Paper 142-2008

# The Early Warning Project: Prewarning and avoiding problems and costly down time in complex industrial processes

Torulf Mollestad: SAS Institute, Oslo N
Fredrik Fossan, Frode Oplenskedal, Pål Navestad: ConocoPhillips, Stavanger N

## ABSTRACT

The goal and focus of the Early Warning Project has been to find early indicators of potential trouble or suboptimal performance of a complex industrial process. If such indicators are found, there may still be time to perform the appropriate actions in order to stabilise the process and avoid the problem. An alert service should pick up on a signal of upcoming trouble as early as possible, and maintain an "alert degree" which varies according to new evidence coming in. Such functionality will be more potent – and complex – than fixed value-based alert generation that are in abundance today.

One test case will be presented, namely the performance of the low-pressure separator on ConocoPhillips' Ekofisk 2/4 J (EKOJ) platform in the North Sea. This is a highly complex process and there are a lot of influences that are not fully understood. The EKOJ receives streams also from other platforms, meaning that performance issues on these, "up-stream" flows may influence the performance on EKOJ or, more specifically, the separator that we focus on.

A method has been designed that, given a set of sensory data (time series showing temperatures, pressures, flows etc.) around the industrial process (on EKOJ itself as well as upstream to EKOJ), extracts knowledge in the form of sets of simple rules. Such rules are intuitively understandable to the experts, but may also be used for real-time monitoring of the working system. Each rule provides diagnostic capability but may also be combined with others to yield combined evidence and an overall probability of imminent problem or failure. Rules may also be used to directly diagnose the situation at hand.

## THE EARLY WARNING PROJECT

### THE GOAL

The Early Warning Project at ConocoPhilips has been undertaken with the goal of finding early indicators of problems coming up in complex industrial processes, to aid diagnosing the situation and solving the problem. Ideally, the system should be able to suggest measures/actions to operators or experts that would be appropriate in order to stabilize the process. Such an alert service should pick up on a signal of upcoming trouble as early as possible, and maintain an "alert degree" which varies according to new evidence coming in.

In the project, we extract knowledge in the form of rules that describe statistically strong relationships between *events* happening in the underlying process. Data mining methods have been utilized in order to extract this knowledge directly from the time series data, which are collected at sensors sitting at different points around the industrial facility. Typically, such sensors measure temperatures, pressures, flows, valve openings etc. at different points around the process. We use the term *event* to describe any abnormality happening on either of these measure points – examples of typical events are "extremely low flow", "sudden jump in temperature" "unstable/volatile pressure", etc.

The purpose of the project, then, has been to find relationships between events such as the above, and problems on the overall process. One emphasis in the project has been interpretability of the results – the knowledge extracted should be made explicit to the domain experts and also readily verified by them. Typically it can be expected that a number of the patterns found will be replicas of general knowledge shared by all users, or even represent trivial knowledge. However, there will typically also be hidden relationships in the data that will prove of high value once revealed. The importance of tight cooperation with domain expertise throughout the project can therefore not be overemphasized.

#### THE TEST CASE

The method has been tested on one selected case – namely the performance of the low-pressure separator on the Ekofisk 2/4 J (EKOJ) platform in the North Sea. The main reason for the choice was that this construction is fundamental to the overall performance of the platform. At certain times the separator may start to hiccup and even grind to a halt. If this happens, the entire industrial process stops with it, and a huge profit loss is the result. Moreover, this is a highly complex process, a lot of influences are not fully understood and there is a lot of potential gain from extracting yet more knowledge from the vast data sets.

The separator itself is essentially a huge pool which takes in a stream of crude oil coming from the wells, and separates them into the essential constituents – oil, water and gas. Three flows are thus output from the separator, each of which is submit to further treatment and refinement on the platform and, finally, on shore. On the input side, the EKOJ receives streams also from other platforms, meaning that performance issues on these, "up-stream" flows may influence the performance on EKOJ or, more specifically, the separator that we focus on. In all, there are a lot of sensors/measure points, both upstream and downstream to the separator that may influence or signify the performance of the separator and thereby the performance of the entire process.

A substantial work was done going through the topology of the platform and isolating those measure points that might potentially say anything about or be relevant to the performance of the separator. Data from each of these was found as so called *tags* (individual time series) in ConocoPhillips' time series data base, Pi. After careful scrutiny, around 115 potentially relevant tags were originally identified. Of prime focus is the water level in the separator. Typically, in times of trouble, the level will sink. Later we will explain how the definition of a problem situation was done – through the definition of a *target* variable.

#### AN OVERVIEW OF THE METHOD

In the project, a process has been laid out to restructure the original Pi data (the tags) into *event-based data*, i.e. new time series that keep track of whether a particular event happened at a certain time point. To do this, a number of basic *event templates* were defined (the framework is designed so that new definitions may be easily added). The current library of event templates includes:

- The trend reaching a high (or a low) level – may be defined in terms of
  - a fixed value – a user/expert specified maximal or minimal value, or
  - a level collected from taking statistics on a local time period, e.g. the 95th percentile
- Volatility – is the trend stable or unstable in a given time window?
- Shocks – large jumps or falls in data values at particular points
- User defined events (independent of tags)

Each of the original Pi tags was combined with *each* of the event templates, giving rise to new binary/ two valued time series (taking either 0 or 1 as values) indicating whether the event occurred (1) at a particular time point – or not (0). About 1000 new event variables were constructed in this way, combining each of 115 tags with each template function; each of these new event variables was represented binary on 10 minute intervals. We also added to the picture the manually generated reports telling when the operators themselves experienced problems - added to the analysis through a (binary) "operator perceived trouble"-variable. In this way, the operator's view of the process may be checked against a more objective indicator of a problem (as will later be known as the target).

