An Introduction to
Alarm Analysis and Design

Iman Izadi ∗ Sirish L. Shah ∗∗ David S. Shook ∗∗
Tongwen Chen ∗∗∗

∗ Matrikon Inc., #1800, 10405 Jasper Ave. Edmonton, Alberta, Canada, T5J 3N4
(e-mail: iman.izadi@matrikon.com)

∗∗ Department of Chemical & Materials Engineering, University of Alberta, Edmonton, Alberta, Canada, T6G 2G6
(e-mail: sirish.shah@ualberta.ca, dave.shook@ualberta.ca)

∗∗∗ Department of Electrical & Computer Engineering, University of Alberta, Edmonton, Alberta, Canada, T6G 2V4
(e-mail: tchen@ece.ualberta.ca)

Abstract: Alarms are essential in every process, system and industrial complex. They are configured to notify the operators about any abnormal situation in the process. In practice, operators receive far more false and nuisance alarms than valid and useful alarms. In this paper, an overview on alarm analysis and design is given. Some of the reasons for false and nuisance alarms are discussed and a few solutions to reduce them are studied. False alarm rate, missed alarm rate and detection delay trade-offs in alarm design are also discussed.

Keywords: Alarm systems; Alarm management; Fault detection; Safety analysis.

1. INTRODUCTION

Large industrial plants contain thousands of sensors and actuators, wired or wireless communication networks with thousands of clients and hundreds or thousands of control loops. All the components of a plant are susceptible to faults which disrupt normal operation of the control system and may result in unsatisfactory performance, instability, failure or even dangerous situations. Due to the increasing complexity of process control systems and the growing demands for quality, cost efficiency and safety, it is important that faults be promptly diagnosed and appropriate remedies be applied. Fault detection has been a very active research area in both academia and industry in the past few decades. A variety of fault detection methods have been developed and are available in the literature [2, 4, 6].

When a fault is detected an alarm is raised (in auditory or visual form) to inform the operator or other responsible individuals about the abnormality in the plant. Alarms are prioritized depending on the impact of the abnormality on the operation and its severity. In the ideal case, for each abnormal situation one and only one alarm should be triggered. This is hardly the case in practice. Operators, almost always, receive far more alarms than they require or can handle. These alarms are either false (an alarm that is raised but nothing is abnormal within the plant, also known as false positive or type I error) or nuisance (an alarm that is true but redundant, for other alarms have already informed the operator about the abnormality). Too many false and nuisance alarms (alarm flooding) distract the operator from operating the plant to the extent that critical alarms may be ignored. This situation can lead to the point that the operator no longer trusts the alarms (“cry wolf” effect) or even shuts down the whole monitoring system. To rectify this situation, there has recently been an increasing interest in industry to address this issue and seek remedies to reduce the number of false and nuisance alarms.

In this paper, an overview of alarm analysis and design will be given. We will first investigate the causes of false and nuisance alarms. Then we discuss how the situation can be improved. Multivariate data analysis, model-based performance monitoring, threshold design and processing of the alarm variable will be discussed as methods that can reduce the number of false and nuisance alarms. We will then discuss the two important trade-offs in alarm design and present a simple case study.

2. WHY PLANTS ARE MONITORED?

Industries reportedly lose billions of dollars every year due to plant problems. Plant problems include a wide range of issues, from minor equipment damage, process abnormalities and bad controller tunings to major plant incidents. Minor problems in the plant will affect the quality or quantity of the final product which will in turn result in lower profit margin. In a worse situation, problems might lead to
unplanned shut-down of the plant. Unplanned shut-downs, although might be triggered by a minor problem such as malfunction or failure of an inexpensive equipment, are one of the most serious situations in plants and will result in major production loss and two to three fold repair costs. To improve the efficiency, reliability and quality and to avoid equipment damage, production loss and unplanned down-time, it is now a routine practice to monitor such processes.

Another reason for effective plant monitoring, which is gaining more and more attention, is environmental issues. With growing concerns about water/air/earth contaminations, environment protection and global warming, industries are urged to monitor their processes and products more aggressively. Governments also impose substantial fines on violations of environmental protection regulations, which will also result in reduced profit margin.

