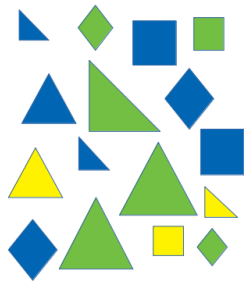


## Understanding and predicting faults in a paper mill production process, using exploratory pattern analytics (EPA)

Despite leading to a loss of output, and therefore being potentially very costly to the business, faults are still common in many production processes. This document will explain how faults can be analysed from sensor recordings using the exploratory pattern analytics (EPA) tool, in order to gain an insight into what happens during a fault and potentially help to predict them before they occur. The EPA tool can be applied to sensor recordings to find patterns describing typical situations in which an event occurs. For example, a pattern might be “Blade pressure” is Low AND “Vacuum strength” is High, in which case a fault is likely to have occurred; this pattern gives a concise summary of a typical situation in which there is a fault, making it easier to *predict* and to *understand* one particular type of problem.

To give a visual example of patterns, consider the shapes in the image below. Let’s imagine that each shape is a datapoint within a dataset. Every shape has several properties, which we can think of as variables and corresponding values.

Example data points:



Example properties:

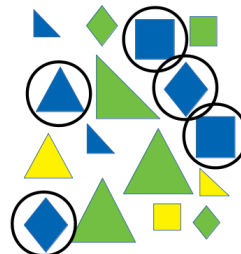
- Colour (Yellow, Green or Blue)
- Number of sides (1,2,3,4,5, etc.)
- Sides the same length (True, False)
- Angles the same size (True, False)

A pattern is a short description of a selection of data points. It contains several conditions, which combine to select a group of data points, also known as the ‘subgroup’ of the pattern. Below is an example pattern, with the subgroup indicated by black circles around the relevant data points.

Example pattern:

Colour = Blue  
AND  
Sides the same length = True

Example subgroup (in circles):



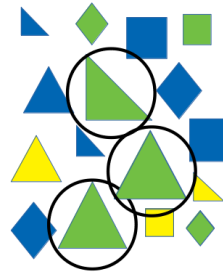
To decide which patterns are interesting, there is a ‘target’, which in this simple example might be the average size of the shapes that are selected by the pattern. The best pattern will be one that maximises the average size of the shapes, whilst including as many data points as possible. Please note that usually there is a trade-off, such that patterns with a smaller subgroup (which select fewer data points) are often able to have a higher average value, whilst patterns with a slightly smaller average are able to have a

larger subgroup. Below is an example of a pattern with a smaller subgroup but where the shapes have a higher average size.

Example pattern:

Colour = Green  
AND  
Number of sides = 3

Example subgroup (in circles):



The remainder of this document will provide a brief demonstration of running an analysis using the EPA tool. The example we will focus upon is a paper mill, which is a production process consisting of several connected steps that ultimately create large rolls of paper. One large sheet of paper is fed through the different machines in succession, which apply necessary processes like extracting moisture or compressing the paper. A common issue is that the sheet of paper breaks during processing, which temporarily halts production. When this happens, the paper must be carefully fed back through the sequence of machines before production can continue, leading to a loss of output before the process is running again. As noted by Ranjan et al.<sup>1,2</sup>, this is estimated to be a problem worth billions of dollars across the entire paper industry.

First, let us consider the data we will be using for our demonstration. The paper mill dataset (<http://bit.ly/2uCIJpG>)<sup>1,2</sup> is freely available after filling in a form, and contains 60 different sensor recordings from across the production process, with measurements taken every 2 minutes, for a period of a month. In addition, there is a variable which indicates whether there is a fault at each moment in time. Measurements were taken at 18398 time points (2-minute intervals), and there were faults at 124 of these points.