At this point, techniques from statistics/data mining were used to find those out of the 1000 event variables

that were best correlated with the target, i.e. tend to take on value 1 before and around a problem. Variable clustering applying principal component analysis has been applied for this purpose – using four hour aggregate data. At this stage in the process, a number of events may be discarded as unrelated to the break down or underperformance of the process. Further investigation on the relationships between events was then done by performing association/sequence analysis on the discrete event data that are now determined relevant. The results of this analysis is a set of rules, typically of the form A => X or A & B => X; the intuitive interpretation of the latter being:

- Given the sequence A & B => X – Upon having observed both events A and B we may increase our belief that X will happen within a limited time interval

In the following we will refer to such rule sets as *models,* being a compact description of a complex reality, one which may be interpreted and be used as a source of knowledge for making decisions. It is important to emphasize that such a rule set is only a model insomuch as it needs to be presented to business experts and clarified with them. Only after having ensured that the rules reflect genuine and non-trivial patterns should they be used in an automatic monitoring scheme. In the course of the project, a number of different reports and graphical displays of the data have been developed to support iterative development of the models. Some of the most important categorisations and visualisations of the data are shown later in this document.

#### THE HIGH LEVEL ALGORITHM

The procedures implemented in SAS Enterprise Guide essentially follow six main steps, namely

1. Imputation and tag selection
    - Linear imputation over intervals of missing values
2. Preparation of tag data – augmentation with statistics
    - Augmenting with time varying statistics
3. Processing point derivatives
    - Defining event templates – each to be combined with each tag
    - Automatic generation of event tags
    - Deletion of non-varying and dispersed event tags
    - Combination of data into compound intervals
    - Definition of a target variable
    - Production of target agreement data set
4. Clustering and processing of clusters
    - Clustering on the target agreement data set
    - Production of reports, performance graphs, tree graphs
    - Selection (automatic or/and manual) of relevant clusters (and their event tags)
5. Association rules mining
    - Extraction of one-item and two-item association rules using Enterprise Miner
6. Postprocessing and visualization
    - Producing rule reports
    - Displaying rules through association graphs

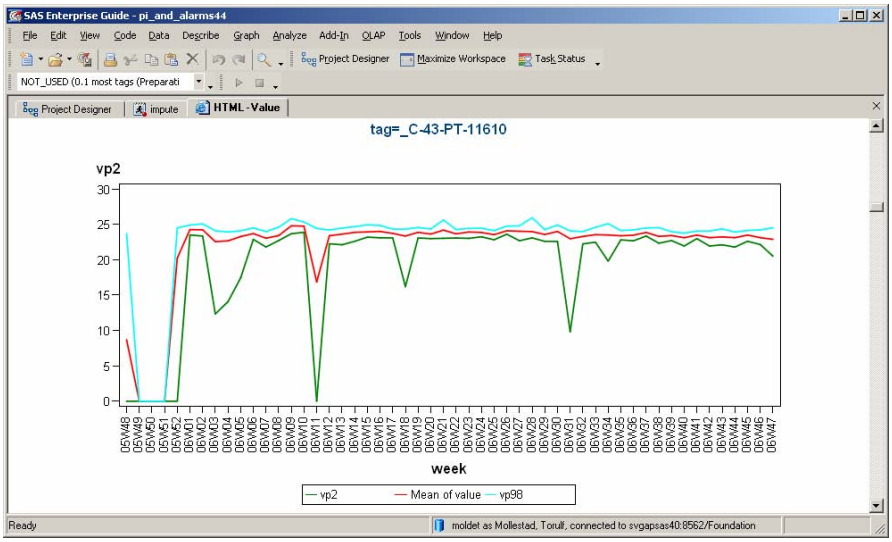
## THE METHOD IN DETAIL

### IMPUTATION AND TAG SELECTION

In this part of the process, the original time series from Pi was made complete by applying a linear, "non-dramatic" interpolation function across gaps of missing values. The resulting data set contained one year's worth of data on ten minute intervals for 115 tags, yielding about 52000*115 distinct measurements.

### PREPARATION OF TAG DATA – AUGMENTATION WITH STATISTICS

In this part of the process, simple weekly statistics were extracted from each of the 115 completed Pi tags. The statistics that were computed include level, volatility and jump statistics in terms of percentiles. The statistics are displayed in graphs such as in the following example for level statistics. The graph shows weekly 2nd and 98th percentile levels and the variations throughout the year:




There are in general great weekly variations; it should be further considered whether weekly statistics is the best choice wrt. representing the overall change over time of the process; our experience suggests that more subtle algorithms based on traversal of the data might be a better choice.