The other and probably the most important reason for plant monitoring is safety. According to the Abnormal Situation Management (ASM) Consortium, petrochemical plants on average suffer a major incident every three years [1]. Besides the average $80 million cost per significant outage, these incidents usually cause human casualties and serious injuries. Many can recall these catastrophic industrial incidents, a few of which are:

- Chernobyl, Ukraine, 1986 (more than 4000 direct and indirect deaths)
- Piper Alpha Oil Rig, North Sea, 1988 (167 deaths)
- Phillips 66 Complex, Texas, 1989 (23 deaths)
- BP Refinery, Texas City, 2005 (15 deaths)

Surprisingly, these incidents occur not usually because of major design flaws or equipment malfunctions, but rather simple mistakes. The Phillips 66 explosion was caused by a check valve that was connected the wrong way, only because the open and close ends of the check valve looked similar. In the Piper Alpha oil rig disaster, the automatic fire-fighting system did not operate because it had been switched to manual. It is important to remember that, although these disasters may have different causes, most of them could have been prevented if a reliable monitoring and alarm system had been in place. In some of these incidents, managers, operators and engineers were criminally charged and sentenced to jail terms. Also, after some of these incidents, governments introduced new safety regulations and laws. In most developed countries, health and safety negligence is a punishable crime.

In the United States, petrochemical industry alone loses between 10 to 20 billion dollars annually, due to abnormal situations. On top of that, industries lose another 10 billion dollars annually towards the cost of equipment repairs, environmental fines, compensations for human casualties and injuries, investigations, litigations, etc. Industries also claim over 2.2 billion dollars annually in equipment damage to insurance companies in the United States [1]. All these issues contribute to industries’ commitment to further improving their monitoring and alarm systems.

3. WHY FALSE ALARMS?

The outcome of a monitoring system is almost always an alarm. An alarm is a notification that an unexpected event has occurred. Traditionally an alarm is defined in terms of a single variable going outside some “normal” or “safe” operating range (Figure 1). The operating range could come either from instrument limits or a certain confidence range of the normal operating conditions. Alarm limits that are set close to instrument limits are used to protect the equipment and avoid hazardous situations. Alarm limits that are set at a certain confidence range (μ ± 3σ for instance) are meant to keep the operation at normal and often optimal conditions. Alarm limits are also known as control limits, check limits or trip points.

Traditionally, only a limited number of selected variables were monitored. These variables either reflect the quality of the process or are important in process operation and safety. Thus, monitoring them gives a good indication of the quality and normal operation of the plant. Statistical process control (SPC) and statistical quality control (SQC) charts (e.g., Shewhart and CUSUM charts) were often used for performance monitoring. SPC and SQC theories are well advanced and still used in some industries [9]. For the overall performance of the process, a few “quality variables” or “key performance indicators (KPI)” that are calculated from process variables are monitored.

Historically, alarms were expensive to add and difficult to implement. In the old days, each alarm had to be hardwired from the sensor to the control room. The space in control room was also very limited due to numerous input/output devices installed. These technical and physical limitations resulted in a limited number of carefully designed and reviewed alarms. The alarms were very reliable and operators could completely trust them.

Today, on the other hand, hardware and software advances have made it easy to add an alarm at practically no extra cost. With the distributed control systems (DCS) available on all modern plants, almost every process variable that can be measured is measured and stored in data historians. So all the variables are available for control and monitoring from the operator’s panel. Adding or configuring an alarm can be as easy as checking a box in the software or adding a few lines of code. What has happened as an outcome of this availability is a large increase in the number of poorly designed alarms. Moreover, usually after every plant incident a new set of alarms are added to the DCS without removing the old ones. As a result, a plant that in the old days had a dozen carefully configured alarms now often has hundreds of alarms. This has significantly reduced the quality and efficiency of alarm systems and
Table 1. EEMUA benchmark

<table>
<thead>
<tr>
<th></th>
<th>EEMUA</th>
<th>Oil and Gas</th>
<th>Petrochemical</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>average alarms per hour</td>
<td>≤ 6</td>
<td>36</td>
<td>54</td>
<td>48</td>
</tr>
<tr>
<td>average standing alarms</td>
<td>9</td>
<td>50</td>
<td>100</td>
<td>65</td>
</tr>
<tr>
<td>peak alarms per hour</td>
<td>60</td>
<td>1320</td>
<td>1080</td>
<td>2100</td>
</tr>
<tr>
<td>distribution % (low/med/high)</td>
<td>80/15/5</td>
<td>25/40/35</td>
<td>25/40/35</td>
<td>25/40/35</td>
</tr>
</tbody>
</table>

is the most significant contributing cause of false and nuisance alarms.

