

Modelling the Risk of Pollution Using Alarm Data and Asset Maintenance Data

Introduction

Pollution can have serious environmental, financial and social consequences not only for the people and area directly affected, for example damage to people's homes and businesses, but also for wider society because of the consequences on surrounding areas. Many pollution incidents are due to asset failures, particularly pumps at pumping stations, and so the risk of pollution can be reduced significantly by carrying out asset maintenance to help minimise the risk of asset failure.

This paper describes a case study of how alarm data from pumps at wastewater pumping stations, asset maintenance data, asset data and pollution data were used to model the risk of pollution from each station (site). The model identifies the factors that determine the risk of pollution. The output of the model is the risk of pollution from each site, enabling water companies to target their maintenance resources to those sites with the highest risks of pollution. The model can be updated frequently, for example weekly, to ensure that it uses the latest information to calculate the current risks of pollution.

Alarm Data

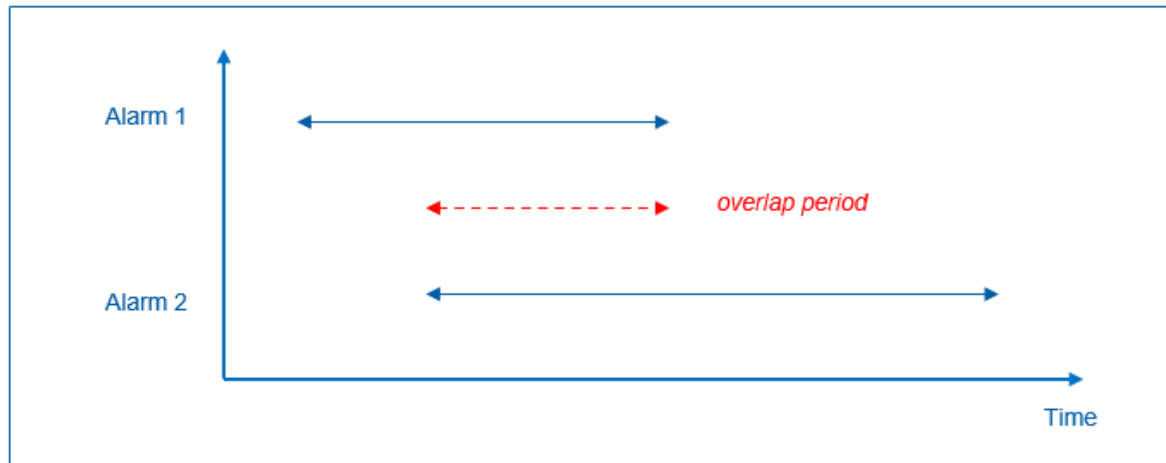
The alarm data were for pumps at 2,834 wastewater pumping station sites for one year. The alarms file had 1,068,040 records, an average of one record every day at each site. The distribution of the number of alarms at the sites was highly asymmetric with half the sites having fewer than 50 alarms triggered during the year and the other sites having between 50 and 62,434 alarms triggered during the year. Similarly, half the alarms lasted less than 3.25 minutes and the other alarms lasted between 3.25 minutes and 503,829 minutes (350 days).

The alarm data had a number of fields but the fields that required most preparatory work were the alarms' start date times and end date times. The very long tail of the highly asymmetric distribution of the number of alarms at the sites meant that many sites, if not all, had overlapping and nested alarm records. This is a major problem that can lead to negative intervals between alarms as shown in Figures 1 and 2, and must therefore be addressed before the data are analysed and modelled. Deduping such records to a set of non-overlapping and non-nested records is a big and complex task because it requires many iterations. The result is one record that captures their true nature and does not coincide with other records for the site.

Overlapping Alarms

Two alarms for a site overlap when the second alarm begins between the start and end times of the first alarm, and ends after the end of the first alarm. Figure 1 shows two overlapping alarms and the overlap period (overlapping can be more than two records deep).