Periods of exception (for instance controlled halting of the process for maintenance) were excluded from the data. Experience tells us that the current definitions of what constitutes such exceptional states should also be put under scrutiny.

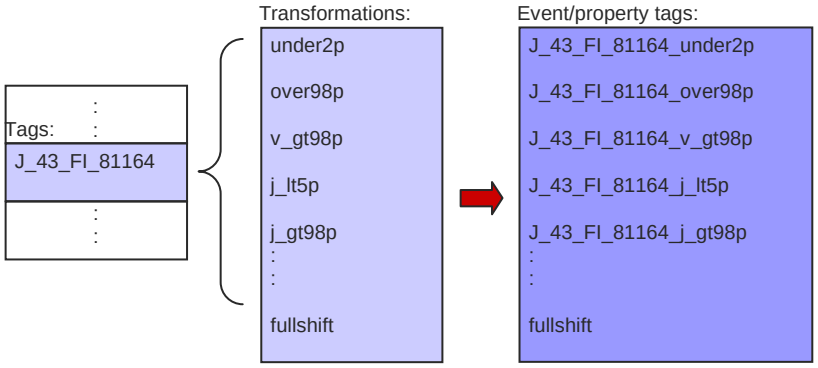
### PROCESSING POINT DERIVATIVES

In this part of the process, a number of event templates were defined, very simple in form, and each of them was then applied to all input tags. One such template function is the following, using the (weekly) 2nd percentile as shown in the graph above:

```
under2p = (value < vp02);
```

Each original tag P would now be combined with *each* such event template. The combination of tag P with template T would yield a new "event tag" named "P_T", the sole purpose of which was to represent an event type "T" (for instance "a low value", "a sharp fall" or "a high volatility") happening on trend/tag "P". For instance, the tag "A_43_PT_P2_under2p" maintains knowledge about one single property of the Pi tag by name "A_43_PT_P2" – namely whether the *event* "under2p" (going under the 2nd percentile, i.e. the level going very low compared to the surrounding data) happened for that tag at a particular time point.

The user is free to define other simple functions like these; care has been taken to make sure that modifications of this type may be done with ease. The Pi tag “P” will from this point on live as a set of binary event tags, “P_over98p” being one of them. Again, about 1000 derived event tags were defined in this process, each on the same frequency as the original data, i.e. each ten minutes. In this light, we now have 52000*1000 binary measurements to keep track of.




Consultation with the process experts suggested the following definition of the target – i.e. the *definition* of a problem on the processor itself: The water level (Tag J_43_LT_00407) falling below 5th (weekly) percentile, followed - within 30 minutes (three ten minute lags) - by sum of water flows measured at valves A and B (J_43_LY_00407_AB) falling below a fixed (predefined) level 4. In the code, the target was represented as four SAS macro variables, making it easy to modify (the interpretation should be obvious given the target definition):

```
%let target1 = J_43_LT_00407_under5p;        /* going under the 5. percentile    */
%let target2 = J_43_LY_00407_AB_underband;   /* going under 4                    */
%let target = low_water_closed;              /* the name of target to be defined */
%let lags = 3;                               /* 30 minutes                       */
```

The target, named “low_water_closed”, was shown to happen (target=1) 22 times in the year covered by the data (dec. 05 to nov. 06). In further processing the data, aggregation of all events into larger time intervals – four hours – was done, this in order to make the system as robust as possible. Thereby, one four hour group would keep track of all events happening within, and further analysis was based on these aggregate data points (each event tag represented now by 2100 rather than 52000 measurements). Moreover, all events within a user defined period after each of the 22 target events are deleted from the data, in order to remove any confusion/exceptional intervals from the data. We do not wish a large number of rules that merely describe consequences of problems post factum.

In the resulting data, a number of event tags could be deleted readily for different reasons, including being too frequent or having a low of correlation with the target. The surviving tags, including the ones shown below, each occur less than the chosen limit percentage, and are thus candidates for contributing to interesting rules. The column to the right shows the frequency of the event in the overall set of around 2100 four hour groups. In this light, the target itself had a frequency of 1.08% (22/2100)

| | |
|---|---|
| A_43_PT_P2_under2p | 0.1121673004 |
| B_43_FX_5682_over98p | 0.1387832743 |
| B_43_FX_5682_under2p | 0.0598859316 |
| B_43_FX_5682_v_gt95p | 0.1102661597 |
| : | |

For each event tag, each of the 2100 data points could be marked a false positive, false negative, true negative or true positive, depending on whether the event happened in conjunction with the target event. This data was then passed into a variable clustering algorithm as explained in the following section.

