# Executive Summary

Most of the water used today in Scotland is collected and stored in man-made lakes called reservoirs. These are made by building a dam across the upper part of a river, allowing the water to collect, ensuring that even if it does not rain there will still be water available.

Before the water in a reservoir can be drank, it first needs to be cleaned to remove silt and any biological contamination present in the reservoir.

The water is passed through mesh screens to remove large debris such as leaves, weeds, and sticks. Clarification is then used to remove impurities in the water- this include coagulation and flocculation, as well as sedimentation which remove any trace chemicals that are undesirable in drinking water. Any impurities that make it through the clarification stage are subsequently removed through filtration.

The pH of the water needs to be controlled after this step, as is water is too acidic it can corrode the metal pipes which it is pumped through, and if it is too alkaline it will leave deposits inside the pipes, leading to blockages.

Finally, and the most important stage for our study, the water needs to be disinfected to ensure that any water-borne diseases are eliminated and that the drinking water meets the Water (Scotland) Act 1980. As failure to provide wholesome water is a criminal offence it is vital that Scottish Water is able to minimise the number of failures that impact the service reservoir’s water quality by predicting and preventing any failure either before they occur or as quickly as possible.

To this end we propose to develop a Bayesian Belief Network that will highlight the key factors affecting the probability of one or more failures occurring in the quality of the drinking water supply, as well as to enable the likelihood of such a failure occurring based on known data about other parts of the water treatment and service reservoir system.

Ultimately the model was able to reveal the main factors impacting the water quality and those which contribute the most to a failure occurring in the Prescribed Concentration or Value of the water. It also gave us insight into areas in the water treatment process that would benefit from deeper explanation, such as the fact the North region has a higher failure rate than the rest of the Scottish water supply, and that certain reservoirs in the reservoir cascade contribute more to a failure in the PCA than other reservoirs. Finally, the model allows for predictions to be made about the probability of a failure occurring based off known facts about the other factors, as the evidence for these factors can be set based off real world observations and provide an up-to-date probability distribution of the odds of a failure occurring.

# Modelling Process

## Data Pre-Processing

Before the data could be used to create a Bayesian Belief Network it was necessary to carry out some data processing. As there were many continuous variables these had to be discretised to allow for them to be used. By making use of GeNie, the columns were analysed for all the values present and then discretised, with the reasoning behind the bins selected explained below:

* Srs\_free\_chlorine – this must be below 1mg per Litre to be considered safe so split into <=1mg and >1mg bins
* Srs\_coliforms – this should be 0 to be considered safe so anything above 0 is flagged as dangerous
* Srs\_icc – this is the number of intact cells in the standard reservoir. A number above 10,000 cells indicates a significant value.
* Srs\_tcc- similar to the above a value above 10,000 is significant
* Srs\_total\_chlorine – this must be below 2mg per Litre to be considered safe
* Srs\_voloutcap – this was split into standard readings (~90% of values) and high readings
* Srs.total\_risk\_factor – this was split into 3 equally wide bins for low, medium, and high-risk factors
* Srs\_average.daily\_flow – similar to voloutcap this was split into standard and high readings
* Srs\_current.storage\_retention – this is split into below 100 and above 100
* Srs\_aow – this is the age of water in the storage reservoir. For testing purposes, it is split into <=7 days, 7 to 21 days, and >= 21 days
* Srs\_stor\_vol – similar to voloutcap, split into standard and high readings
* Srs\_stor – similar to aow and split into same time brackets
* Srs\_colony\_22 – the number of CFU allowed in 1 mL of water at 22⁰C- should be less than 100 to be classed as safe, higher than 100 classed as danger
* Srs\_colony\_37 – similar but at 37⁰C and only 10 CFU per mL allowed
* Wtw\_free\_chlorine – identical to srs\_free\_chlorine
* Wtw\_total\_chlorine - identical to srs\_total\_chlorine
* Wtw\_ colony\_22 – identical to srs\_colony\_22
* Wtw\_colony\_37 – identical to srs\_colony\_37
* Wtw\_coliform - identical to srs\_coliform

Missing values were filled in for each column, either with the average value present for that column, or for variables where a bad result could have a significant impact on the water quality they were flagged as significant to ensure that they could be further investigated if testing was carried out again.

## Variable Selection

Initially wtw\_mains.power was excluded from the variables as it has a consistent value of 1 for all readings and therefore should have no effect on the other variables. Srs\_fails\_in\_previous year was also dropped as the cumulative total fails were also available and this would reduce the risk of multicollinearity. The srs\_sample\_month was also dropped as to model any time-based dependencies would require a Dynamic Bayesian Network which is beyond the scope of this assignment.

After dropping these more obvious variables we then used the steps module in R to systematically remove variables one at a time by comparing models with different combinations of variables and rejecting any that failed the F test. After carrying this out we were left with the following variables which we decided to proceed with:

* srs\_pcv.failure
* srs\_bacto.fails.risk.factor
* srs\_ecoli
* wtw\_phosphate.dosing
* srs\_icc
* srs\_colony22
* wtw\_manganese.removal
* srs\_fails\_in\_any\_previous\_year
* wtw\_sourcetype
* wtw\_raw.water.pre.treatment
* wtw\_floctype
* srs\_coliform
* wtw\_ozonation
* srs\_ooa
* srs\_region
* wtw\_membrane
* wtw\_chloramination
* srs\_total.chlorine
* srs\_free.chlorine
* srs\_aow
* wtw\_slow.sand.filters
* wtw\_ph.adjustment
* wtw\_raw.water.storage
* srs\_voloutcap
* wtw\_floc
* wtw\_rgf
* srs\_average.daily.flow
* srs\_stor srs\_current.storage.retention
* srs\_stor\_vol
* wtw\_total.chlorine
* srs\_sr.chain.risk.factor
* srs\_condition.risk.factor
* wtw\_clarification

## Structure Learning

To prepare the dataset for structure learning the categorical variables were factorised, while the variables containing numbers were converted from int type to numerical type.