	A	B	C	D	E	F	G	H	I
1	time	Fault?	RSashScanAvg	CT#1 BLADE P	P4 CT#2 BLADE	Bleached GWD P	ShwerTemp	BlndStckFloTPD	C1 BW SPREAD
2	5/1/99 0:00	0	0.37666549	-4.5964348	-4.0957558	13.4976875	-0.1188297	-20.669883	0.00073248
3	5/1/99 0:02	0	0.47572049	-4.5425018	-4.0183588	16.2306585	-0.1287327	-18.758079	0.00073248
4	5/1/99 0:04	0	0.36384849	-4.6813938	-4.3531468	14.1279975	-0.1386357	-17.836632	0.01080348
5	5/1/99 0:06	0	0.30159049	-4.7589338	-4.0236118	13.1615665	-0.1481417	-18.517601	0.00207548
6	5/1/99 0:08	0	0.26557849	-4.7499278	-4.3331498	15.2673405	-0.1553137	-17.505913	0.00073248
7	5/1/99 0:10	0	0.38125349	-4.6117458	-4.0850718	14.1431955	-0.1625007	-16.494255	0.00073248
8	5/1/99 0:12	0	0.31332549	-4.5302098	-4.1209308	18.6814645	-0.1696717	-15.329521	0.00073248
9	5/1/99 0:14	0	0.39640149	-4.6990468	-4.0741928	21.3076545	-0.1768587	-13.937523	0.00073248
10	5/1/99 0:16	0	0.34268849	-4.5535058	-4.1888848	22.8920355	-0.1840457	-13.630058	0.00073248
11	5/1/99 0:18	0	0.45825249	-4.6395608	-4.2465468	23.1609565	-0.1912177	-11.910331	0.00073248
12	5/1/99 0:20	0	0.39375649	-4.8119978	-4.1164548	27.5508485	-0.1984047	-8.3857651	0.00073248
13	5/1/99 0:22	0	0.40494649	-4.6177988	-4.1874618	25.0356515	-0.2055757	-10.844688	0.00073248
14	5/1/99 0:24	0	0.46700049	-4.4560498	-4.3165398	30.2554085	-0.2127627	-5.3095011	-0.0093385
15	5/1/99 0:26	0	0.43343049	-4.4547488	-4.2242328	32.7386845	-0.2199497	-6.5157391	-0.0093385
16	5/1/99 0:28	0	0.46963149	-4.3980648	-4.0284598	29.8572755	-0.2271217	-6.7996141	-0.0194095
17	5/1/99 0:30	0	0.42656349	-4.3461838	-4.1815608	32.1860415	-0.2343087	-5.2274701	-0.0201755

<sup>1</sup> Chitta Ranjan (2020). *Understanding Deep Learning: Application in Rare Event Prediction*. Connaissance Publishing.  
<sup>2</sup> Chitta Ranjan, Mahendranath Reddy, Markku Mustonen, Kamran Paynabar, & Karim Pourak. (2019). Dataset: Rare Event Classification in Multivariate Time Series.

To use the EPA tool, the user will identify some variable of interest (called the ‘target’ variable) and will extract patterns that help to explain that variable. The goal of the EPA process will be to understand in what circumstances the target is extreme. With a numeric target, this means finding circumstances in which the value is exceptionally high (or exceptionally low) on average. For a non-numeric target, this means looking for circumstances when a particular value is especially likely to occur. In our case, the target variable will be the indicator of whether or not a fault has occurred, and we will search for circumstances in which a fault is especially likely.

Alongside the target variable, it is crucial to have other variables that potentially will help to explain the target variable. These should be measurements that naturally accompany the target, and variables that seem irrelevant to the target should be excluded. The paper mill dataset conveniently provides 60 relevant variables from across the production process.

In order to search for patterns that help to predict a fault before it happens, we construct a target variable that indicates whether a fault will begin within the next 10 minutes. Patterns found using this target will describe situations in which a fault is especially likely to occur soon, and this could help an expert gain a better understanding of when faults occur and why.

To understand the EPA tool better, it is worth knowing how patterns are chosen. Patterns identify a ‘subgroup’ of rows within a dataset. The subgroup associated with each pattern can be evaluated according to its size and how extreme the target value is (for example, how likely some event is within the subgroup compared to across the dataset overall). The formula for doing this evaluation is called a ‘quality function’, which will decide the relative importance of the subgroup being large compared to it having a particularly extreme target value. Additionally, it is possible to search for patterns with large subgroups versus searching for patterns with smaller subgroups.

The EPA tool is integrated within the [Di-Plast Data Analytics tool](#). It appears as one of the options on the left-hand menu within the Data Analytics tool, as shown in the screenshot below.