**CLUSTERING AND PROCESSING OF CLUSTERS**

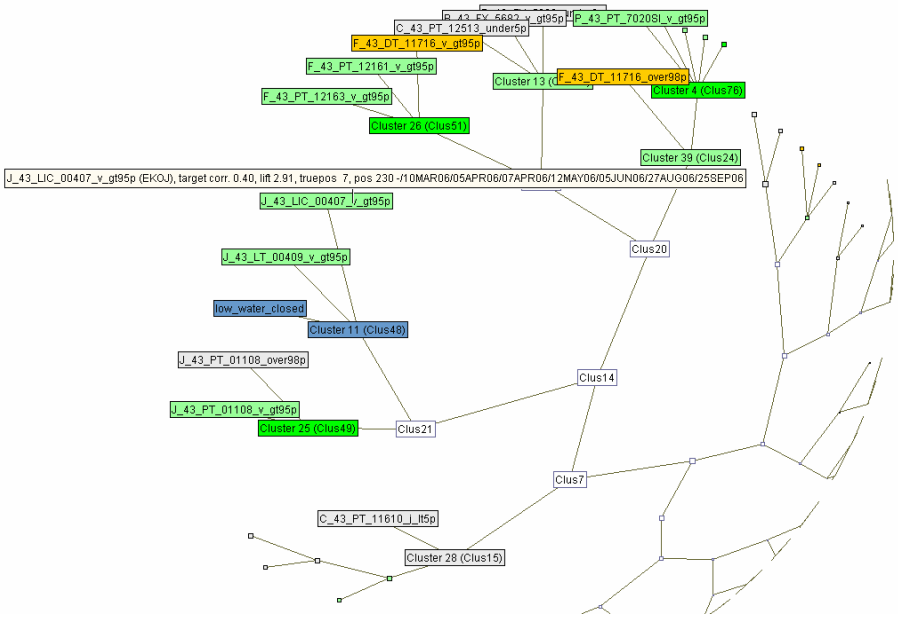
In this part of the process, the variables are clustered according to their agreement on the target variable. By clustering in this manner, we make sure that the clusters are focused around the target event; in other words that event tags that behave similarly wrt. the problem times will tend to be clustered together. For instance, for cluster 10;

| *Tags:* | *Problem times (only dates shown here):* |
|---|---|
| J_43_PT_01286_j_lt5p | 01DEC05/25JUL06/09OCT06 |
| J_43_PT_00312_j_gt95p | 01DEC05/25JUL06/09OCT06 |
| J_43_PT_00412_v_gt95p | 01DEC05/25JUL06/09OCT06 |
| J_43_PI_80196_v_gt95p | 01DEC05/25JUL06/09OCT06 |

These four events all occurred *together with the target* on those specific three – out of 22 – days that experienced problems. Which kind of knowledge may be extracted from this? This way of processing the data has the benefit that it is possible to assign interpretations to the clusters, using knowledge from the operators and process experts about what actually happened on those particular times. Certain combinations of events (represented by clusters) would indicate to the experienced personnel that a problem of type X is about to come up and that pressing button Y and pulling lever Z is the appropriate remedy. After study of the reports and displays produced it is evident that the algorithm groups together events that behave similarly – in general yet with particular emphasis on the 22 times when the target is 1.

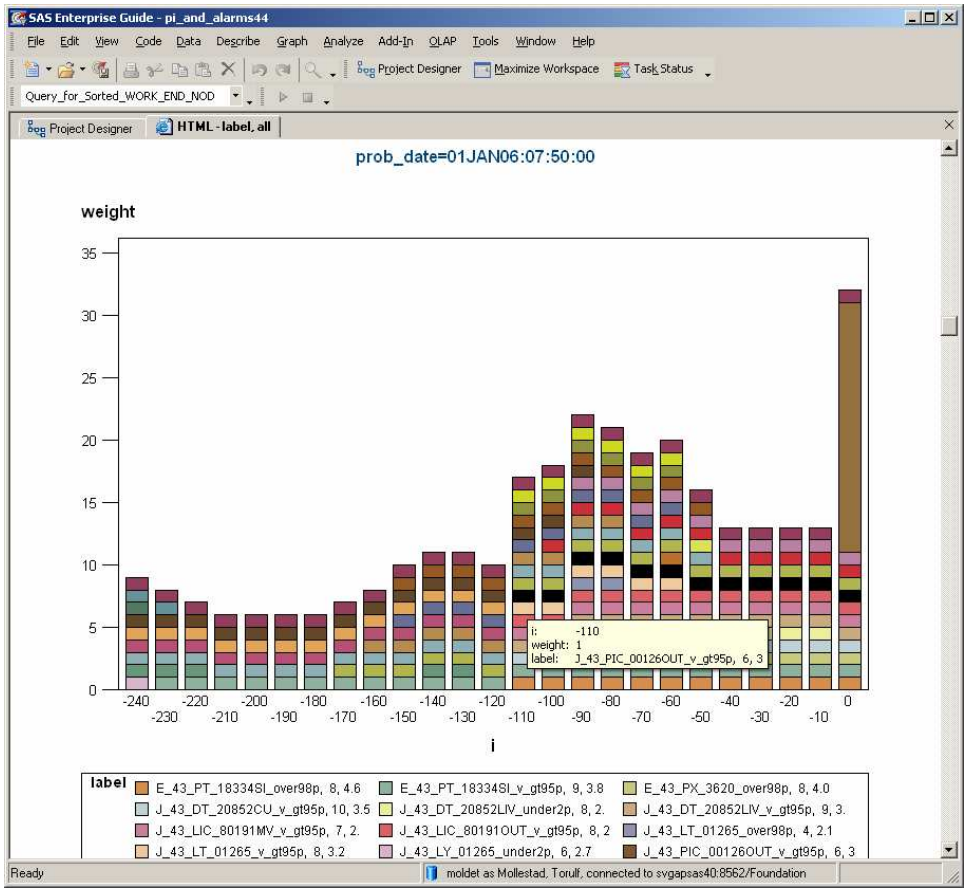
When studying reports on how the respective clusters and the events contained in them behaved wrt. the 22 problem times, a number of patterns were evident. It was clear from the data that clusters did indeed tend to gather around specific problem times. This is not to say, however, that particular dates are "assigned" uniquely to one cluster – several clusters may contribute to explaining a particular turn of events. We also noticed that some of the 22 problem times were more "active" than others, i.e. that very many of the events happen at these points. It should be considered whether these incidents represent normal running conditions at all – if not, they should be removed from the data at an earlier point in the process.