Several algorithms were used to produce the models. These include:

* Grow-Shrink
* Incremental Association
* Fast Incremental Association
* Interleaved Incremental Association
* Max-Min Parents and Children

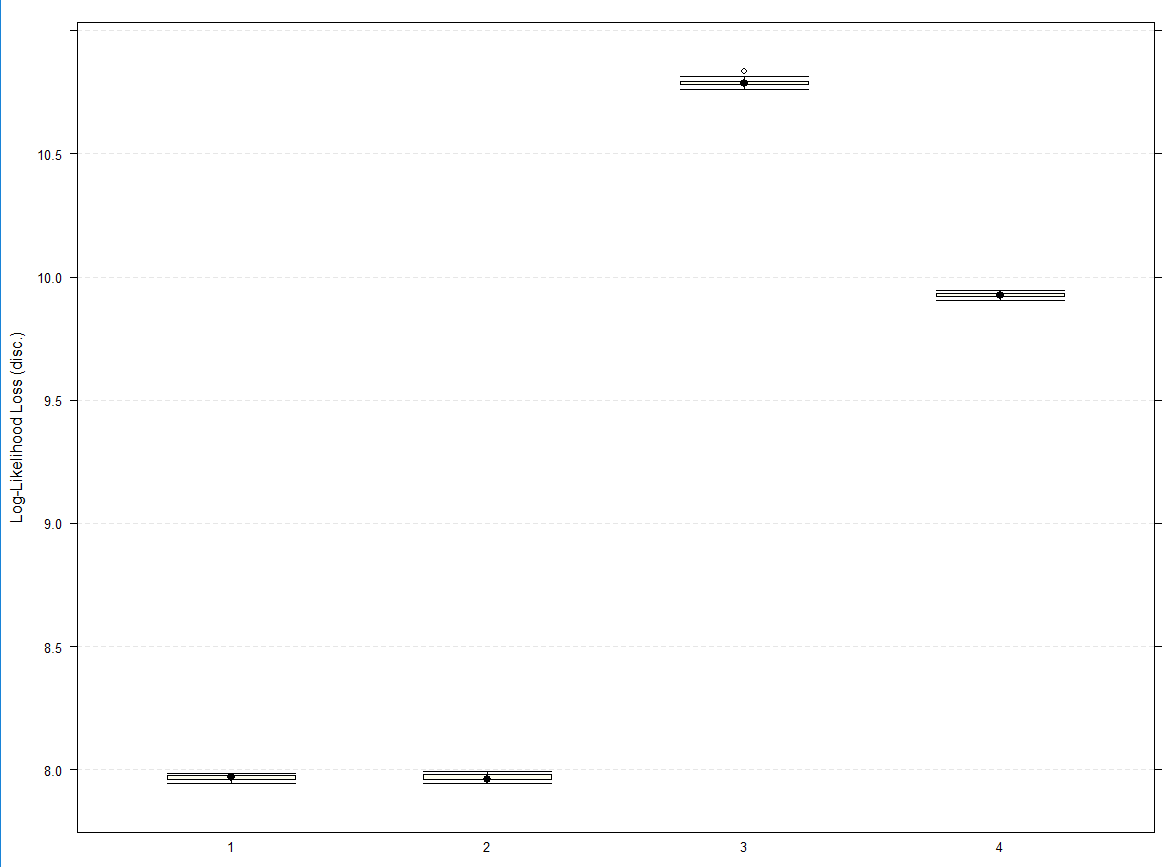
The reasons we learn the structure are twofold: To help us better understand the model itself, by allowing us to learn the dependences between its variables, as well as to estimate the statistics of the underlying distribution and infer further information about the model.

The main types of structured learning approaches are constraint-based learning, which performs independent statistical tests to ascertain a set of dependencies that exist in the data, and score-based learning which scores each node based on the data and searches for a network structure to maximise the score.

