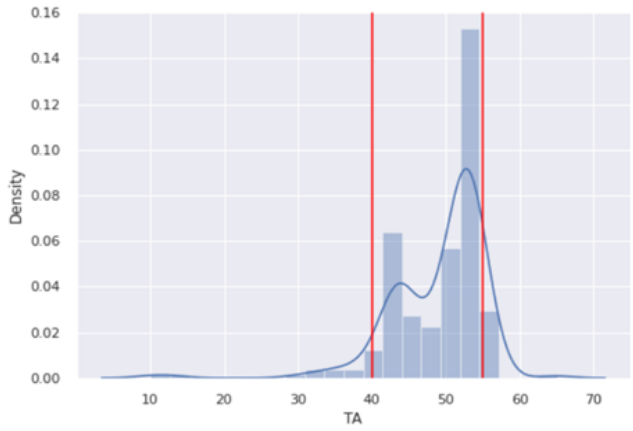

Figure 12. Diagnosis of the Ammonia temperature

Table 4. Statistical indicators <<Ammonia temperature>>

CP	Pp	Cpk	Krt	Skw
51.85	123.96	51.75	-9.25	-2.2

In the diagnosis of the ammonia temperature, it is observed that the majority of the histogram and the curve of gausse are in the tolerance range (Figure 12). We see that the process is very capable of producing the quality required by the standards, since the Cp is equal to 51.85, which is greater than 1.33, in addition it is well regulated because the Cpk that shows a value of 51.75 is almost equal to Cp. Moreover, we have for this process, a potential of 123.96, but we do not need to increase these performances since the Cp is greater than 1.33 (Table 4).

Regarding the state of the process, In marks a form of wear and some form of saturation at the upper limit of tolerance.

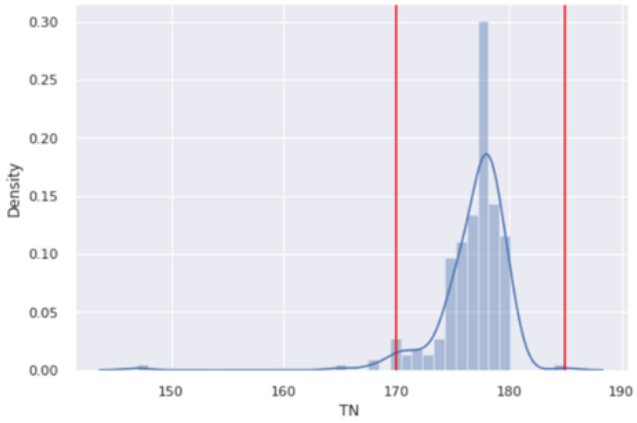

Figure 13. Diagnosis of the Neutralizer temperature

Table 5. Statistical indicators <<Neutralizer temperature>>

CP	Pp	Cpk	Krt	Skw
101.53	180.26	101.42	-30.26	-4.02

In the diagnosis of the neutralizer temperature, it is observed that the histogram and the curve of gausse are almost totally within the tolerance range (Figure 13). We see that the process is very capable to produce the quality required by the standards, since the Cp is equal to 101.53, which is greater than 1.33, in addition it is well regulated because the Cpk that shows a value of 101.42 is almost equal to Cp. Moreover, we have for this process, a potential of 180.26, but we do not need to increase these performances since the Cp is greater than 1.33 (Table 5).

Regarding the state of the process, In marks a form of wear and a form of saturation at the upper limit of the tolerance.

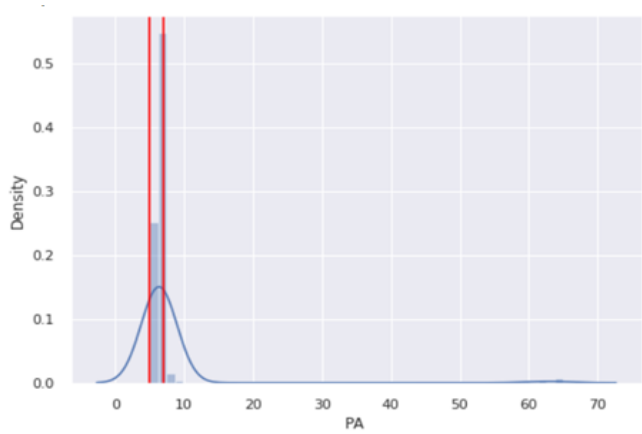

Figure 14. Diagnosis of the Ammonia Pressure

Table 6. Statistical indicators <<Ammonia Pressure>>

CP	Pp	Cpk	Krt	Skw
44.33	141.53	44.28	-50.35	7.22

In the diagnosis of the ammonia pressure, it is observed that the histogram and the curve of gausse are almost totally within the tolerance range (Figure 14). We see that the process is very capable to produce a quality required by the standards, since the Cp is equal to 44.33, which is greater than 1.33, in addition it is well regulated because the Cpk that shows a value of 44.28 is almost equal to Cp. Moreover, we have for this process, a potential of 141.53, but we do not need to increase these performances since the Cp is greater than 1.33 (Table 6).