**Interreg North-West Europe Di-Plast**  
European Regional Development Fund

**Select a Module**

Choose one of the analytics options: ○

- Home
- Data Inspection
- Data Comparison
- Feature Selection
- Classifier
- Exploratory Pattern Analytics
- Final Data Report

**JADS** Jheronimus Academy of Data Science  
Bug reports and suggestions welcome

### Welcome to the Data Analytics Tool

This dashboard is developed by Jheronimus Academy of Data Science (JADS) and Osnabrück University (UOS). It is developed for the Di-Plast project. More information on the Di-Plast project can be found [here](#). For information on this dedicated tool, we advise reading our [wiki](#) page containing all the necessary information for installation and interpretation.

The Data Analytics tool consists of several modules that analyse parts of your dataset. It is of great importance that the data used in this tool is properly validated. For validating the data, we advise you to check our data validation tool that can be accessed [here](#).

**The tools are:**

- Data Inspection: Inspect your data to check if it is in the right format, and that the right variables are there.
- Use this module to compare the measurements of the same variable/sensor on two different datasets.
- Feature Selection: Reduce the size of your dataset by spotting unnecessary variables, but also by spotting the important ones!
- Classifier: If you have something you want to predict in your dataset, use this module to predict classes or numerical values. For example, predict if this product is of bad quality.
- Exploratory Pattern Analytics: Explore your dataset by finding interesting patterns in your data.
- Final Data Report: a downloadable HTML-form with all sorts of information about your dataset.

👉 Select a tool from the dropdown menu on the left.

See Disclaimer

Developed in the Di-Plast project in collaboration with:

**JADS** Jheronimus Academy of Data Science

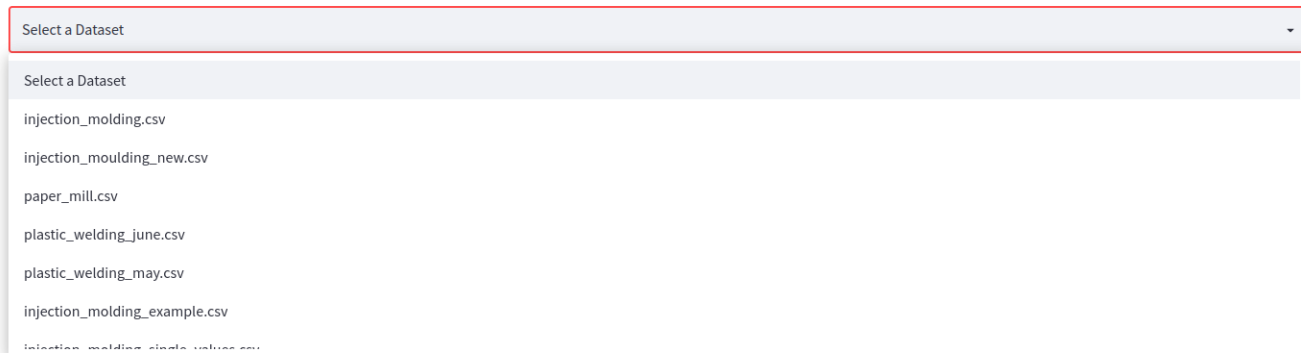
**UNIVERSITÄT OSNABRÜCK**

To perform the analysis, the first task is to load the data by selecting it from the drop-down menu in the EPA tool.

## Exploratory Pattern Analytics (EPA)

Welcome to the Exploratory Pattern Analytics (EPA) tool. This tool makes it possible to find interesting groups of data points within your data, which are described through simple patterns. More information, and guidelines for the tool (available through the link provided in the "Tool guideline and access" section), are provided in the [EPA tool page of the Di-Plast Wiki](#).

Which dataset do you want to view?



The screenshot shows a web interface for selecting a dataset. At the top, there is a question: "Which dataset do you want to view?". Below this is a dropdown menu with the text "Select a Dataset" and a downward arrow. The dropdown is open, showing a list of dataset names: "injection\_molding.csv", "injection\_moulding\_new.csv", "paper\_mill.csv", "plastic\_welding\_june.csv", "plastic\_welding\_may.csv", "injection\_molding\_example.csv", and "injection\_molding\_single\_value.csv".