One way of presenting all clusters – and similarities between them - includes an interactive network depiction as shown below. The green colour represents the degree of correlation with the target, itself shown dark blue. The event tags coloured yellow are those that have a high lift (>3) wrt. the target. The graph may be turned around and investigated further; each node in the graph may be set up to cross reference a document, for instance a description of the tag or giving details of a particular problem.



Certain clusters may be discarded at this point, either because they are statistically uncorrelated to the target (low correlation and low lift) or after careful consideration by the expert. The corresponding event tags are then removed (added to a list of variables to be removed that is maintained as a macro variable in the project code) the process is rerun and the development proceeds in iterations. At the point when no further variables may be safely, and with certainty, removed, the clusters of interest (i.e. the event tags contained in those clusters) should be exported to SAS Enterprise Miner for association rule analysis.

For each time of trouble, a full profile of what happened before the problem may be presented to the business experts. For instance, the graph below shows the events that occurred within 240 minutes (four hours) of the problem that occurred on Jan 1st, 07.50 in the morning. The horizontal axis shows number of minutes before the problem occurs – the column to the very right thus represents the exact time that the problem struck, i.e. at 07.50 (the tall box to the right represents the target, whereas each of the smaller boxes signify events happening before). The selected tag (J_43_PIC_00126OUT_v_gt95p) is marked black with a yellow pop-up box indicating that this event first appeared 110 minutes before the problem occurred (by the strict definition of the target), and persisted throughout until the problem struck. This phenomenon might be interesting – or it might not. Only removal of information that is certainly of no interest and then careful scrutiny of the remaining patterns may reveal the all-important rules that may be used to warn of future problems in this highly complex process.




We found that there were a number of different profiles of problems and the processes that lead up to them, the significant event tags may be different from time to time. Yet, between subsets of the 22 problem profiles (corresponding to clusters computed in the last section), there were similarities to be found. This phenomenon suggests that there are certain *classes* of problems that may occur, and scrutiny of these would help in cataloging the potential faults that typically occur and to define action points to counter each one of them. It is our contention that understanding the properties of these different groups and the event tags that are most specific to them represents a great potential for understanding the overall

process, including which types/classes of problems that may – and do – happen to the low pressure separator on the J platform.

**ASSOCIATION RULES MINING**

In this part of the process (done using SAS Enterprise Miner), association rules were generated according to two schema, namely single item (single antecedent or A=>X) and double item (double antecedent or A&B=>X) – a large number of rules were generated in the process, but only rules that mapped into the target, i.e. of the form "A=>Target"and "A&B=>Target" were selected and reported on. Rules with three or four items on the left side may also be computed by Enterprise Miner, but this was not done at this point. It is, moreover, important to notice that the association rules algorithm will typically also extract rules that map into the items A and B themselves, thus potentially going back in time to the earliest possible indications of a problem.

**POSTPROCESSING AND VISUALIZATION**

In this part of the process, the rules generated are imported back into the project. As mentioned above, one- and two item rules that mapped onto the target event were extracted using SAS Enterprise Miner. Important metrics were computed for each rule, including its confidence, support and lift (the latter being the factor that the probability of the target increases by, given that the event on the left side of the rule occurred). The highest lift for a single item rule was 6.36, i.e. the immediate probability of the target increases by that factor at the moment when the left side event is observed. The table below shows the lift, the number of true positives, number of false positives and the total number of positives, respectively for some of the events. For instance, the first event in the list (J_43_TT_01461_under2p – corresponding to the rule J_43_TT_01461_under2p ==> low_water_closed ) happened 90 times in total, and 6 out of 22 times when the target happened, this event happened immediately preceding it. The conditioned probability 6/90 is 6.36 times bigger than the unconditioned (uninformed) probability of the target event, namely 22/2100.