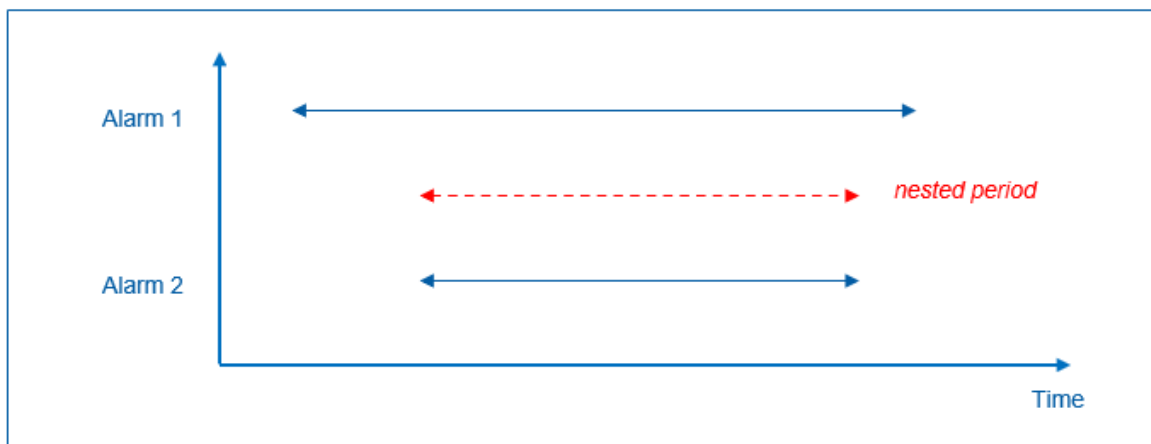
Figure 1: Overlapping Alarms



Nested Alarms

Two alarms for a site are nested when the second alarm begins and ends between the start and end times of the first alarm. Figure 2 shows two nested alarms and the nested period (nesting can be more than two records deep).

Figure 2: Nested Alarms



The number of iterations and therefore time required to dedupe overlapping and nested alarms depends mainly on the depth of overlapping, depth of nesting, and extent and number of the alarms that overlap and are nested. In general, as the number of alarms for a site increases, these factors increase and so sites with large numbers of alarms require more iterations than sites with fewer alarms.

Asset Data and Pollution Data

The asset data describe the process and function at each site. As with the alarm data, the maintenance data required deduping. The pollution data showed the sites that caused pollution and the times of the pollution. Pollution (the target variable in the models) is recorded as a binary variable.

File Used for Modelling

After each source file was prepared, they were merged and the merged file then sorted for each site by the start date time of the events (alarm, maintenance and pollution). This ensured that the cause effect order of the events was reflected in the file so that it had the correct structure for analysis and modelling. The source files had 1,145,335 records and the merged file had 91,239 records, a 92% reduction. This number gives an idea of the extent of the work required to transform the data to the correct form. All the following figures and tables are for the deduped data.

Figure 3 shows the distribution of the number of alarms triggered, Figure 4 shows the distribution of the number of planned works and Figure 5 shows the distribution of the number of reactive works.