With a dataset selected, the next important step is to choose a type of analysis to perform. The different possibilities are described within the tool, as can be seen in the screenshot below. For the current example, we will select “Event detection” since we are interested in faults, which can be treated as events. This will allow us to investigate the circumstances in which faults occur.

### Type of analysis

Depending on the type of analysis, different options for how to perform the analysis will be shown. Please select an option below.

*Classification* looks at what makes one class different from others. For example, distinguishing one particular product from other products, or distinguishing recycle from virgin material, or distinguishing one type of outcome for a process. More generally, this is possible when the target is a non-numeric variable. *High average* aims to find situations in which there is a high value for some numeric variable. For example, identifying circumstances in which a quality score, or a physical property, tends to be high. More generally, this is possible when the target is a numeric variable. *Event detection* tries to understand events in a recording over time. For example, looking for faults or quality issues which occur at specific times. This option is appropriate when there is a variable indicating when an event occurs.

Please select the option that best describes the analysis type

- Classification
- High average
- Event detection
- Other

Filtering the data before performing the analysis is also possible. This step is optional, and will depend on the data and the type of analysis being performed.

For data that consists of measurements over time, it is possible to select a specific time window to analyse; this might be interesting if there is additional knowledge (not included in the data) that the user wants to take advantage of, for example to focus on an individual manufacturing run when a particular product was being produced. Note that this will only be possible if the data consists of measurements over time. The next possibility is to choose data points based on the value of one of the variables. For example, this could be interesting when analysing a manufacturing process and there are moments when the machinery is not in use. Perhaps there is a variable measuring the power usage, and in this case it would be possible to filter out moments when the power usage is zero. The final possibility is useful when the dataset is large and the EPA tool is taking a long time to run. This makes it easy to obtain a smaller dataset by simply selecting data points that are evenly spaced throughout the dataset. It is only helpful when there is a problem with processing the full dataset.

Finally, it is important to note that filtering might not be desired at all; in this case, simply make sure that the tick-box labelled “Would you like to filter the data before analysis” is unticked.

## Filtering

Sometimes it is interesting to focus the analysis only on certain points in the data. To achieve this, filtering is possible. This step is entirely optional.

This step can also be used to reduce the size of the dataset if the analysis is taking too long to complete.

Would you like to filter the data before analysis?

## Time selection

Sometimes it is interesting to focus on a particular period within the data, for example because you are analysing a manufacturing process and want to focus on the period when production is happening rather than periods of rest. Here it is possible to provide a start and end time in order to select data to analyse. A variable from the dataset can be viewed to help with this choice (e.g. a variable representing the state of the system). By default, the entire dataset is selected.

Variable to visualise to help choosing a time period:

Variable to display

## Filter on variable

Here, it is possible to filter based on the value of a variable of choice. This could be any variable in the data. As an example, for a manufacturing process, there might be a variable that indicates whether the process was running or not, and it might be desirable to filter out all data points where the process was not in use. The range of values to keep must be provided. Other values will be removed before analysis.

Variable to use for filtering:

Choose a variable

## Filter to reduce dataset size

For larger datasets, the exploratory pattern analytics process can take a long time. If this happens, it is possible to remove data points in order to speed up the process. This is not recommended unless the EPA tool is running very slowly.

Data points that are evenly spaced throughout the dataset will be kept, and the rest will be discarded. For example, keeping 1 of every 4 points means that the 1st, 5th, 9th, 13th, etc. data points will be kept.

Interval:

1

1 out of every 1 data points will be kept.

Next, some key options for the pattern search can be provided. The target variable should be specified, and the target value for this variable should also be selected if using a non-numeric target. The quality function can be chosen to favour smaller or larger subgroups, and the minimum size can also be set. Finally, there is an option to remove patterns that select essentially the same subset of data as other patterns, basically removing duplicates in the results.