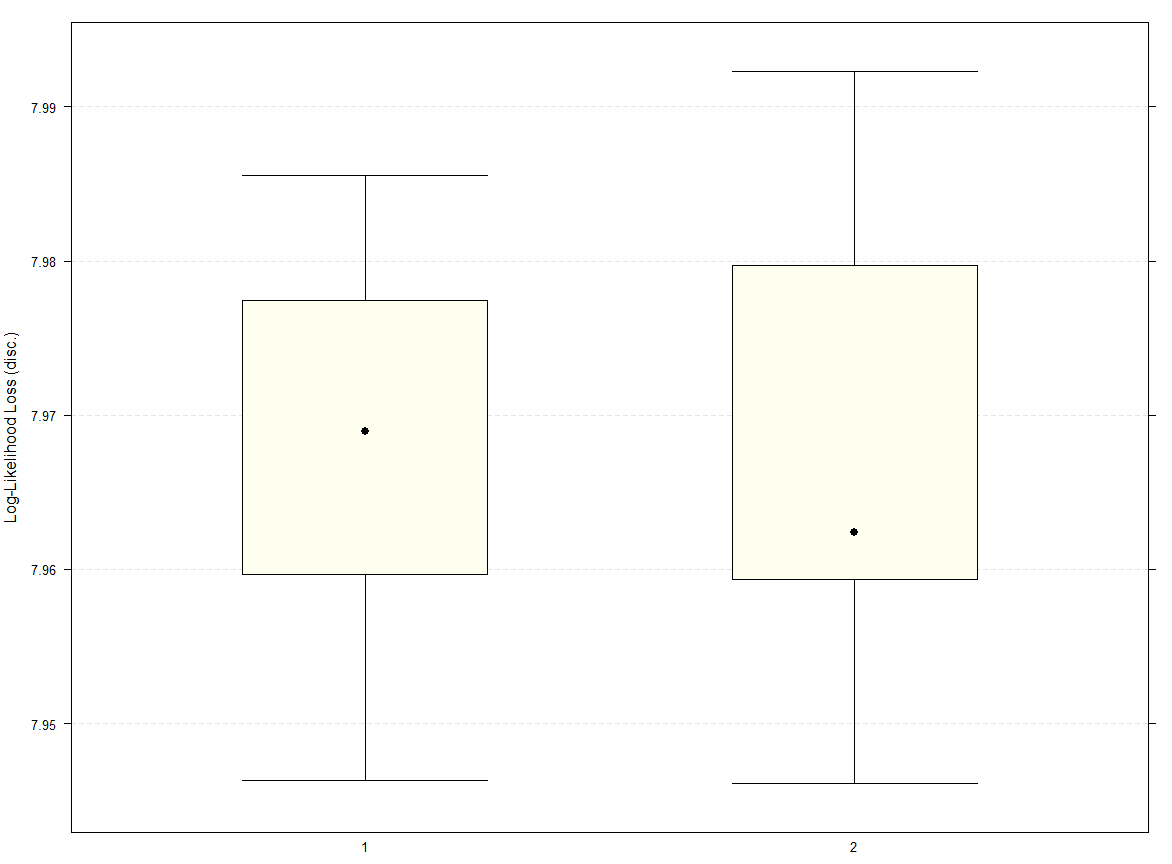
We created several constraint-based and score-based variables. The structures and associated algorithm details can be seen in Appendix A.

## Algorithm Selection

To decide which algorithm to ultimately proceed with, we made use of k-fold validation with 10 folds. Using this method, the original sample is randomly divided into 10 equally sized smaller samples, with one being retained as the validation data for testing the model and the other 9 used to train the model. This is then repeated 10 times to ensure that each sub-sample is used as the validation data at least once, with the results then averaged to produce a single estimation for the model.



As can be seen from the graph, the hill climbing, and tabu search algorithms performed significantly better than growth shrink or incremental association, suggesting that the score-based methods produced better models for this dataset. We replotted the chart with just hill-climbing and tabu search for a better comparison:



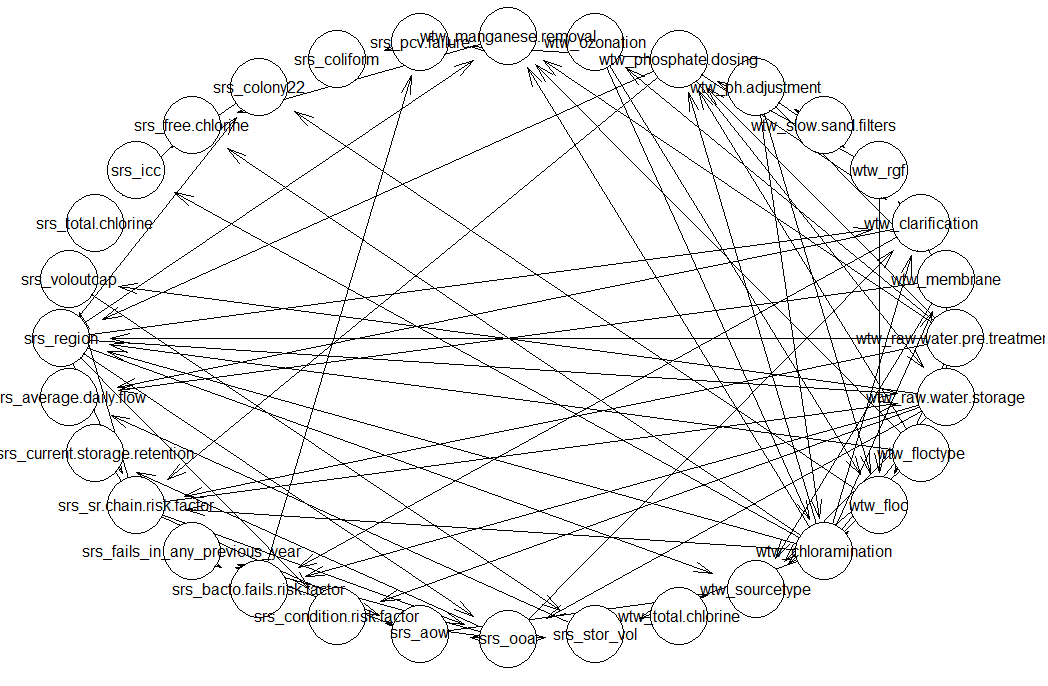
While both methods exhibit extremely similar performances, the hill climbing algorithm produced a model with a lower maximum loss than the tabu search. We decided to proceed using this model as our standard for further tests and analysis.