Figure 3: Distribution of the Number of Alarms Triggered

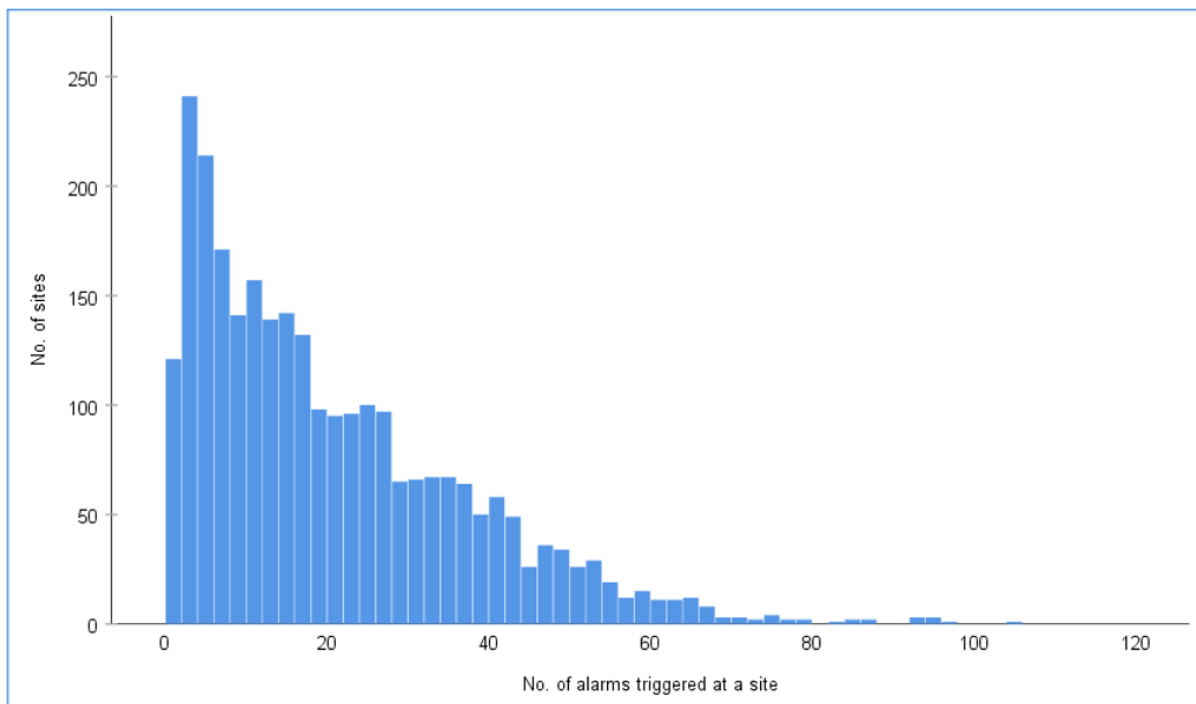


Figure 3 shows that the distribution of the number of alarms triggered at the sites is skewed with a long tail (the mean is 20.2 and the maximum is 105).

