Using percentile analysis for determination of alarm values

R.C. Lake, P.A. Agutter and T. Burke
Vivendi Water Partnership, Blackwell House, Three Valleys Way, Bushey, Herts, WD23 2LG.
(E-mail: Richard.Lake@gup.co.uk; Paula.Agutter@gup.co.uk; Terry.Burke@gup.co.uk)

Abstract The use of statistically derived alarm values for the minimisation of the risk of Cryptosporidium entering the water supply has been recommended in the report of the Bouchier expert group. Values derived using the mean plus standard deviations can be skewed by spikes, including those caused by filter washes. Indeed, theoretically, the use of standard deviations should only be used with normally distributed data. It is shown in this paper that turbidity and particle counting data is not normally distributed and that values derived by means of percentile analysis are more sensitive than using the mean plus standard deviations. Percentile analysis is able to produce alarms for data that is clearly different from normal operation, enabling smaller deviations from the usual to be identified. The values derived from one set of data are also shown to be appropriate for periods of around three months.

Keywords Particle counting; percentiles, statistical process control, turbidity

Introduction
Recommendations in the reports by the Badenoch expert group (Department of the Environment, 1990 and 1996) were made to establish good operational practices for minimising the risk of Cryptosporidium entering the water supply. However, it was not until the report by the Bouchier expert group (Department of the Environment, 1998) that specific recommendations regarding the evaluation of water quality changes were made. These recommendations, along with those in the UKWIR Guidance Manual (UKWIR Ltd, 1997–1998) have led to an increase in the monitoring of water treatment works as the number of on-line turbidimeters and particle counters have increased. As the amount of data collected on a water treatment works increases, more powerful data handling tools to interpret this data are needed, either on-line or as a historical data set.

On-line Statistical Process Control (SPC) can be carried out by the use of CUSUM charts. These are derived by plotting the cumulative difference of a measured parameter from the normal value against time. This approach has been examined within Vivendi Water Partnership (Agutter et al., 2001) and elsewhere in the UK water industry (Hall et al., 2000; Wetherill and O’Neill, 2000). However, there are drawbacks to this technique in that a number of different approaches can be taken to determine the “normal” value. Until this is better understood, the technique will not be used in real time. Also, by generating these graphs on a real time basis, the workload of the plant operative is increased, increasing the risk that less care is given to data analysis. Therefore, until this technique can be developed into a more user friendly form, its use will be limited.

SPC using historical data sets as a starting point has been widely used in the manufacturing industries since the 1920s, with the research team of Dr. Walter A. Shewart within Bell Telephone Laboratories (Dillon et al., 2000). The Shewart chart was not only used to maintain the quality of the product, but also to minimise waste production and increase profitability. Since this time, these charts have changed very little, with the mean and standard deviation being used to set the control values.

The UKWIR report promoting the use of SPC for managing Cryptosporidium risk
(UKWIR Ltd, 1998–1999) suggests using the mean plus two standard deviations is for a warning limit, with the mean plus three standard deviations for an action limit. These values are similar to those used in manufacturing industry (Dillon et al., 2000) and are based on the historical data that are derived from the process. However, these values, whilst preferable to adopting a “global” value, have little theoretical basis for the turbidity or particle count of drinking water. Applying standard deviation analysis to a data set infers that the data is normally distributed about the mean, which is typically the case in the manufacturing industry. For process streams at a water treatment works, this is rarely the case, with the data typically being skewed positively (Agutter et al., 2001) (i.e., a longer tail to the higher values than the lower). This not only skews the standard deviation, but also increases the mean value.

Due to the positive skew shown by the normal distribution curve, percentile charts have been considered. Percentile values remain unaffected at lower levels by spikes or filter washes giving a good indication of the water quality stability. Hargesheimer et al. (1998) demonstrated that the stability of a filter run could be assessed by comparing values at the 50th, 90th, 95th and 98th percentiles. The performance of filters against inlet conditions, coagulant dosing and seasonal factors could be identified. Percentile values can also be used to remove wash data from the statistical analysis of filters (Edwards et al., 2000). Also, the USEPA has proposed as part of the Interim Enhanced Surface Water Treatment Rule that the 95th percentile for turbidity should not exceed 0.3 FTU.