When performing an “Event detection” type of analysis, there is also an extra option to include moments leading up to an event. The user specifies a time period leading up to each event to also include in the analysis. If this option is used, then the analysis will also search for patterns that describe the moments leading up to an event. In our current example, the target variable of the data has already been pre-processed to include the 10 minutes leading up to a fault. For this reason, we choose not to use the “(Optionally) also include earlier time points that happened within” and “Unit of time” fields, meaning that no additional pre-processing is performed.

# Settings

Target variable:

Within\_10

Target value:

True

(Optionally) also include earlier time points that happened within (please specify number and unit):

0

Unit of time (leave blank if you do not want to include earlier time points):

Optionally choose columns to ignore (leave blank to use all columns):

Choose an option

Quality function:

Smaller subgroups

Minimum size for subgroups:

10

Suppress 'duplicate' subgroups that overlap with previous subgroups by more than:



In this analysis, patterns for smaller subgroups will be used. Smaller subgroups are potentially interesting because they could point to situations that then commonly lead into a fault (or pose a danger of developing into a fault). Of course, there are multiple reasons why a fault may occur, and these smaller subgroups give a hint about one particular situation (rather than all of the situations) that often leads to a fault.

After selecting parameters, the analysis will run automatically, and the best-performing patterns will be shown in a table.

## Top patterns

The table of results shows a list of the best patterns found, along with some measures of quality. Each pattern chooses up to three variables and includes a condition for each of those variables. These combine to select points within the dataset. The points selected by a pattern are called its 'subgroup'.

For nominal targets, the percentage of subgroup members that belong to the target class is shown. This number, along with the size of the subgroup, is used to calculate the quality score of the pattern. For extra information, the precision (what proportion of the points selected by the pattern in fact belong to the target group), the recall (how much of the target group is selected by the pattern), and the F1-score (a combination of precision and recall) are provided as extra quality measures. Estimated 5% and 95% confidence intervals are shown for precision, recall and F-1.

For numeric targets, the average value for the target variable is shown. This number, along with the size of the subgroup, is used to calculate the quality score of the pattern. For extra information, the "Hedge's G" measure is also shown. This gives an indication of how large the difference is between the points selected by the pattern and the rest of the dataset. Larger numbers indicate a greater difference. Estimated 5% and 95% confidence intervals are shown for Hedge's G.

	id	Pattern	% of Subgroup that Are Target Class	Size	Quality Score	Precision (lower CI)	Precision (upper CI)	Recall (lower CI)	Recall (upper CI)	F-1 Score (lower CI)	F-1 Score (upper CI)
50	*A*	P4 CT#2 BLADE PSI < -6.83 AND -5.44 <= COUCH VAC AND RS BW SPREAD MD < 0.91	27.5	367	4.56	0.116	0.56	0.0459	0.255	0.0596	0.334
59	*B*	P4 CT#2 BLADE PSI < -6.83 AND -0.00 <= HorzSlcPos < 0.00 AND CouchLoVac < 0.25	48.5	103	4.54	0	1	0	0.156	0.0456	0.261
15	*C*	P4 CT#2 BLADE PSI < -6.83 AND -5.44 <= COUCH VAC AND ShwerTemp < 0.33	36.7	215	4.84	0.0893	0.802	0.0226	0.217	0.0448	0.317
22	*D*	P4 CT#2 BLADE PSI < -6.83 AND ShwerTemp < 0.33 AND -28.95 <= UpprHdTmpRL	34.5	229	4.65	0.0893	0.801	0.0226	0.217	0.0439	0.313
95	*E*	P4 CT#2 BLADE PSI < -6.83 AND ShwerTemp < 0.33 AND -1190.89 <= MachSpd	32.1	246	4.45	0.0821	0.671	0.0226	0.217	0.0438	0.306
60	*F*	P4 CT#2 BLADE PSI < -6.83 AND ShwerTemp < 0.33 AND -4.87 <= RS BW SCAN AVG	33.5	233	4.54	0.0584	0.751	0.0203	0.217	0.0437	0.305
98	*G*	P4 CT#2 BLADE PSI < -6.83 AND -1.51 <= RSashScanAvg < 0.77 AND -1.56 <= 4DryrDraw < 0.10	43	128	4.43	0	0.859	0	0.152	0.0423	0.252
9	*H*	P4 CT#2 BLADE PSI < -6.83 AND -0.00 <= HorzSlcPos < 0.00 AND CT#1 BLADE PSI < -5.96	81.4	43	5.08	0	1	0	0.115	0.0422	0.201
8	*I*	P4 CT#2 BLADE PSI < -6.83 AND -1.56 <= 4DryrDraw < 0.10 AND -5.44 <= COUCH VAC	48.5	130	5.09	0.12	1	0.0164	0.163	0.0416	0.265
65	*J*	P4 CT#2 BLADE PSI < -6.83 AND -1.56 <= 4DryrDraw < 0.10 AND -179.96 <= FanPmpSpd	39.6	159	4.52	0.0881	0.78	0.0164	0.163	0.0416	0.263