Figure 4: Distribution of the Number of Planned Works

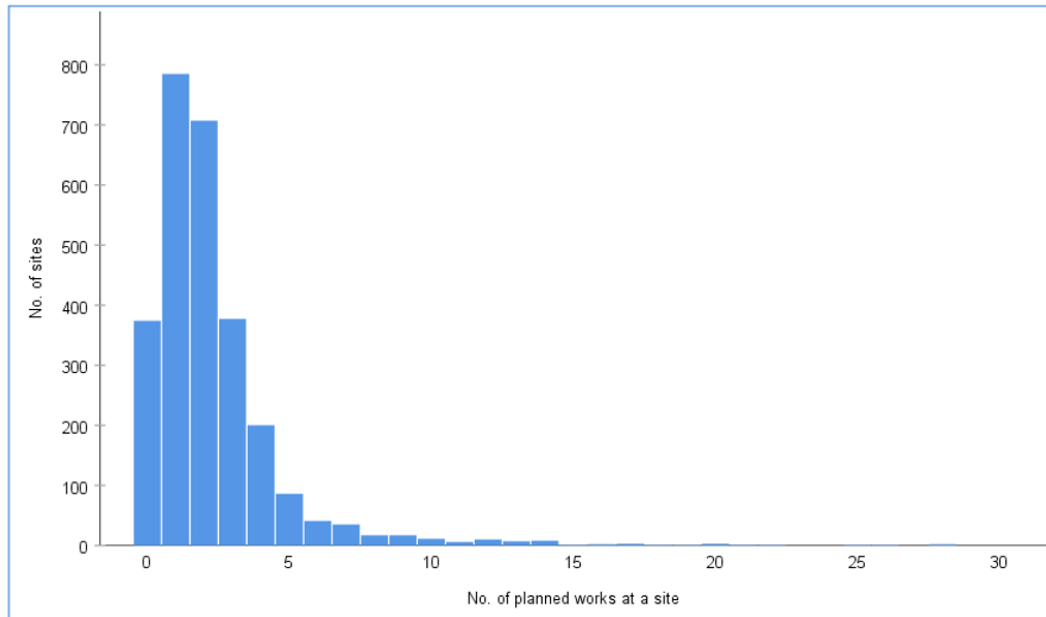
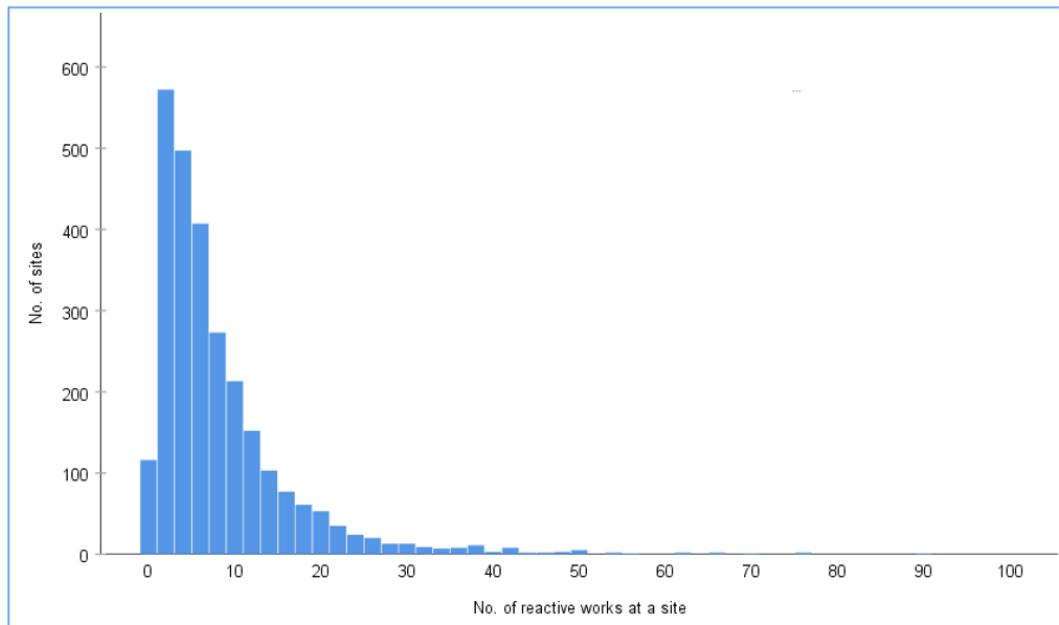


Figure 5: Distribution of the Number of Reactive Works



Comparison of Figures 4 and 5 shows that they have broadly the same profiles but different scales.

Figures 6 and 7 show how the number of pollution incidents from a site depends on the number of planned works and the number of reactive works.

Figure 6: Variation of the Number of Pollution Incidents with the Number of Planned Works