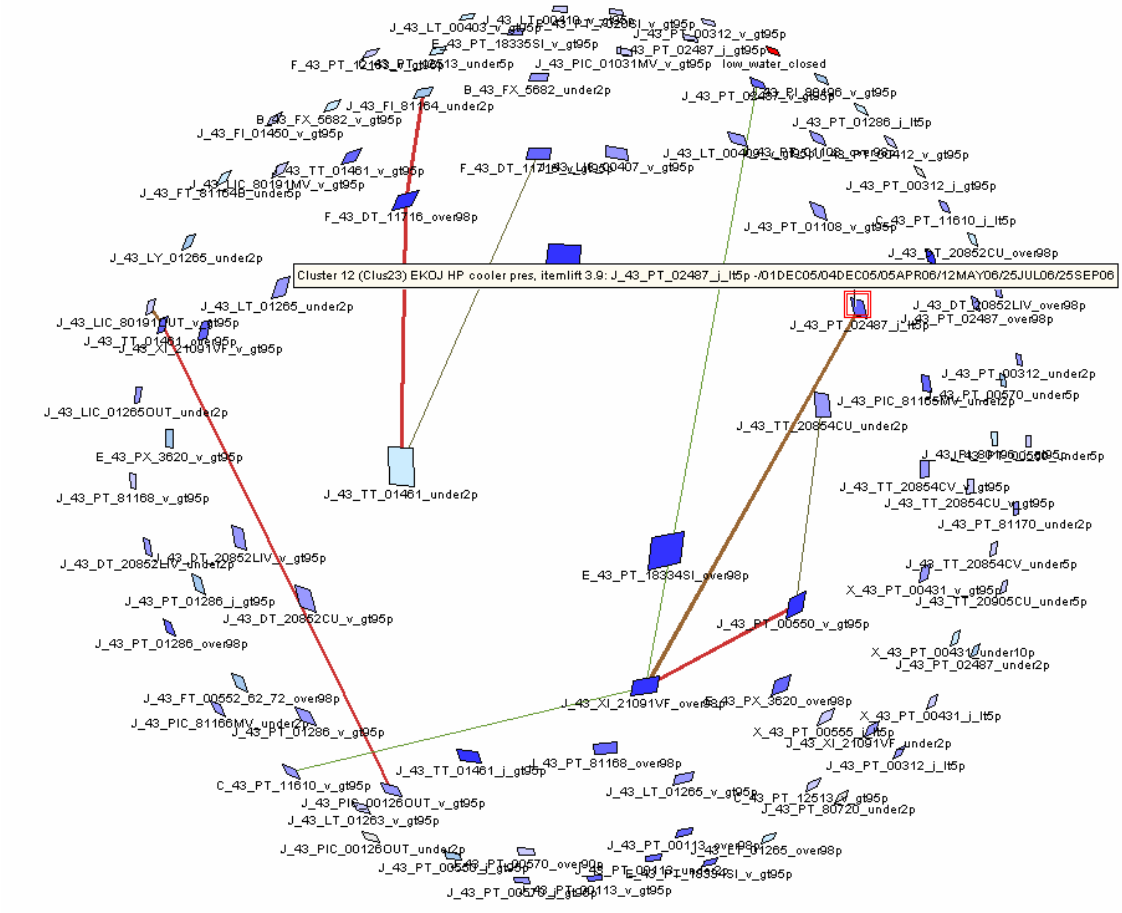
| | | | | |
|---|---|---|---|---|
| J_43_TT_01461_under2p | 6.3576 | 6 | 16 | 90 |
| J_43_TT_01461_j_lt5p | 5.2980 | 9 | 13 | 162 |
| J_43_XI_21091VF_over98p | 5.0581 | 8 | 14 | 155 |
| J_43_PT_00550_v_gt95p | 4.9303 | 6 | 16 | 116 |

The following table shows the corresponding table for *two item* left side rules. Notice that the lift in these cases is much higher. For instance, the lift of the first rule below suggests that the belief that a problem is coming up should increase by a factor of 38.5 if the combination of (two) events shown in the left side occurs. Computing 1.08*38.5 yields 41.67[1] – the updated probability of a problem conditioned on the evidence (the confidence of the pattern - the variable "Conf" below). So, given that the two left side events occur, there more than 40 percent chance that a problem is under way – based *on this piece of evidence only*. If other rules are matched at the same time, the probability should be updated accordingly.