This table shows various useful pieces of information, including the size of the subgroups found and estimated confidence intervals on various performance metrics. Clicking on the table will expand it to show all the columns. Looking at this table, we find that one of the best patterns (labelled \*A\* in the table) is:

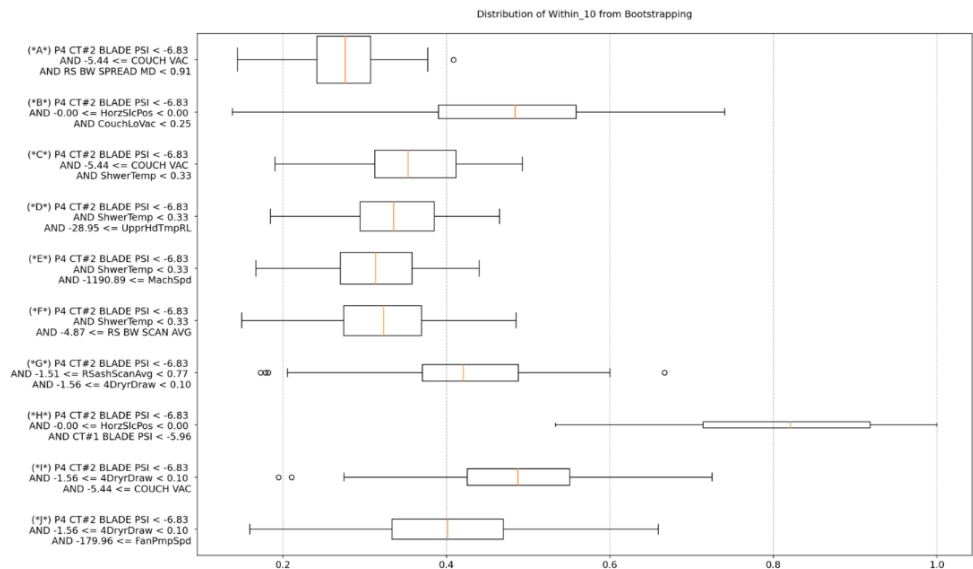
- P4 CT#2 BLADE PSI < -6.83
- -5.44 <= COUCH VAC
- RS BW SPREAD MD < 0.91

This suggests, for example, that the blade pressure is low and the couch vacuum is high typically when a fault is likely to occur. An expert could investigate this particular combination of sensor measurements and ask why they are so likely to precede a fault, whether there is a causal relationship, and whether there is any way to reduce the likelihood of a fault developing.

To accompany the table of patterns, there are two extra visualisations that make it easier to compare subgroups. The first shows the expected variability for the target value, meaning how much it changes across different samples of sensor measurements. This is depicted through boxes in a box plot, with wider boxes implying greater variability.

### Plotting the distribution of the target value

This visualisation shows the expected variability for the target value, meaning how much it changes across different samples of measurements. For nominal target variables, 'target value' means the proportion of subgroup belonging to the target class, and for nominal target variables, it means the average value of the target variable. This is depicted through boxes in a box plot, with wider boxes in the x-direction implying greater variability. The orange line shows the target value on average across different samples. How many points are selected by each pattern is also shown (i.e., its size), with thicker/taller boxes in the vertical direction meaning that a pattern selects a greater number of points on average.



In the second visualisation, patterns are connected to each other by how much their members overlap. If two patterns select similar subsets of data (have similar subgroups), then they have a strong link between them and appear closer together. Overall, this visualisation takes the form of a network diagram.