Assume that in a plant, each alarm is configured based on a confidence range and has a probability of false alarm $\alpha$. For each variable this probability is very small and thus the number of false alarms for each variable is negligible. However, if alarms are configured on $n$ process variables, the probability of false alarm for all the variables is $1 - (1 - \alpha)^n \approx n\alpha$. Therefore, the probability of false alarms grows approximately linearly with the number of monitored process variables. So if alarms are configured on hundreds or thousands of process variables (which is the case in practice), the overall probability of false alarm is not insignificant anymore. Although every individual alarm might be configured very well, the entire alarm system will be quite inefficient. On the other hand, because of interactions of process variables, their correlation is high. Thus, a single fault can cause many process variables to exceed their limits. Therefore, the operator is flooded with many alarms that are true but all of them are triggered by the same fault (nuisance alarms). Notice that, once an alarm is configured in the system, removing it is not easy and requires extensive paperwork and approval of several plant authorities.

There are other factors that contribute to false and nuisance alarms:

- poor design and poor tuning of alarms;
- different operating states of the plant;
- processes change over time;
- equipments changes degradation and wear;
- seasonal changes;
- plant-wide oscillations;
- noise and outliers.

After an alarm is raised operators usually intervene. The operator should first understand the alarm and acknowledge it. Then he/she has to detect the cause of the alarm, determine how significant it is and take necessary corrective actions to bring the operation back to normal and clear the alarm. There are several standards on alarm systems (e.g., EEMUA 191 [3] and ISA S18.02 [5]) that suggest guidelines on the maximum number of alarms an operator should receive per hour. According to the EEMUA guideline, which is widely accepted, an average operator takes about 10 minutes to process and properly respond to an alarm. Therefore, an operator should not receive more than six alarms per hour during the normal operation of the plant. As mentioned before, this is rarely the case in practice. Various studies show that the number of alarms each operator receives is far more than the standard, depending on the industry and their alarm generation policy.

Industry standards also have suggestions on average standing alarms at each time, peak alarms per hour during a plant upset and priority distribution (percentage of low, medium and high priority alarms). Table 1 compares the industry standards for alarm systems with typical surveyed values from different industry sectors.

4. HOW TO REDUCE FALSE ALARMS?

As discussed in previous sections, false and nuisance alarms are very common and reducing them is one of the biggest challenges industries face. There are many different contributing factors and as many (if not more) improvements that can be made. The problem and its remedies can be addressed at different levels in the plant, from management and engineering to operation and maintenance.

There are a variety of products and services on the market that are designed to help industries better design, manage and maintain their alarm systems. For instance, alarm rationalization is a costly and time consuming service that is intended to improve the quality of alarm system by carefully reviewing and reconfiguring all alarms in the plant. Many plants also have their alarm philosophy and an alarm management team that makes sure their alarm system is in compliance with their own philosophy and common alarm standards.

In the following sections we discuss a few techniques that can significantly reduce the number of false and nuisance alarms:

- multivariate process monitoring;
- model-based process monitoring;
- threshold design;
- data processing.

4.1 Univariate vs. Multivariate Monitoring

With the availability of DCS in almost every industrial plant, hundreds to thousands of variables are measured at relatively fast rates. This huge amount of data is used for control and monitoring of the plant. The data is also stored in data historians usually at a slower rate. Although many variables are measured in a plant (almost every variable that can be measured), they are not all independent. The statistical rank of data is usually very low, and adding more variable does not change this rank. Variables are correlated due to known or unknown relationships and also due to sensor redundancies in the plant. Because of
this correlation, when an abnormality occurs in the plant, it affects many process variables. Monitoring all these variables will result in many nuisance alarms. On the other hand, because of low signal-to-noise ratio, each variable carries very little information. To extract more information from variables, multivariate methods are required. Multivariate methods are based on the idea that the behavior of a process can usually be expressed by a few independent variables known as latent variables. Latent (or virtual) variables are obtained as a linear combination of raw process measurements. Several methods are available in the literature to obtain latent variables, including the popular principal component analysis (PCA) and partial least squares (PLS) methods. In multivariate monitoring techniques, instead of individual process variables, the latent variables are monitored, resulting in fewer alarms for the same abnormality [8].