Although useful in providing comparisons between filter runs, the use of a specific percentile value for the setting of alarm values has little merit. Therefore, in this paper, using percentile charts to determine alarm values is considered. This methodology will be compared against the methodology of mean plus standard deviations proposed previously (UKWIR Ltd, 1997–1998), using data collected from a treatment works.

**Methods**

A process flow diagram for this site where the data was collected is presented in Figure 1. This site deals with raw water from three sources, all of which are somewhat karstic. Raw water A has significant surface influence and is thus treated as a surface water site, as inlet turbidity can reach 80 FTU. Treatment is by coagulation and clarification, rapid gravity filtration, followed by GAC contactors. The water from the other two sites is similar (Raw water B), and this is treated by means of coagulant addition in the pipework immediately upstream the GAC contactors (Direct Contact Filtration (Hutchison et al., 1999)). The sources blend prior to super and de-chlorination, with the flow from Raw water B being twice that from Raw water A.

Turbidity data is collected continuously from the outlet of the GAC contactors, using HF Scientific turbidimeters, which are calibrated monthly. Data is stored on-site using Penny

---

**Figure 1** Process flow diagram for the site
and Giles Trendview 5+, from where it can be downloaded into a spreadsheet format. Particle count data has also been collected on line from the combined outlet of the GACs, using Hach 1900 WPC particle counters, calibrated at appropriate intervals, with the data being logged onto Aquaview software.

**Results and discussion**

A typical set of particle count data collected over a period of 17 days from the combined outlet of the GAC contactors is shown in Figure 2. It can be seen that there are four narrow spikes in the data, relating to filter washes. There is more frequent washing of these GAC contactors due to their use as Direct Contact Filters for Raw Water B. The four contactors which are used in this way wash on a 96 hour cycle, i.e. one per day. There are two other areas of interest: a prolonged period of higher particle counts and a spike, which is longer than the normal spike seen during a wash.

Statistical analysis of the data is shown in Table 1. From this, it can be calculated that the alarm values, using mean plus two or three standard deviations (UKWIR Ltd, 1997–98), would be: Warning Alarm – 38 particles/ml, Action Alarm – 54 particles/ml.

As stated in the Introduction, the use of mean and standard deviations as the alarm limits is based on the assumption that the data is normally distributed. In Figure 3, the distribution of the data is presented.

It is clear from this figure that there is a significant positive skew to the distribution of the particle count data, which is similar to the distribution for turbidity data that has been seen previously (Agutter et al., 2001). In addition, for a normally distributed data set, the mean, median and mode averages should be at the same value. In this case, the mode is 2.4, the mean is 5.6 and the median is 2.7. The mean is thus more than double the other average measures, showing the effect of the positive skew.

The percentile analysis chart for this data is presented in Figure 4. This is prepared by numerically ranking the data against a percentile value. From this graph, it is possible

---

**Figure 2** Particle count data from the combined GAC outlet

**Table 1** Basic statistical analysis from the particle count data in Figure 1

<table>
<thead>
<tr>
<th>Particles/ml</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean average</td>
<td>5.6</td>
</tr>
<tr>
<td>Mode average</td>
<td>2.4</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.1</td>
</tr>
<tr>
<td>Maximum</td>
<td>489.6</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>16.2</td>
</tr>
</tbody>
</table>
to identify the normal value at which the process operates. For all but the most erratic of processes, this is around the mode value. The stability of a process can be ascertained by comparing percentile values to the normal operating value (Hargesheimer et al., 1998). It is stated above that the median (i.e. the 50th percentile) and the mode are very close in value. Comparing other percentile values to the normal running of the process shows that in this case the process is stable for 90% of the time.

What is of more interest for the purposes of statistical process control on a water treatment works is the value at which the process moves significantly away from the normal value. It is at these values that the process alarms should be set as the plant runs higher than is normally expected. From the percentile curve in Figure 4, it can be seen that this has been identified as being 8 particles/ml (at the 90th percentile). The curve then moves further away from normal running at 17 particles/ml, between the 96th and 97th percentiles. These values should be used as the warning and action alarm limits respectively.

These values are derived purely by inspection, rather than being linked to a particular percentile value. Indeed, while the percentile value at which this takes place is of interest, this is only to know how long the plant would be in alarm for the period under consideration.