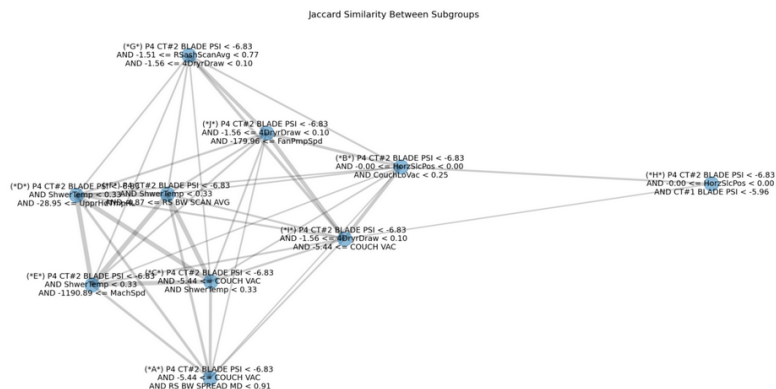
### Overlap between patterns

After discovering patterns, it is possible that two different patterns might essentially be different ways of describing the same points in the data. In this case, it might be useful to know that they are closely related.

On the other hand, patterns might have an extreme target value for different reasons. If two patterns select quite different subgroups, then there might be different reasons they are interesting, and it could be worthwhile to investigate them both separately in greater detail.

In this visualisation, patterns are connected to each other by how much their subgroups overlap. If two patterns select similar subsets of data (they have similar subgroups), then they have a strong link between them and appear closer together. Overall, this visualisation takes the form of a network diagram.

Only draw edges when overlap is greater than:





At this point, there may be patterns that are particularly interesting. Knowledge of the problem domain is especially important in interpreting the results and deciding which patterns are interesting enough to investigate further. The EPA tool makes it possible to examine the most interesting patterns in more detail. A visualisation is provided that compares subgroup members to non-members for one specific pattern. The target variable, the variables used to define the pattern (selector variables), and additional variables that are most clearly different between members and non-members are shown. This makes it possible to see additional information about the pattern, and understand more about the circumstances in which the pattern occurs.

## Focus on a specific pattern/subgroup

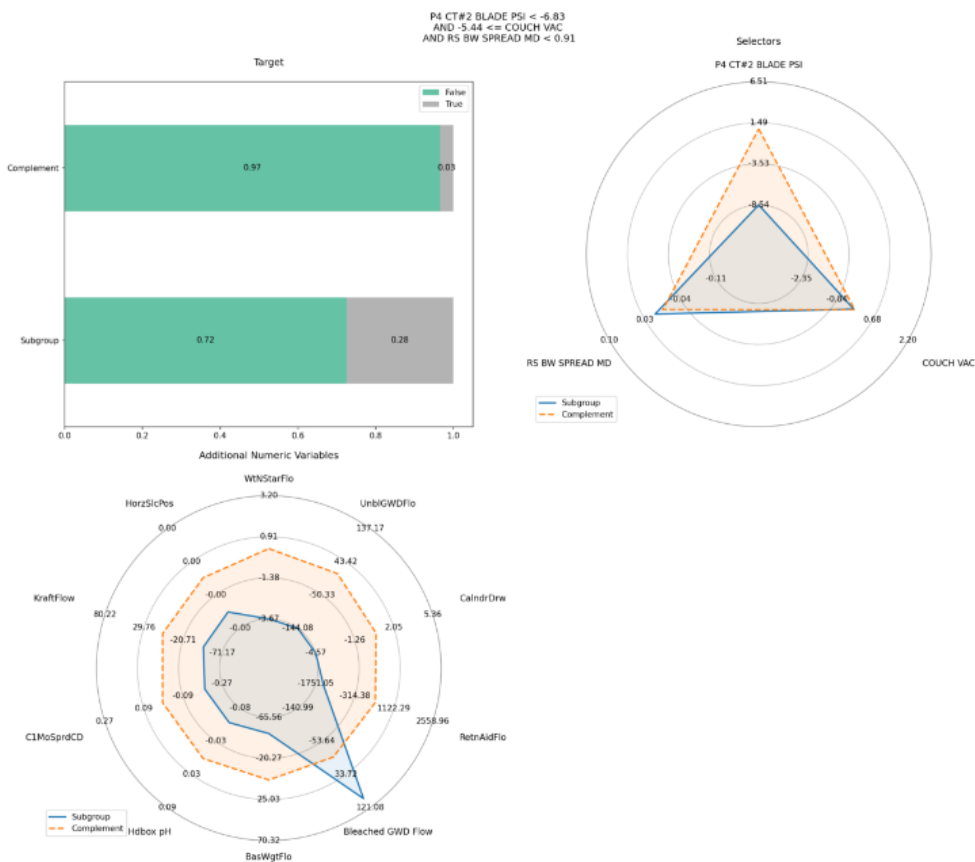
At this point, there may be patterns that are particularly interesting. The EPA tool makes it possible to examine these in more detail. This visualisation compares subgroup members (points selected by the pattern) to non-members (these non-members are also known as the 'complement') for one specific pattern.