The hill-climbing model was used to estimate the probability of it correctly predicting whether there would be any failures in the PCV, which resulted in the following confusion matrix:

|  |  |  |
| --- | --- | --- |
|  | **observed** | |
| **predicted** | **0** | **1** |
| **0** | 2611 | 15 |
| **1** | 0 | 36 |

The model was able to correctly predict every instance in which a failure occurred, with no false negatives predicted and only 15 out of 2611 true negatives predicted as a failure, an impressive F1 score of 0.827 calculated using the below equation:

The final model is shown below:



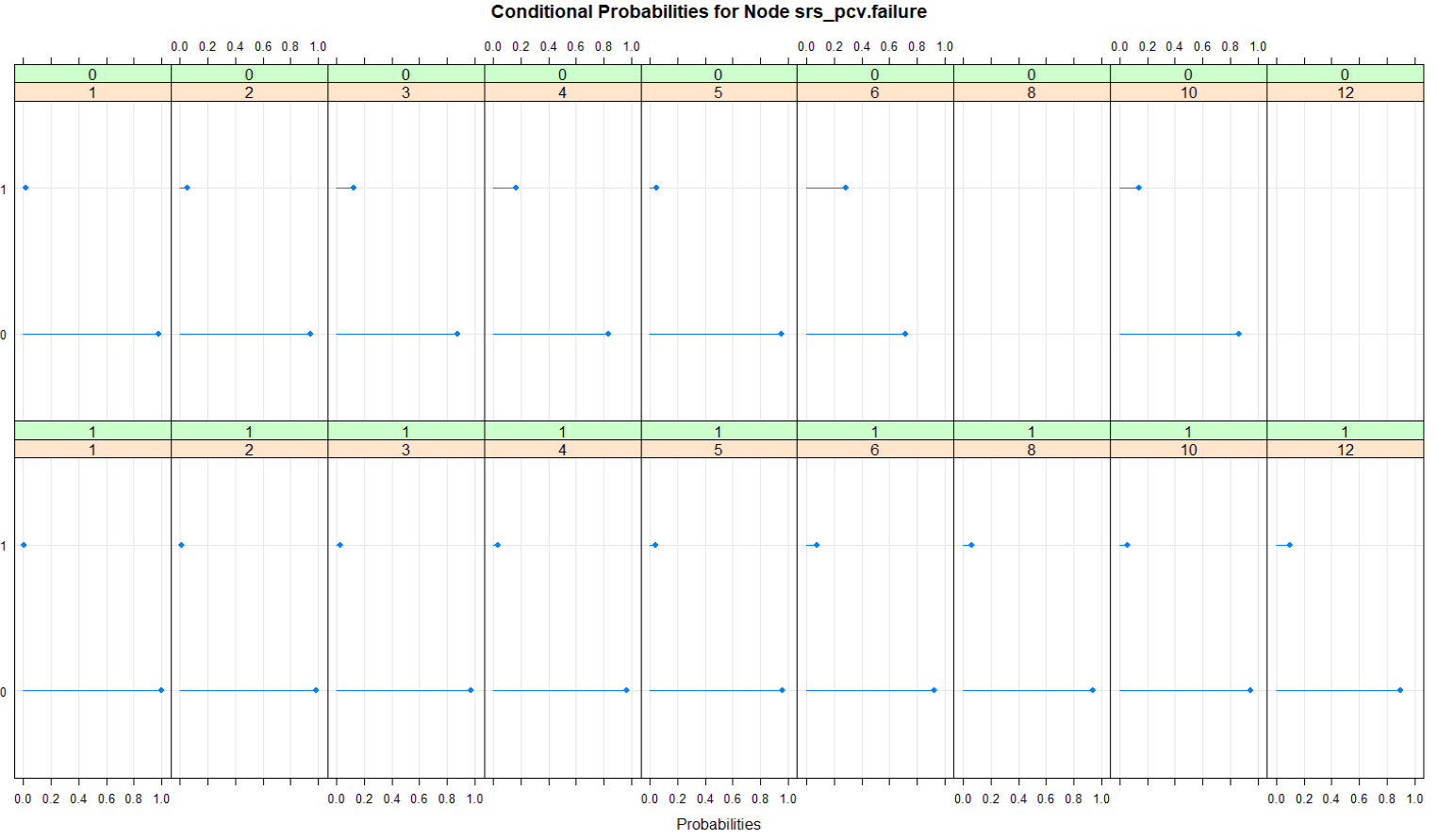
## Parameter Learning

Parameter learning allows use to simplify the model down to the effects each parameter has on the other parameters. Think of it like building several smaller models each with only one node and its parents. There are two estimators commonly used for this, Maximum Likelihood estimators, which measure the maximum entropy present, and Bayesian posterior estimators which minimises the loss present based off the structure of the overall network and any information already present in the data itself.

However Maximum Likelihood estimators are unstable for multivariate models (ours has almost 30) and so based off the standards used by computer scientists, we will use Bayesian posterior estimates to produce more sensible parameter estimations.