Figure 2 illustrates a classic example of multivariate performance monitoring. Two variables A (horizontal plot, top left) and B (vertical plot, bottom) are plotted with their corresponding control limits. If these two variables are monitored individually, then the point indicated by  is outside the control limits for both variables. Therefore two alarms will be raised. But if variable A is plotted against variable B (top right), it will become clear that the two variables are correlated. The ellipse represents a confidence range of the joint distribution of the two variables, and the point indicated by  is indeed within the joint confidence range, so it is not an abnormal point. It can also be seen that the point indicated by  is outside the joint confidence range and thus is an abnormal data. However, since this point is within the control limits of both variables, univariate monitoring was not able to detect the abnormality.

4.2 Model-based Performance Monitoring

In model-based performance monitoring, instead of individual process variables, deviation of the process from its model is monitored. Figure 3 illustrates the block diagram of a model-based monitoring scheme. The actual behav-

Fig. 2. Multivariate vs. univariate monitoring

ior of the process is compared to the expected behavior predicted from some model of the process. The difference is called prediction error. If the model is perfect and no abnormalities exist in the process, the error is just noise. If the prediction error is not noise, then either the model is not good or there is an abnormality in the plant. By utilizing a well configured model, monitoring prediction error instead of individual variables will be more efficient and accurate in detecting abnormalities. One can even say individual variable monitoring is a special case of model-based monitoring, where the model is the normal operating point of the variable. The model has a central role here, so every effort put into obtaining a better model is worthwhile. The model can be a mathematical quantitative model or a heuristic qualitative model. There are several methods for modelling processes. The most reliable models are first principles models. First principles models are either based on laws of physics (Newton’s laws of motion, laws of thermodynamics, mass/energy balance, Maxwell’s laws of electromagnetic, etc.) or engineering concepts (Ohm’s law, pump curve, etc.). First principles models are very robust (laws of physics can not be violated) and deviation of the plant from the model indicates definite process problems. On the downside, obtaining these models needs a good understanding of the process (expert’s knowledge) which is not always available or even possible.

An alternative and more common approach is data-driven modelling. The availability of huge amounts of historical data through DCS systems particularly facilitates obtaining data-driven models. Data-driven models include dynamical models (estimated using system identification techniques) or statistical models (PCA, PLS, time series models, etc.). First principles and data-driven modelling techniques can also be combined as historical data can be used to estimate parameters of a first principles model. Different parts of a plant might also be modelled in different ways. Qualitative models (fuzzy logic models, neural networks, rule-based models, etc.) are also used and are becoming more popular.

Depending on the context, the “prediction error” may have other names. In PCA analysis for example, it is called the “squared prediction error (SPE)”, and in model-based fault detection literature it is known as the “residual signal”. Also notice that a model always involves a number of process variables, so model-based performance monitoring can be categorized within multivariate methods.

4.3 Threshold Design

Decision on raising an alarm is always a binary decision: either an alarm is raised or not. So, no matter what technique is used, the alarm is always raised based on a single variable. Obviously, it does not mean that the single variable is an actual process variable. It could as well be
Consider the alarm variable in Figure 4. As it can be seen, from about sample 1000, the mean of the variable changes. Assume that the first and second portions of the data represent normal and abnormal data respectively. The ideal monitoring system should raise no alarm for the first portion (normal data) and one sustained alarm for the second portion (abnormal data). So, an alarm raised in the first part will be a false alarm. Also if somewhere within the second part the alarm is cleared then this would be considered a missed alarm.

For this alarm variable, if the lower threshold is selected (solid line in Figure 4), there will be many false alarms (many points from the first part go over the limit). However, since the majority of abnormal points are higher than the threshold, very few alarms will be missed. On the other hand, if the higher threshold is selected (dashed line in Figure 4), the number of false alarms will be reduced significantly, but many alarms will also be missed. Therefore, a tight alarm limit will result in many false alarms but a few missed alarms; and a loose alarm limit will result in a few false alarms but many missed alarms. This constitutes the first major trade-off in alarm design: false alarm rate vs. missed alarm rate by changing the threshold. Threshold design is always a compromise between false and missed alarm rates. Depending on the situation and the operator's preference, the threshold can be quantitatively designed to achieve certain false alarm and missed alarm rates.