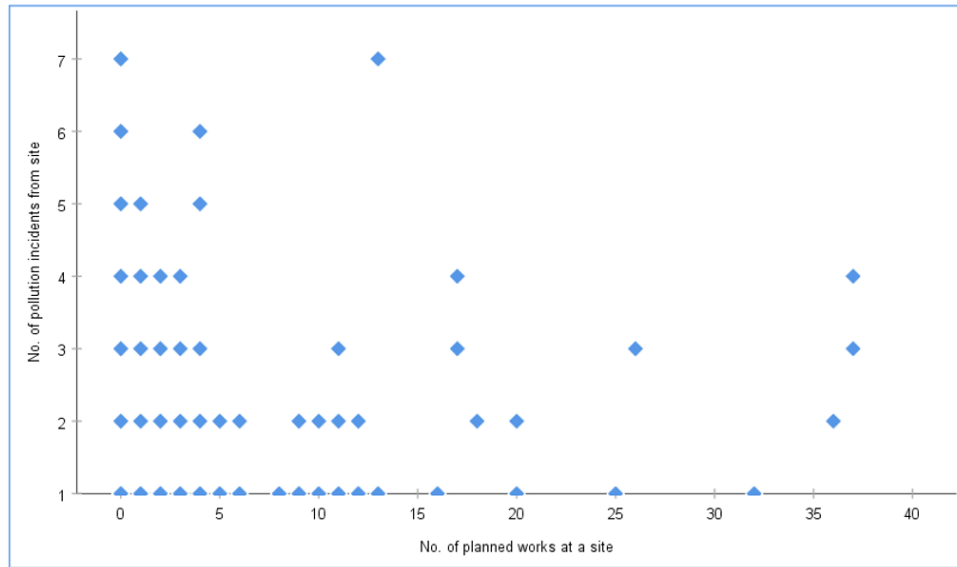
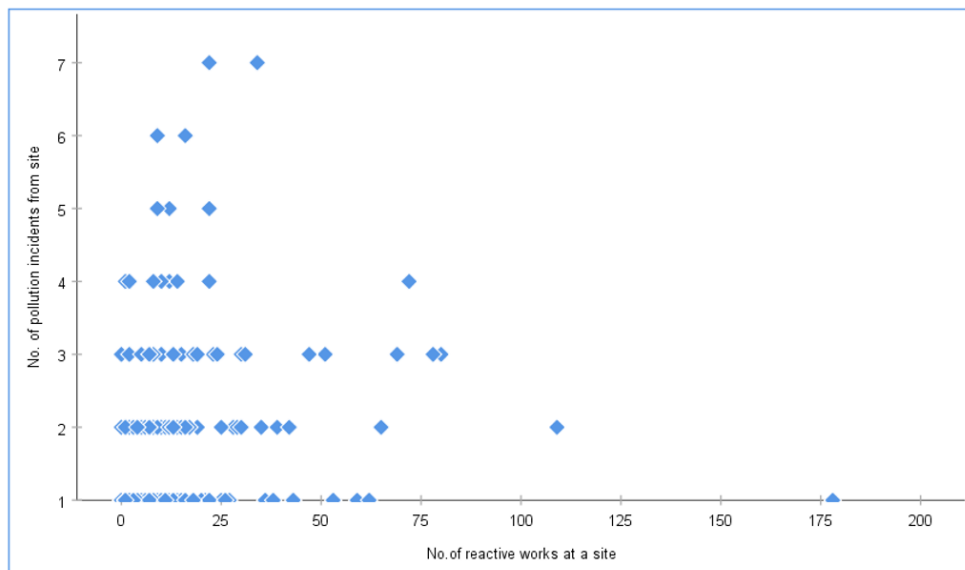