Regarding the state of the process, a form of wear and a form of saturation at the lower limit of the tolerance are marked.

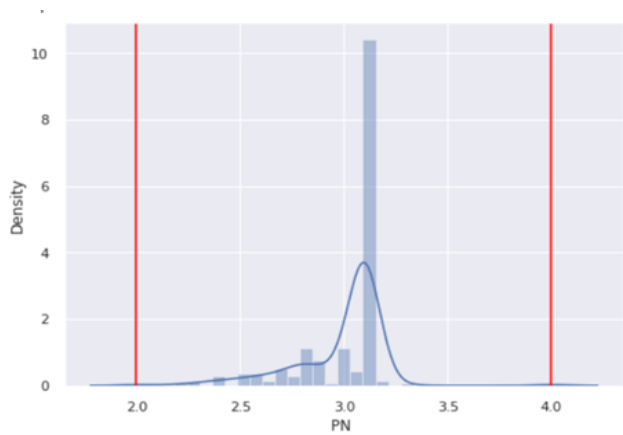

Figure 15. Diagnosis of the neutralizer pressure

Table 7. Statistical indicators <<Neutralizer pressure>>

CP	Pp	Cpk	Krt	Skw
1527.49	3043.67	1527.32	-4.77	-1.32

In the diagnosis of the neutralizer pressure, we observed that the histogram and the curve of gausse are totally within the tolerance range (Figure 15). We see that the process is very capable of producing the quality required by the standards, since the Cp is equal to 1527.49, which is greater than 1.33, in addition, it is well regulated because the Cpk that shows a value of 1527.32 is almost equal to Cp. Moreover, we have for this process, a potential of 3043.67, but we do not need to increase these performances since the Cp is greater than 1.33 (Table 7).

Regarding the state of the process, a form of wear and a slight form of saturation at the upper limit of the tolerance are marked.

5.2 Anomaly detection

After the general diagnosis of the process, it became clear to us that the problem of the process lies in the flows of the neutralizer. That is why we applied two commonly used algorithms to this type of data, in order to choose the best algorithm to monitor and detect this problem. After applying the algorithms, we obtain the following results:

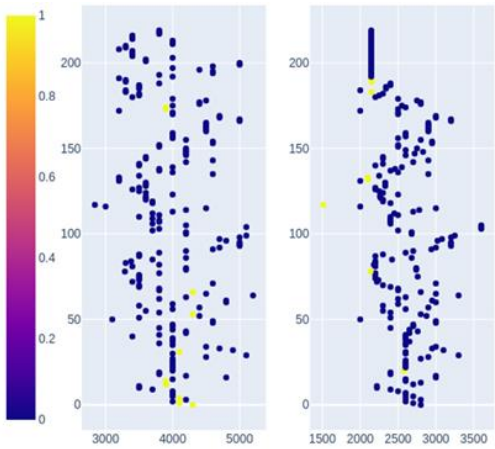

Figure 16. Connectivity-based local outlier

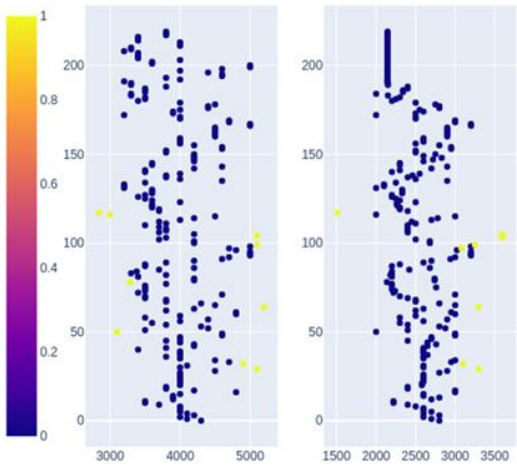

Figure 17. Isolation Forest

Table 8. The accuracy of the algorithms

	NH3 NFR	HNO3 NFR
Connectivity-Based Local Outlier	99%	91%
Isolation Forest	82%	82%