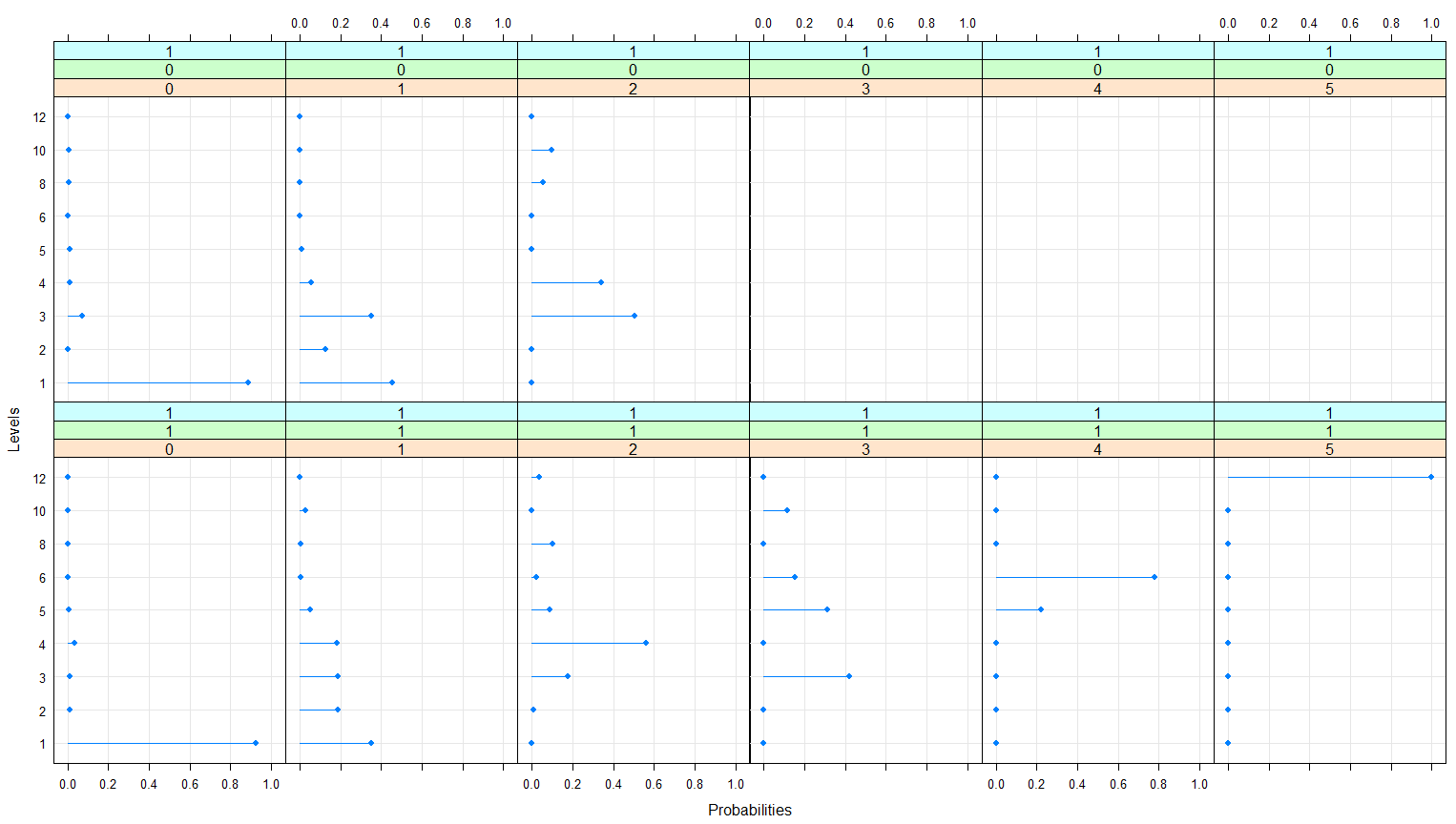
We carried out parameter learning for the following parameters:

Source Reservoir PCV Failure



The two factors affecting pcv\_failure are wtw\_phosphate.dosing and the srs\_bacto.fails.risk.factor. As can be seen in the graph, the higher the risk factor gets the more likely it is that the pcv will fail, rising from a 0.05% chance of failure up to nearly 15% chance at a risk factor of 10. This is mitigated somewhat when phosphate dosing is implemented, dropping the failure rate at a risk factor of 12 down to 10%.

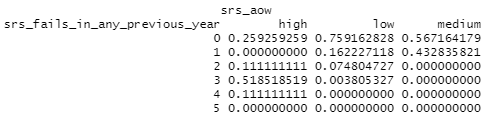
Bacto Fail Risk Factor



The Bacto Failure Risk Factor is impacted by whether clarification is carried out during water treatment, whether or not raw water is stored during water treatment and the number of fails the previous year.

Predictably the more failures the previous year the higher the risk factor is, however, less apparent is the effect the other two factors have on the risk. If raw water is stored then the chance of a higher risk factor being assigned is increased, contrarily if a sedimentation or clarification stage is present then the risk is reduced somewhat.

Source Reservoir Failures in Previous Years



The main factor affecting the number of failures in a year seemed to be most dependant on the age of the water, with water stored for longer than the standard test period of 21 days having a significantly higher probability of causing more than 1 failure in the source reservoir’s quality.

# Results of Analysis

## Inference

After learning both the structure and gaining some insight into the parameters of our Bayesian network, the next step would be to carry out inference, as this will be the main purpose of the entire modelling process. Specifically, in our model we want to answer the question, for given evidence (this is the value of a known variable) about different combinations of the variables, what is the likely state of pcv\_failure?

We have a choice of inference methods, exact or approximate. Exact algorithms are more iterative in nature and provide more reliable answers but take a significantly longer time to run, whereas approximate algorithms don’t always provide the best answer, but they do scale better for a large number of variables. As our model contains many variables, we will resort to using approximate methods.

Approximate algorithms are based off Monte Carlo simulations which sample from the global distribution of the entire network. The three most popular sampling methods are importance sampling, logic sampling and Markov Chain Monte Carlo (MCMC). We will experiment with and compare these three methods to determine which is the most applicable to our network.