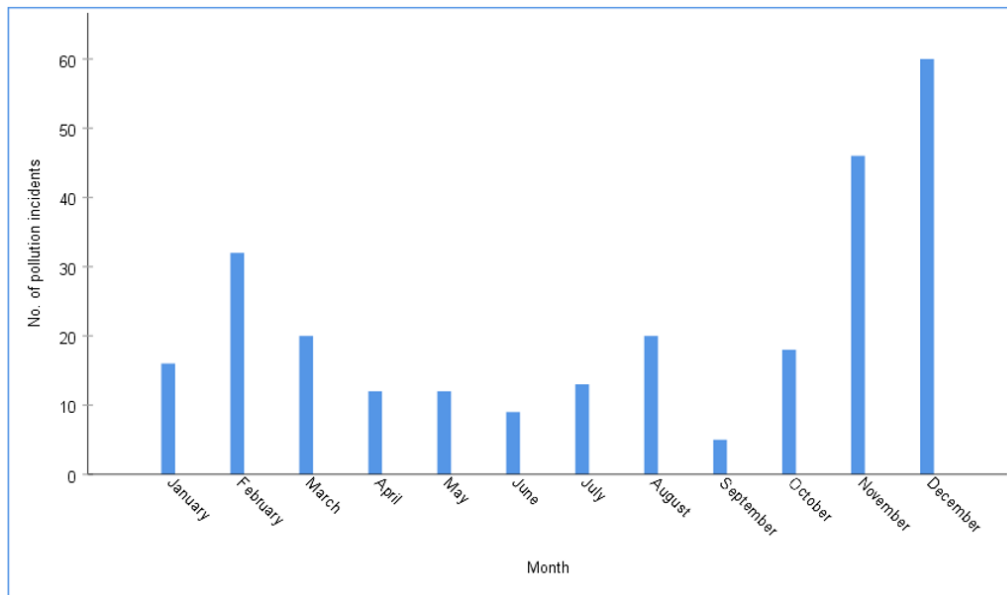
Figure 7: Variation of the Number of Pollution Incidents with the Number of Reactive Works



Figures 6 and 7 show that there are many more reactive works than planned works and that planned works are much better than reactive works at reducing the number of pollution incidents. This result is as expected and suggests that the different types of maintenance is a good predictor of the risk of pollution.

Figure 8 shows how the number of pollution incidents depends on the month. As expected, it shows that wetter months tend to have more pollution incidents than drier months.

Figure 8: Variation of the Number of Pollution Incidents with Month



Risk of Pollution Models

Risk of pollution logistic regression models for Sewage Handling (SH) and Preliminary and Storm Treatment (PST) were developed, with pollution as the binary target variable. Table 1 shows the models.

Table 1: Risk of Pollution Models

Predictor	Sewage Handling	Preliminary and Storm Treatment
Constant	1.670	0.947
Site group type	overflow and wastewater catchment: -0.708 other: 0	overflow and wastewater catchment: -1.871 other: 0
Alarm unit process	drainage: -0.690; other: 0	
Alarm process function	outstation: -0.545; other: 0	
Previous event	planned/reactive work: -0.685 alarm/pollution: 0	planned/reactive work: -1.143 alarm/pollution: 0
No. of planned works at site	-0.033	-0.197
No. of prev. pollutions at site	0.157	0.544
No. of alarms at site	-0.327	0.545

Model Interpretation

Negative coefficients in regression models mean that as the predictor increases the target variable decreases, and positive coefficients mean that as the predictor increases the target variable also increases. Therefore, as *no_planned_works_site* increases the risk of pollution decreases, and as *no_of_prev_pollutions_at_site* increases the risk of pollution increases.

Predictors with more than one coefficient are interpreted differently because the coefficients show the relative effects of the categories, not their absolute effects. For example, consider *previous_event*. Relative to the effect of alarm/pollution (coefficient 0), the negative coefficient for planned/reactive work means that planned work and reactive work reduce the risk of pollution by 0.685 units. The results are independent of which category is the reference category. Similar conclusions can be drawn for *site_group_type*, *alarm_unit_process* and *alarm_process_function*.

The coefficients of *previous_event* and *no_planned_works_site* show that planned work or reactive work at PST sites is much more effective at reducing the risk of pollution than planned or reactive work at SH sites.

The different signs of the coefficients of *no_of_alarms_at_site* suggests that in general alarms from SH sites are acted on but alarms from PST sites are not acted on and so more attention is paid to alarms from SH sites than alarms from PST sites. This inference is borne out by comparing the coefficients of *no_of_prev_pollutions_at_site* (0.157 for SH and 0.544 for PST). This shows that a history of pollution at PST sites increases the risk of more pollution than a history of pollution at SH sites.