Figures 16 and 17 presents the results obtained using the Connectivity-Based Local Outlier and isolation forest methods. The input parameters are the flow rates of NH3 neutralizer and HNO3 neutralizer on the X-axis, and the amount of data on the Y-axis. The output is the probability of anomaly. The ideal values of the hyperparameters for the cof are (k = 100) and for forest isolation are (n_sample = 0.1, r = 30, n_dim = 1 and α = 0.01). For these parameters the tolerance intervals are 3000 to 5000 and 2200 to 4200 respectively, we see that the dispersion of HNO3 is more centred than NH3 and it can be seen that the values which are out of tolerance are rare. This leads to the explanation that the values out-of-tolerance are not the ones making the parameters incapable and it is clearly seen in the cof technique unlike the isolation forest method, this explanation is also confirmed by the workers of this company.

Table 8 presents the values obtained from both methods. According to the Table 8, we observe that the COF method is more efficient than the isolation forest. This observation leads us to choose the method of COF to follow this process in order to avoid the problems generated in these parameters.

6. CONCLUSIONS

In this article, we have made a complete diagnosis of the process of neutralization of ammonia by nitric acid. To do this, we relied both on the SPC approach and techniques for detecting anomalies. In the first step, we defined the statistical indicators that they used in the general diagnosis, the second step, we define the process on which we work. Then we developed our diagnostic model which consists of two parts:

- The first part: works on a general diagnosis of the system and the discovery of the parameters that suffer from problems affecting the quality of the product
- The second part: works on the application of techniques of machine learning to the parameters that are not capable for detected the anomaly, comparing the results and choosing the best algorithm to monitor and solve the problems affecting the final product.

At the end of our study, we reached the results that the parameters: the ammonia temperature, the neutralizer temperature, the ammonia pressure and the neutralizer pressure are in good working order, on the other hand, we have the parameters NH3 neutralizer flow rate and HNO3 neutralizer flow rate have a capacity to be improved and developed without investment in order to produce the quality required by the standards

we propose the implementation of the Statistical method

Process Control within the company and also the use of the Connectivity-Based Local Outlier algorithm in order to avoid all the disruptions that we have seen during our research.

Nevertheless, we hope to have other opportunities for putting this system into practice in another process, whether in the same processing sector or in other industrial sectors and this within the framework of applied research.

REFERENCES

[1] Cui, Y., Liu, Z., Lian, S. (2022). A survey on unsupervised industrial anomaly detection algorithms. arXiv preprint arXiv:2204.11161. https://doi.org/10.48550/arXiv.2204.11161

[2] Gupta, M., Gao, J., Aggarwal, C., Han, J. (2014). Outlier

detection for temporal data. Synthesis Lectures on Data Mining and Knowledge Discovery, 5(1): 1-129. https://doi.org/10.2200/S00573ED1V01Y201403DMK008

[3] Thakkar, A., Lohiya, R. (2020). A review of the advancement in intrusion detection datasets. Procedia Computer Science, 167: 636-645. https://doi.org/10.1016/j.procs.2020.03.330

[4] Pourhabibi, T., Ong, K.L., Kam, B.H., Boo, Y.L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decision Support Systems, 133: 113303. https://doi.org/10.1016/j.dss.2020.113303

[5] Tellis, V.M., D'Souza, D.J. (2018). Detecting anomalies in data stream using efficient techniques: A review. In 2018 International Conference on Control, Power, Communication and Computing Technologies (ICCPCCT), Kannur, India, pp. 296-298. https://doi.org/10.1109/ICCPCCT.2018.8574310

[6] Salehi, M., Rashidi, L. (2018). A survey on anomaly detection in evolving data: [with application to forest fire risk prediction]. ACM SIGKDD Explorations Newsletter, 20(1): 13-23. https://doi.org/10.1145/3229329.3229332

[7] Chalapathy, R., Chawla, S. (2019). Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407. https://doi.org/10.48550/arXiv.1901.03407

[8] El Sibai, R. (2018). Sampling, qualification and analysis of data streams. Doctoral dissertation, Sorbonne Université; Université Libanaise.