We began with the two variables directly connected to the pcv\_failure node- wtw\_phosphate.dosing and srs\_bacto.fails.risk.factor, to see what effect they had on the probability of failure.

|  |  |
| --- | --- |
| Probabilities with Phosphate dosing | Probabilities without Phosphate dosing |

When phosphate dosing is present the chance of failure is 0.012, whereas when it is not present it increases up to 0.051, a fivefold increase.

Comparing the different bacto fails risk factors yielded the following probabilities:

|  |  |  |  |
| --- | --- | --- | --- |
| Prob at Risk level 1 | Prob at Risk level 2 | Prob at Risk level 3 | Prob at Risk level 4 |
| Prob at Risk level 5 | Prob at Risk level 6 | Prob at Risk level 8 | Prob at Risk level 10 | Prob at Risk level 12 |

As expected, the highest chance of failure comes with a risk level of 12. Counter-intuitively though the second highest chance of failure comes at risk level 6, with both risk level 8 and 10 having lower failure rates. This suggests that the factors affecting the bacto fail risk should be examined in more detail, particularly what effect they have on the pcv failure chance.

We then tested out different values for tw\_raw.water.storage, wtw\_clarification and srs\_fails\_in\_ any\_previous\_year to see what effect they had on the chance of failure:

|  |  |
| --- | --- |
| Probabilities with raw water storage | Probabilities without raw water storage |
| Probabilities with water clarification | Probabilities without water clarification |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Probabilities with 0 failures in past year | Probabilities with 1 failure in past year | Probabilities with 2 failures in past year | Probabilities with 3 failures in past year | Probabilities with 4 failures in past year | Probabilities with 5 failures in past year |

When raw water is stored it nearly doubles the chance of a failure occurring (although this is only from 1 to 2 % so still a small effect overall). Water clarification has a minimal effect on the chance of failure, as when it does not take place the chance of failure increases by only 0.2%, while the number of failures in the past year has the biggest effect, varying the chance of failure by up to 10% at 5 failures in the past year- because of this we will test the factors that affect the number of failures in past years.

The two factors affecting this are the order of appearance of the water in the service reservoir system (there are 8 smaller reservoirs in the system) as well as the age of the water, which is divided into up to 7 days, between 7 and 21 days and more than 21 days. The results were the following:

|  |  |  |  |
| --- | --- | --- | --- |
| Prob in reservoir 1 | Prob in reservoir 2 | Prob in reservoir 3 | Prob in reservoir 4 |
| Prob in reservoir 5 | Prob in reservoir 6 | Prob in reservoir 7 | Prob in reservoir 8 |

It seems that in general the risk of failure decreases as the water moves through the reservoirs, with the last reservoir effectively removing the chance of failure (only a 0.07% chance). This suggests that the age of the water is were most of the risk of failure is coming in this case:

|  |  |  |
| --- | --- | --- |
| Prob with low water age | Prob with medium water age | Prob with high water age |

Again, oddly, water with an age of 7 to 21 days has a higher chance of leading to failure than water less than 7 days old and most surprisingly, water older than 21 days old which has the lowest effect on failure of the three ages. We will therefore investigate the factor linked to the age of the water; srs\_sr.chain.risk.factor. This is however composed of several other factors which we will more closely investigate: whether the water has been chloraminated, whether the water was pre-treated using screens or which region (north, south, east, and west) the water originated from.

Adjusting these variables had minimal effect on the failure rate of the pcv, except for one- when comparing the 4 regions, 3 of the 4 had similar rates of failure (~0.13) except for the north region which had a rate of 0.18, a noticeable increase, suggesting more attention should be paid to water originating from this region.

The only variable connected to the region that we have not already examined is wtw\_floctype, which suggests that pcv failure could possibly be related to the type of floc used, or more likely the lack of a floc at all.

# Conclusion

To conclude, we have determined several key factors that affect the likelihood of a failure occurring in the Prescribed Concentration or Value. These are:

* Whether or not Phosphate dosing for plumbosolvency control is present
* If the raw water is stored prior to treatment
* If a clarification or sedimentation stage is present
* The number of failures that have occurred in the previous year
* How long the water is stored for
* Whether a chloramination process was used
* If raw water pre-treatment such as screens or microstrainers are present
* Whether or not the flocculant is dosed, and if so with iron or aluminium

The model also suggests that further investigation should be carried out on water from the North region of Scottish Water’s supply as the chance of failure in this region is higher than the other three. It could also be worth looking deeper into differences between each service reservoir in an SR cascade as which reservoir the water originates from.

The model has some limitations that must be mentioned:

* Currently, the model cannot handle continuous data, and therefore any future data that is fed into it will need to be discretised in the same manner as the original data before it can be input, which could possibly affect the structure of the learned network and any predictions that arise from it.
* The data fed into the model contained many variables, so it was difficult to select which ones to include and which to exclude. It is possible that important values have been dropped and therefore the model may not fully account for all the factors affecting the pcv failure.
* As the model is based off the Scottish Water service and we are not experts on the subject, it is possible that important connections are missed, or meaningless ones are added. To alleviate this expert knowledge could be incorporated to create an expert network with these connections restored/removed.

By addressing these limitations, we can improve the model and the usefulness of any results it produces.