4.4 Processing the Alarm Variable

Regardless of the accuracy of the model or quality of the design, the alarm variable carries some noise. Because of this noise, comparing the alarm variable with a threshold may cause false alarms. To improve the alarm quality and reduce the number of false alarms, the alarm variable should be processed (Figure 3). Processing reduces the noise in the variable, and facilitates threshold design. Filtering, deadband and alarm delay are simple processing techniques that if utilized properly can reduce false and missed alarm rates. Here we briefly discuss filtering. Alarm delay and deadband will be discussed elsewhere [7].

Filters can reduce noise, remove bad or unwanted data, extract features, separate frequency components and modify statistical distributions. Several filters such as moving average and exponentially weighted moving average (EWMA), are very common in industrial data analysis and processing. An EWMA filter is defined as

\[
y(i) = \alpha x(i) + (1 - \alpha)y(i - n)
\]

where \(x(i)\) is the sequence of original raw data, \(y(i)\) the sequence of filtered data, \(\alpha\) the filter parameter and \(n\) the order of the filter.

Consider the alarm variable in Figure 4. Figure 5 shows the same variable after applying an EWMA filter with \(\alpha = 0.1\) and \(n = 1\). The filtered data has much less noise and the threshold shown in the figure can almost perfectly generate the ideal alarm.

The downside of using a filter is the delay it introduces in raising the alarm. Because of noise reduction property of filters, when the alarm variable changes due to an abnormality, the change will appear with some delay in the filtered data. And the more effective the filter is in reducing noise, the more delay will be introduced. This is the second important trade-off in alarm design: accuracy (or sensitivity) vs. latency by changing the filter parameters. A well designed filter will achieve the desired false alarm and missed alarm rates with minimum detection delay.

5. CASE STUDY: PUMP PERFORMANCE MONITORING

A heavy duty pump is used to pump sea water to an oil processing facility. A number of variables are measured in this pump, including the output flow rate and the head pressure which are used to monitor performance of the pump. Figure 6 shows averages of flow and head pressure measurements in one minute intervals for a period of about 30 hours. It is difficult to observe if any fault has occurred in this period by looking at the trends of variables only. To switch to multivariate performance monitoring, we plot the pump curve (scatter plot of head pressure vs. flow) as shown in Figure 7.

By investigating the pump curve, we observe that around the middle of the curve, there are some discrepancies from normal behavior (some of the points are above the normal curve). The discrepancies appear to be at the end of the
data set. To study the fault quantitatively, we first need to model normal pump behavior (model-based performance monitoring). A pump curve can be modelled as a quadratic function

\[ H = a_0 + a_1 Q + a_2 Q^2, \]

where \( H \) is the head pressure and \( Q \) is the flow. After fitting a second order polynomial to the data and using the actual flow measurements, we can calculate the expected head pressure (Figure 7, dashed). The prediction error (absolute value of the difference between the actual and the predicted values) can now be monitored as a measure of pump performance (Figure 8).

From the prediction error, it is clear that towards the end of the data set, the pump operation is not normal (the head pressure is different than what it is expected to be). A simple 3-sigma check limit will detect this fault, but it will also generate a number of false alarms and will miss some alarms. The alarm will also be chattering around the fault occurrence time. Notice that these values are averages, and therefore they have already passed through one level of filtering. In real time implementation, the instantaneous values will be used and therefore the number of false alarms will be more.

By investigating the histogram of the prediction error (Figure 8, right), we observe that:

- there is a small overlap between normal and abnormal data, so even if the “optimal” threshold (at 0.35) is selected we still have false and missed alarms.

To make the false and missed alarm rates as small as possible, we filter the data. Figure 9 shows the prediction error filtered with a moving average filter of size 15.

The data is now smoother and a threshold at 0.35 can almost completely separate the normal and abnormal data, resulting in yet fewer false and missed alarms. This is achieved at the cost of a delay in detecting the fault (accuracy-latency trade-off), as a moving average filter of size 15 delays the alarm anywhere between 1 to 15 samples.

6. CONCLUSIONS

Industrial plants face a large number of false and nuisance alarms that can overshadow all the advantages of monitoring and alarm systems. In this paper we discussed some of the challenges in the design and analysis of alarm systems. We reviewed some of the common causes of false and nuisance alarms and suggested a few remedies.

We briefly discussed two major trade-offs in alarm design:

- detection delay and false alarm rate trade-off
- false alarm rate and missed alarm rate trade-off

REFERENCES