[9] Chabchoub, Y., Sibai, R.E., Fricker, C. (2019). Bike sharing systems: a new incentive rebalancing method based on spatial outliers detection. International Journal of Space-Based and Situated Computing, 9(2): 99-108. https://doi.org/10.1504/ijssc.2019.104220

[10] Gupta, S., Modgil, S., Gunasekaran, A. (2020). Big data in lean six sigma: A review and further research directions. International Journal of Production Research, 58(3): 947-969. https://doi.org/10.1080/00207543.2019.1598599

[11] Agarwal, S., Suchithra, A.S., Singh, S.P. (2021). Analysis and interpretation of rainfall trend using Mann-Kendall's and Sen's slope method. Indian J. Ecol, 48: 453-457.

[12] Marcos, H.F., Cejudo, A., Martinez-Romo, J., Pérez, A., Araujo, L., Lebea, N., Oronoz, M., Casillas, A. (2022). Approximate nearest neighbour extraction techniques and neural networks for suicide risk prediction in the CLPsych 2022 shared task. In Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, Seattle, USA, pp. 199-204. https://doi.org/10.18653/v1/2022.clpsych-1.17

[13] Beck, G., Duong, T., Lebbah, M., Azzag, H., Cérin, C. (2019). A distributed approximate nearest neighbors algorithm for efficient large scale mean shift clustering. Journal of Parallel and Distributed Computing, 134: 128-139. https://doi.org/10.1016/j.jpdc.2019.07.015

[14] Ning, J., Chen, L., Chen, J. (2018). Relative density-based outlier detection algorithm. In Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence, Shenzhen, China, pp. 227-231. https://doi.org/10.1145/3297156.3297236

[15] Ghiassi, M., Saidane, H., Oswal, R. (2021). YAC2: An α-proximity based clustering algorithm. Expert Systems with Applications, 167: 114138. https://doi.org/10.1016/j.eswa.2020.114138

[16] Bansal, M., Sharma, D. (2021). A novel multi-view clustering approach via proximity-based factorization targeting structural maintenance and sparsity challenges for text and image categorization. Information Processing & Management, 58(4): 102546. https://doi.org/10.1016/j.ipm.2021.102546

[17] Abraham, A., Dutta, P., Mandal, J.K., Bhattacharya, A., Dutta, S. (2018). Emerging technologies in data mining and information security. Proceedings of IEMIS-2018. https://doi.org/10.1007/978-981-13-1498-8

[18] Basysyar, F.M. and Dwilestari, G. (2022) House price prediction using exploratory data analysis and machine learning with feature selection. 11-21.

[19] Wawage, P. and Deshpande, Y. (2022) Real-time prediction of car driver's emotions using facial expression with a convolutional neural network-based intelligent system. International Journal of Performability Engineering, 18: 791–7. https://doi.org/10.23940/ijpe.22.11.p4.791797

[20] Xue, L., Hou, Y., Wang, S., Luo, C., Xia, Z. and Qin, G. (2022) A dual-selective channel attention network for osteoporosis prediction in computed tomography images of lumbar spine. 30-39.

[21] Dubey, R., Gunasekaran, A., Childe, S.J., Fosso Wamba, S., Roubaud, D., Foropon, C. (2021). Empirical investigation of data analytics capability and organizational flexibility as complements to supply chain resilience. International Journal of Production Research, 59(1): 110-128. https://doi.org/10.1080/00207543.2019.1582820

[22] Oakland, J.S. (2007). Statistical Process Control. Routledge. https://doi.org/10.4324/9780080551739

[23] Burr, I.W. (2018) Statistical Quality Control Methods. Routledge. https://doi.org/10.1201/9780203738528

[24] Wang, Y., Li, K., Gan, S. (2018). A kernel connectivity-based outlier factor algorithm for rare data detection in a baking process. IFAC-PapersOnLine, 51(18): 297-302. https://doi.org/10.1016/j.ifacol.2018.09.316

[25] Xu, D., Wang, Y., Meng, Y., Zhang, Z. (2017). An improved data anomaly detection method based on isolation forest. In 2017 10th International Symposium on Computational Intelligence and Design (ISCID), 2: 287-291. https://doi.org/10.1109/ISCID.2017.202

[26] Sirisha, A., Chaitanya, K., Krishna, K.V.S.S.R. and Kanumalli, S.S. (2021) Intrusion detection models using supervised and unsupervised algorithms - A comparative estimation. International Journal of Safety and Security Engineering, 11: 51–8. https://doi.org/10.18280/ijsse.110106

[27] Radhakrishnan, M., Boruah, S. and Ramamurthy, K. (2022) EEG-Based Anamoly Detection for Autistic Kids-A Pilot Study. Traitement Du Signal, 39: 1005-1012. https://doi.org/10.18280/ts.390327