# References

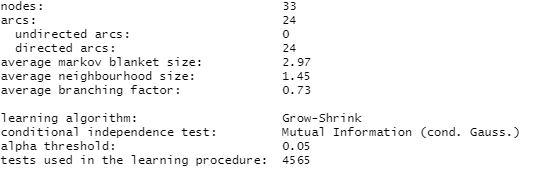
* Scottish Water (2021) *Water Treatment,*Available at: *https://www.scottishwater.co.uk/help-and-resources/education/all-about-water/water-treatment*
* New York State Department of Health (July 2017) *Coliform Bacteria In Drinking Water Supplies*, New York: New York State Department of Health.
* Mercian Science (2021) *TVC (Total Viable Count),*Available at: [*https://www.mercianscience.co.uk/our-applications/tvc-total-viable-count/*](https://www.mercianscience.co.uk/our-applications/tvc-total-viable-count/)
* Royal Society for Public Health (1 March 2017) *Principles of Water Supply Hygiene*, United Kingdom: Water UK.
* Scottish Water (2021) *Chloromination FAQs,*Available at: *https://www.scottishwater.co.uk/help-and-resources/faqs/water-faqs/chloramination-faqs*
* World Health Organization (1996) 'Health criteria and other supporting information', *Guidelines for drinking-water quality,*2(2)
* O. AbdulrahmanAdeleke et al. (2019) 'Principles and Mechanism of Adsorption for the Effective Treatment of Palm Oil Mill Effluent for Water Reuse', *Nanotechnology in Water and Wastewater Treatment,* pp. 1-33.