Both sets of statistically derived alarm levels are shown with the raw data in Figure 5. Comparing these values, those derived using percentile analysis are much lower than those derived by using the mean plus standard deviation methodology. This means that more alarms would be triggered using percentile analysis. It is thus important that the extra alarms triggered would be justifiable, and represent a period where the plant may be at increased risk.
As stated above, and indicated on Figure 2, there are two areas of particular interest for this data set. These are the extended spike in the middle of the period under consideration and the extended period of higher particle counts at the beginning of the data period. For the large spike, both sets of analysis would have triggered an alarm at approximately the same time. There is thus little advantage in either method for this event.

The main difference between the sets of analysis comes from the extended period of higher particle counts. The standard deviation analysis would not have triggered an alarm (either warning or action) despite the level of particle count being considerably higher than usual. The percentile analysis in this case would have registered a warning alarm during most of this period, with the action alarm also being triggered for a short time early on. This indicates that the percentile analysis is much more sensitive to the particular data set than standard deviation. However, it must be stressed that even the use of standard deviation set alarm limits for specific plants is preferable to setting blanket alarm values on a company wide basis.

Although the data presented above is derived from particle counting, the same principles apply to turbidity data. Indeed, as turbidimeters are much more prevalent at water treatment sites, these values can be used for examining a much greater quantity of raw data for setting alarm values. By using data over a greater period, the alarm values can remain in place for a longer period. In Table 2, percentile derived turbidity alarm values for seven consecutive two to three week periods are presented. These values were taken from the autumn to early winter period, when water quality is highly variable on the inlet to the sources from an individual Direct Contact Filtration contactor (Raw Water B on Figure 1). Despite the variability of the water coming onto the plant, the alarm values are consistently around 0.1 FTU for the warning alarm and 0.25 FTU for the action alarm. However, the actual percentile at which these values are seen can vary between the 87th and the 99.9th for the

Table 2 Percentile derived turbidity alarm levels over seven consecutive periods

<table>
<thead>
<tr>
<th>Period</th>
<th>Warning alarm (FTU)</th>
<th>Percentile when triggered</th>
<th>Action alarm (FTU)</th>
<th>Percentile when triggered</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Oct to 17 Oct</td>
<td>0.12</td>
<td>93</td>
<td>0.22</td>
<td>99.8</td>
</tr>
<tr>
<td>18 Oct to 31 Oct</td>
<td>0.12</td>
<td>87</td>
<td>0.26</td>
<td>97</td>
</tr>
<tr>
<td>1 Nov to 17 Nov</td>
<td>0.11</td>
<td>99.9</td>
<td>0.24</td>
<td>99.94</td>
</tr>
<tr>
<td>18 Nov to 30 Nov</td>
<td>0.10</td>
<td>88</td>
<td>0.26</td>
<td>99.3</td>
</tr>
<tr>
<td>1 Dec to 16 Dec</td>
<td>0.10</td>
<td>91</td>
<td>0.27</td>
<td>99</td>
</tr>
<tr>
<td>17 Dec to 31 Dec</td>
<td>0.09</td>
<td>98.5</td>
<td>0.25</td>
<td>99.3</td>
</tr>
<tr>
<td>1 Jan to 15 Jan</td>
<td>0.09</td>
<td>99.9</td>
<td>0.18</td>
<td>99.94</td>
</tr>
</tbody>
</table>
warning alarm and the 97th and 99.94th for action alarms. This depends upon the stability of the filters during the period being considered. Indeed, it is feasible that an alarm value may not be reached during a period of running, if there is a period of completely stable operation.

The closeness of these results has two consequences. Firstly, alarms derived for one period can be reasonably applied to the following period. Secondly, the alarm values do not need to be derived too frequently, thus not adding greatly to the duties of the site operator. These values are derived by inspection. Therefore, to derive the values on a frequent basis, particularly if one is looking to set individual alarms on a bank of filters at a large site, would be very time consuming. These results suggest that alarm values that are calculated on a quarterly basis would continue to be appropriate.

**Conclusions**

Percentile analysis has been shown to be more sensitive in indicating plant performance, and thus the setting of alarm values, than standard deviation analysis. It does not have the disadvantage of being skewed by high values coming from spikes, which increase both the mean and the standard deviation of the data. This means that percentile analysis can clearly identify events that would otherwise be missed.

Although the method requires visual inspection of sorted data to derive values, these values have been shown to vary very little over a three month period. Therefore, there is only the need to review these values on a regular, say quarterly, basis.

**References**