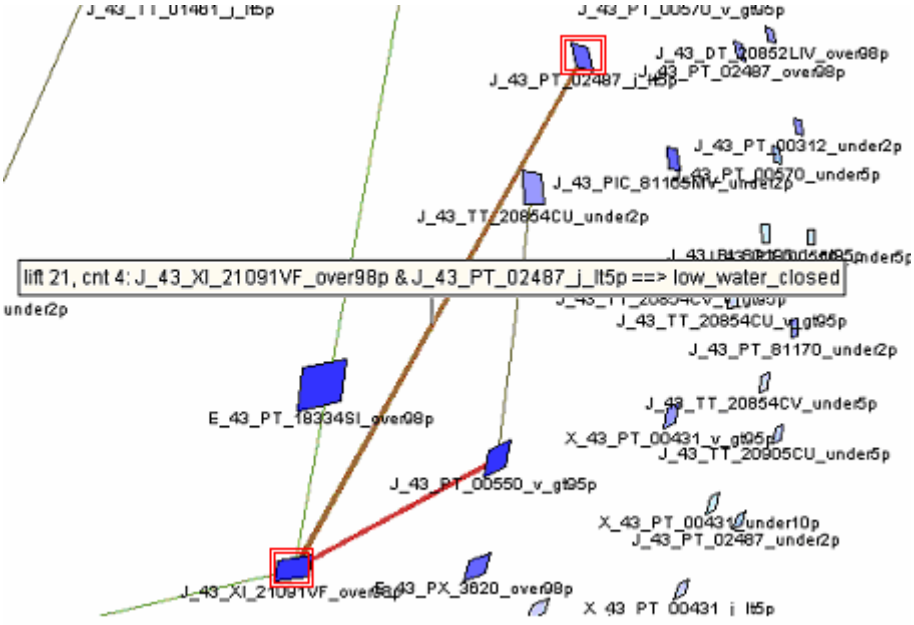
| RULE | EXP_CONF | CONF | SUPPORT | LIFT |
|---|---|---|---|---|
| J_43_XI_21091VF_over98p & J_43_PT_02487_over98p ==> low_water_closed | 1.08 | 41.67 | 0.12 | 38.56 |
| J_43_PT_01286_over98p & J_43_PT_00550_under5p ==> low_water_closed | 1.08 | 40.00 | 0.10 | 37.02 |
| J_43_PT_02487_under2p & J_43_PT_02487_over98p ==> low_water_closed | 1.08 | 40.00 | 0.05 | 37.02 |
| J_43_XI_21091VF_over98p & J_43_PT_01286_over98p ==> low_water_closed | 1.08 | 35.71 | 0.12 | 33.05 |
| J_43_PT_01286_over98p & J_43_PT_00570_under5p ==> low_water_closed | 1.08 | 33.33 | 0.07 | 30.85 |
| J_43_XI_21091VF_under2p & E_43_PT_18334SI_over98p ==> low_water_closed | 1.08 | 33.33 | 0.07 | 30.85 |
| J_43_PT_01286_over98p & J_43_LIC_00407_v_gt95p ==> low_water_closed | 1.08 | 28.57 | 0.05 | 26.44 |
| J_43_TT_20854CU_under2p & J_43_PT_02487_over98p ==> low_water_closed | 1.08 | 27.78 | 0.12 | 25.71 |
| J_43_PT_02487_over98p & J_43_PT_00312_under2p ==> low_water_closed | 1.08 | 27.27 | 0.07 | 25.24 |
| J_43_XI_21091VF_over98p & J_43_PT_01286_j_lt5p ==> low_water_closed | 1.08 | 26.67 | 0.10 | 24.68 |

[1] Remember the uninformed probability of the target (expected confidence) of 1.08

The generated rules may be displayed in a graph that incorporates information from both tables above, showing both the important characteristics of the single item rules, and the improved decision capabilities arising from combining events into two item rules. The first figure below shows the properties of the event pointed to, i.e. event J_43_PT_02487_j_lt5p (marked in red - an abrupt negative jump on that Pi tag). This single event (and the corresponding one item rule) has a lift of 3.9. Now, assuming that event J_43_PT_02487_j_lt5p is interesting, exploration from this starting point may be endeavored.



The event J_43_PT_02487_j_lt5p is connected by an arc to event J_43_XI_21091_over98p (marked in red at the bottom of the following figure – the values reaching a very high level) which on its own has a lift of 4.7 (verified by just passing the cursor over the node). However, the brown line between them represents the combination rule – which itself has a lift of 21, i.e. the combination of the two events sends a much stronger signal wrt. the probability of an upcoming problem.




As described above, the process of knowledge extraction should be performed in iterations and in close cooperation with domain experts. When event tags that are irrelevant to the target have been removed, the search for rules may continue, and the rules should be graded according to their interestingness. Rules that are regarded as interesting and that have a high statistical impact on the target may in turn be put into production, in a system as described in the following.

#### A PRODUCTION SYSTEM USING THE EXTRACTED KNOWLEDGE

The set of selected and verified rules/sequences constitutes a model for an early warning system on the low-pressure separator and its problems. Such a model may be put into production (having designed appropriate user interfaces) in several ways, including;

- Monitoring based on single patterns, simply alerting when events occur that satisfy the left sides of rules. This corresponds to the way alerts are being tackled today, with the added knowledge of the (historically derived) statistical strength of each signal
- Monitoring based on single patterns, taking into consideration partial satisfaction of left sides of rules – gives itself an alert level, yet warns which is the most significant future indication of a fault. In other words – given the current state – what are the events that would most strongly influence the probability of an immediate problem? The infrastructure for doing this is available.
- Monitoring based on combining signals from different rules into one unified risk score (which may itself be seen as a trend). The relative strength of different rules firing/not firing in a given situation is contracted into one single unified risk score. Several methods may be applied, including Bayesian networks or Dynamic Bayesian networks.
- Like the above, yet also providing decision support of which symptoms to focus on, how similar situations have been handled in the past and a suggestion wrt. how to best handle the problem, in other words a decision support system.

The models have not been set into production at this point, but the topic of setting association/sequence

models from petroleum data into production has been studied in the master's thesis work of two NTNU students under the supervision of the first author of this document. The result of this work is promising and has been made available to ConocoPhillips.