The target variable, the variables used to define the pattern (selector variables), and additional variables that are most clearly different between members and non-members are shown. These respectively appear in the top-left, top-right and bottom panels of the visualisation. This makes it possible to see additional information about the pattern, and understand more about the circumstances in which the pattern occurs.

In the top-left, the distribution of values for the target variable is shown. For nominal targets, a different set of horizontal boxes is used for the subgroup and the complement. For numeric targets, the subgroup is indicated by a solid blue line and the complement is indicated by a dashed orange line. In the remaining panels, the subgroup is also indicated by a solid blue line and the complement by a dashed orange line.

Pattern to focus on:

(\*A\*) P4 CT#2 BLADE PSI < -6.83 AND -5.44 <= COUCH VAC AND RS BW SPREAD MD < 0.91



Finally, since the paper mill data comes from a process that happens over time, we can focus on particular moments at which a pattern occurs, to see what happens to different sensor recordings before, during, and after. After selecting a single pattern, the user of the EPA tool can then select a particular moment when the pattern occurs, from a drop-down list. The target variable is shown, along with the other variables that are most clearly different between subgroup members and non-members. The moment at which the pattern occurs is indicated by a red rectangle in the background.

## Specific subgroup members

Finally, if the data comes from a process that happens over time, we can focus on particular moments at which a pattern occurs, to see what happens to different variables before, during, and after. After selecting a single pattern, you can now select a particular moment when the pattern occurs, from the drop-down list below. The target variable is shown, along with the other variables that are most clearly different between subgroup members and non-members. The moment at which the pattern occurs is indicated by a red rectangle in the background.

Subgroup member to inspect:

1999-05-28 06:50:00

Also display earlier time points that happened within:

20

Also display later time points that happened within:

10

Unit of time:

Minutes



The examples above should show by now that the EPA tool can find various types of patterns. What these patterns do is inform the user about what the data can tell us about a particular question, in a format that is easy to interpret and understand. In this way, the patterns provide information about what is contained within the data, and they are subject to the limitations of the data. Therefore, it is important that the EPA tool is used in combination with expertise about the subject matter, i.e., experts who understand the sensor recordings and the manufacturing process, who can interpret the outputs of the tool.

One additional thing to note when performing the analysis is that the EPA tool supports an iterative approach. This means that the EPA process can be run to obtain initial patterns, which then might suggest changes like adding or removing variables or refining the search parameters, before re-running the process. When taking this iterative approach, it is best to involve someone who understands the phenomenon being studied and who will be able to make use of the insights provided by the tool.

To summarise: in this demonstration, we have considered the problem of faults within a production process, something that is costly to many manufacturing businesses. The EPA tool was used to identify patterns that help to explain and predict faults within a production process. These patterns provide useful information to an expert who might be trying to diagnose the faults and reduce the likelihood of them occurring. This demonstration has also shown that there are different ways to search for patterns, for example it is possible to look for patterns identifying a larger or smaller subgroup, and it is possible to focus either on moments when a fault is occurring or moments that precede (and potentially predict) a fault. Ultimately, the EPA tool provides insights from the data, to support an expert who is trying to reduce the impact of faults; if the expert is successful, this would imply cost savings to the manufacturer.