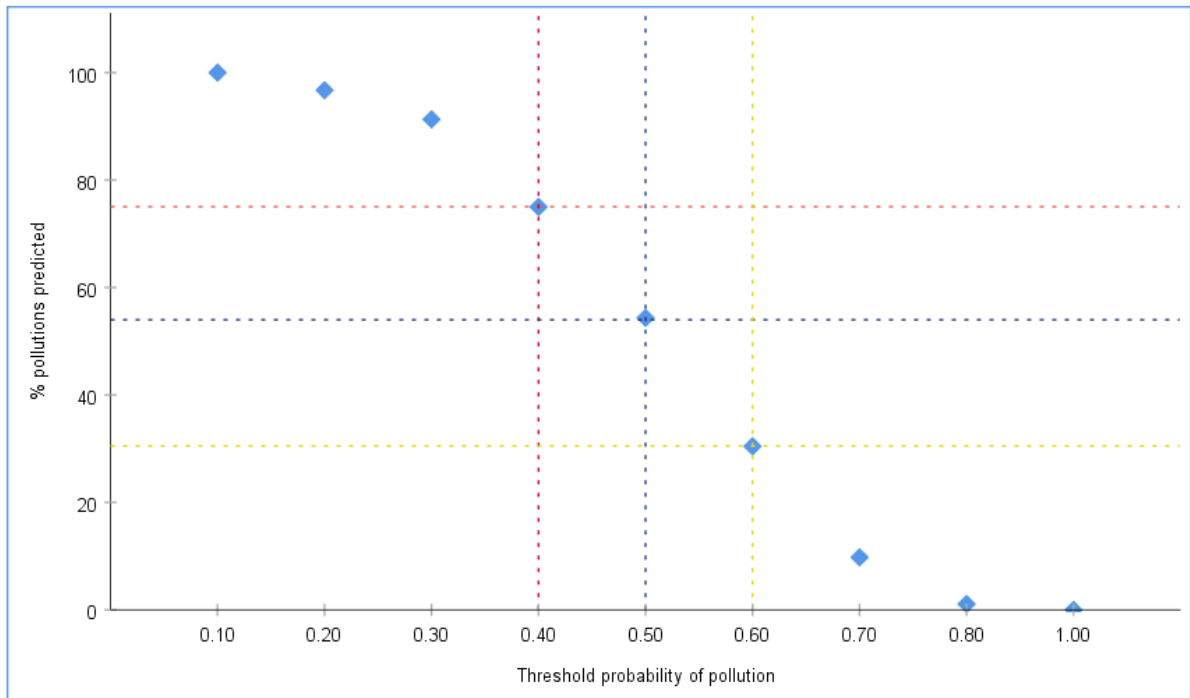
These conclusions are in line with experience and practice because planned maintenance is carried out to reduce the risk of asset failure and thereby the risk of pollution, and sites with a history of pollution have higher risks of pollution (because they have not had maintenance).

Threshold Risk of Pollution

Even though the target variable in logistic regression models is binary, the output of the models is a probability (the risk of pollution in this case). It is usual to specify the threshold probability in logistic regression models as 0.5 so that the predicted outcome for cases (sites in this case) with output probabilities smaller than 0.5 is the non-event (no pollution in this case) and the predicted outcome for cases with output probabilities greater than 0.5 is the event (pollution). Sometimes, particularly when the consequences of 'failure' are severe, the threshold probability used is calculated from a trade-off of the gains against the costs of the outcomes for different threshold probabilities.

Figure 10 shows how the percent of pollutions predicted varies with the threshold probability for SH sites. For example, if the threshold probability is 0.40, 75% of the pollutions are predicted correctly.

Figure 10: Variation of Percent Pollutions Predicted with Threshold Pollution Probability



Conclusion

This paper shows how risk of pollution models can help water companies gain insight and understanding into the causes of pollution and identify sites with high risks of pollution. The models clearly show that the best way of reducing the risk of pollution is to carry out maintenance at sites with high risks of pollution at that time.

Dr Atai Winkler

e: atai.winkler@pamanalytics.com

m: 07817 263016

w: pamanalytics.com