## SUMMARY AND FUTURE WORK

This project has investigated the potential in using event-based knowledge to diagnose a complex industrial process. The result is a framework that may ultimately be used to avoid production problems or to minimize consequences of such problems. The framework should be used in an iterative process that, in close cooperation with domain experts, focuses on revealing important and interesting rules and patterns. There are three main benefits from this work, namely:

- Providing new ways of looking at the data, thereby aiding the expert user to develop his/her knowledge and understanding of the process, and a basis for learning for novice users.
- Finding relationships that may be formalised as rules that may in turn be put into production, providing a continuous monitoring of the process and warning of problems coming up. This would also ensure that the operator may focus on the important events rather than considering the overflow of information that is currently written into the event log.
- Using extracted rules and relationships to provide decision support, reducing dependency on tacit knowledge and yielding organizational learning from the process.

### EXPERIENCES

The methodology described in the earlier sections has been specified and implemented in full.

- A set of rules has been found, each member in itself describing a strong relationship to the chosen target variable. When rules combine, even stronger evidence results, and the nature of the combination may be used to give an overall probability score – an estimated probability of the target problem coming up.
- The rules found and presented were intuitive/logical in the view of the process expert. As expected, a number of rules were found to be fairly trivial, but interesting relationships, deserving further investigation, were also uncovered.
- A large number of rules were concentrated around event tags from the J platform itself, whereas we have reason to believe that adding more data from earlier in the process would be benefitial. Only a few Pi tags from other platforms upstream to the separator had been included in the input data set, and information from these processes do therefore occur only infrequently in the rule set. It seems now that such rules would potentially be of great interest and that time series data from upstream to EKOJ needs to be included more liberally.
- The expert judged some of the rules as reflecting operator's reactions to the (problem) situation, i.e. representing consequences of the problem rather than indications of an upcoming problem. What this means is essentially that the target variable is wrongly defined, possibly pointing to time points *inside* the development of the problem, plausibly even after the operator has taken the appropriate steps to avoid a crisis. If this is the case, the problem profiles would show a lot of events that are really consequences of the problem or of the corrective measures, rather than indications of an upcoming undesirable situation.
- The algorithm is successful at picking out rules that describe the overall situations. It was commented that a number of the rules presented would seem to be connected with maintenance operations routinely done at regular intervals. When these operations are performed, the water level goes down and the target event will necessarily be a consequence. This indicates that the input data set used included time periods that were not representative of normal operation.

It is our belief that the approach is valid, and that the method may be applied to a great number of challenges in a spectrum of different industries; we are considering other areas where the method may be applied successfully. However, a most important criterion for the ultimate success of such an undertaking as this will always be that domain experts take a level of ownership to the process, and that access to them is ensured at regular intervals throughout the course of the project.

In this work, emphasis has been put on ease of interpretation to operators and domain experts, and a strengthened coupling to diagnostics and decision support will provide actionable knowledge to the needy at the right time.

### ISSUES FOR FURTHER WORK

All through the project, the need for close involvement from domain expertise has been evident. The process needs to be performed in iterations, and we found a number of potential modifications that need to be considered at each iteration; some of these considerations are given below. The points describe ways of improving the precision of the current work, but are generally valid, i.e. they reflect general observations that will need to be addressed in any project of this nature.

- Continuous improvement of the precision of the target variable (this would potentially have a very high impact). The evidence suggests that the target variable may currently be defined to strike at a point too late into the development of the problematic situation. Careful reconsideration and redefinition of the target variable is in order so that it signals the earliest possible point in time where the process is about to enter a non-desirable state.
- Also with respect to the definition of the target, it is not vital that the target signals real (experienced) problems. Rather, a better target would signal *any undesirable situation*, including even those times where the problem was avoided. In other words, any "close shave" situation should ideally be included. The definition of the target variable should be made with this, more liberal, view in mind.
- The tags used as input to the process are for the most part selected from the J platform itself. It was uncovered that data from other platforms are in lack in the input, suggesting that an investigation into Pi for other potential input variables would be in order. In general, all (time series) variables that are deemed potentially relevant to the target should be included in the process.
- Insignificant event tags should be removed from the data. The domain expert, knowing the details of the physical process, could tell us which signals were trivial and therefore which tags could be immediately removed. This process should proceed, making sure that no interesting event tags are removed in error, and ultimately resulting in a set of tags/rules that are of use to the domain expert (the problem of irrelevant tags would however also be diminished by a more careful definition of the target event)
- The observations from data and comments from the expert seem to indicate that certain time periods that are included in the input data should not be. Careful scrutiny of which periods constitute abnormal situations should be done. Such periods will typically include shutdowns and recovery periods, but maybe also times when smaller and routine operations are performed.

## REFERENCES


- Joacim Lunewski Christensen: *A Framework for Discovering Interesting Rules from Event Sequences with the purpose of Prewarning Oil Production Problems*, M.Sc. thesis, Norwegian Institute of Science and Technology NTNU, 2007
- Per Kristian Helland: *A Survey of Combining Association Rules for Pre-Warning of Oil Production Problems*, M.Sc. thesis, Norwegian Institute of Science and Technology NTNU, 2007


## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:


Torulf Mollestad, Ph.D
SAS Institute Norway
Work Phone: +47 907 24867
E-mail: nortmo@nor.sas.com