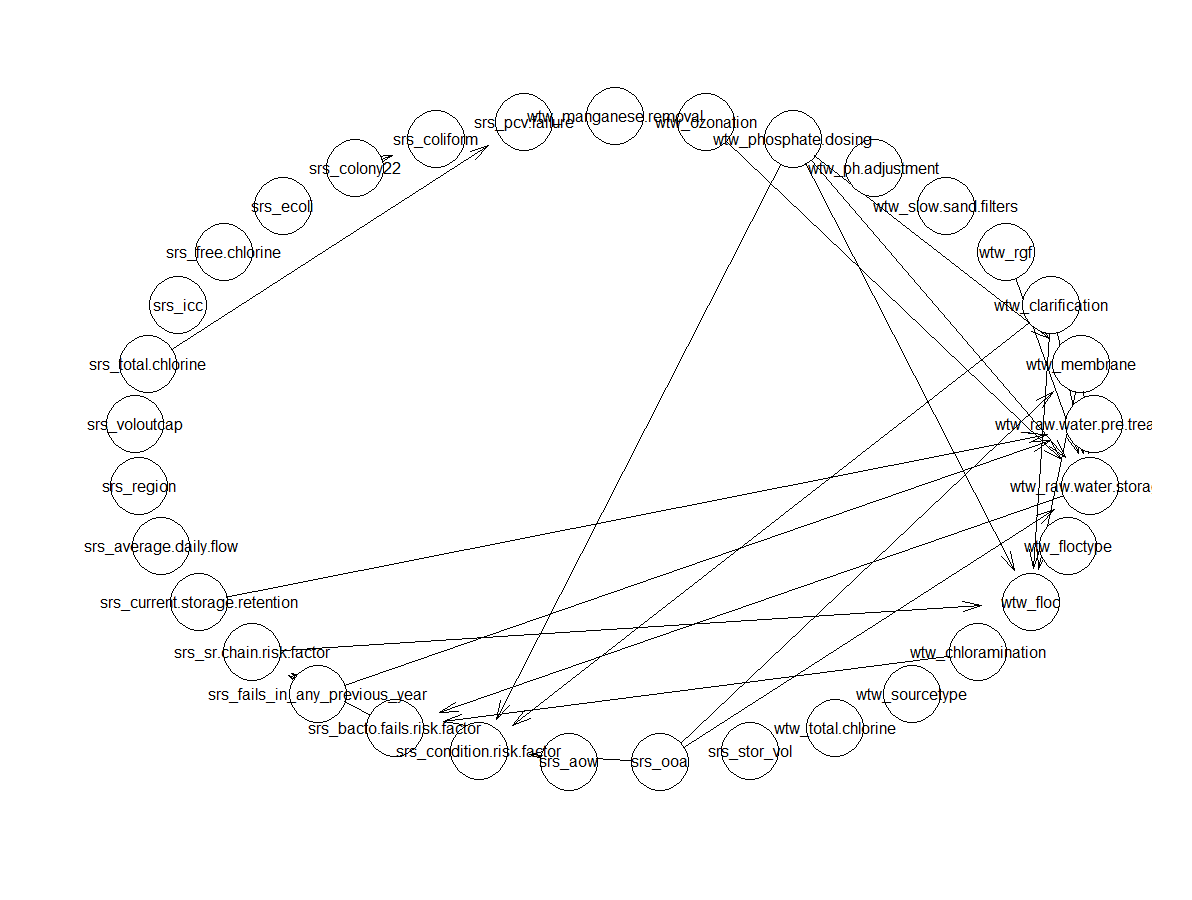
# Appendices

## Appendix A: Structure Learning Models

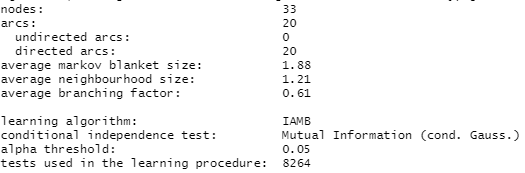
### Constraint Based Models

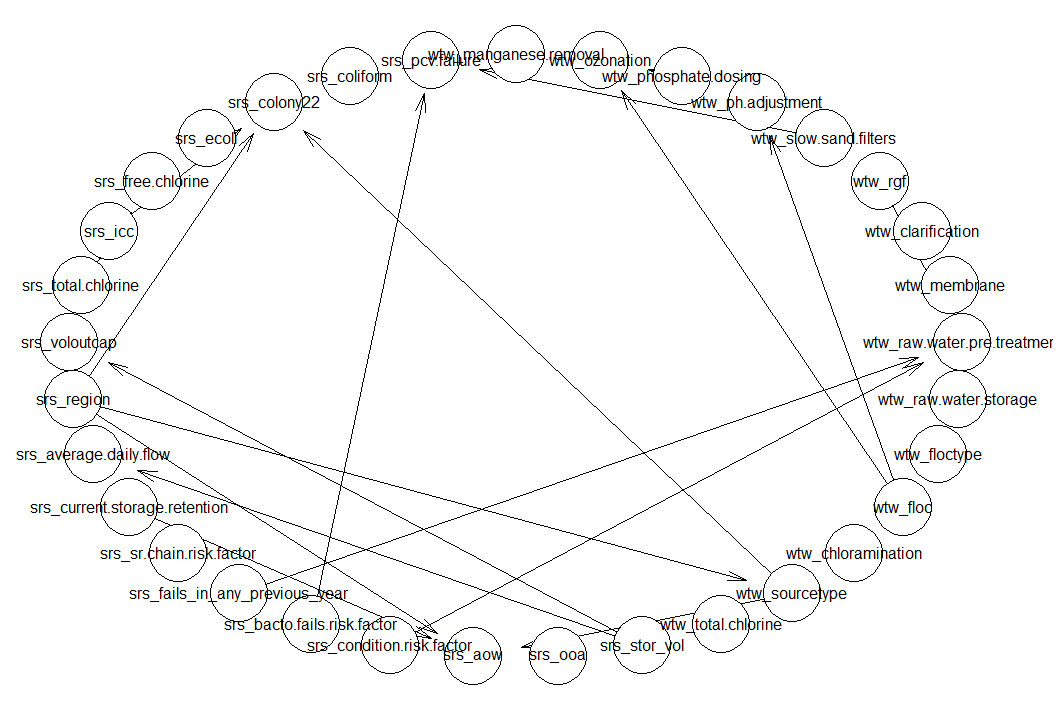
Grow-Shrink Algorithm



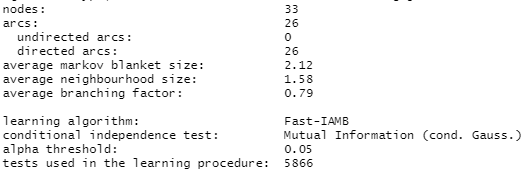


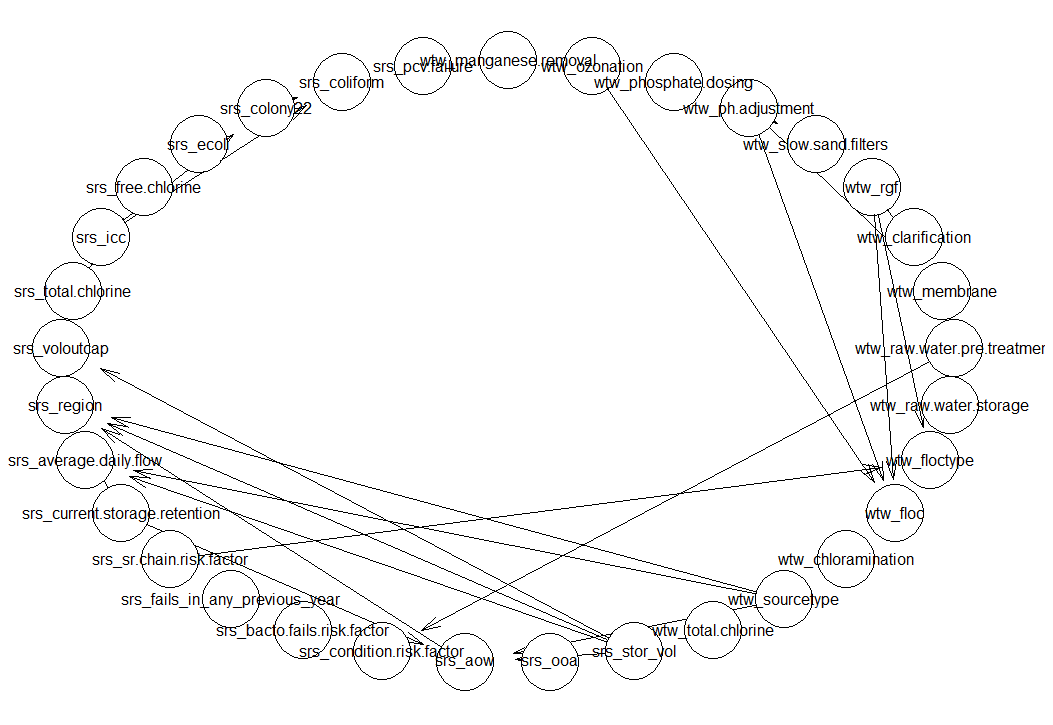
Incremental Association



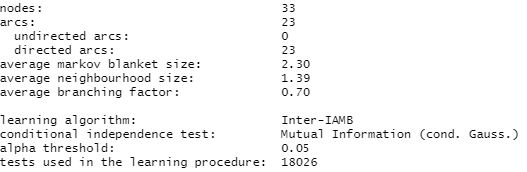


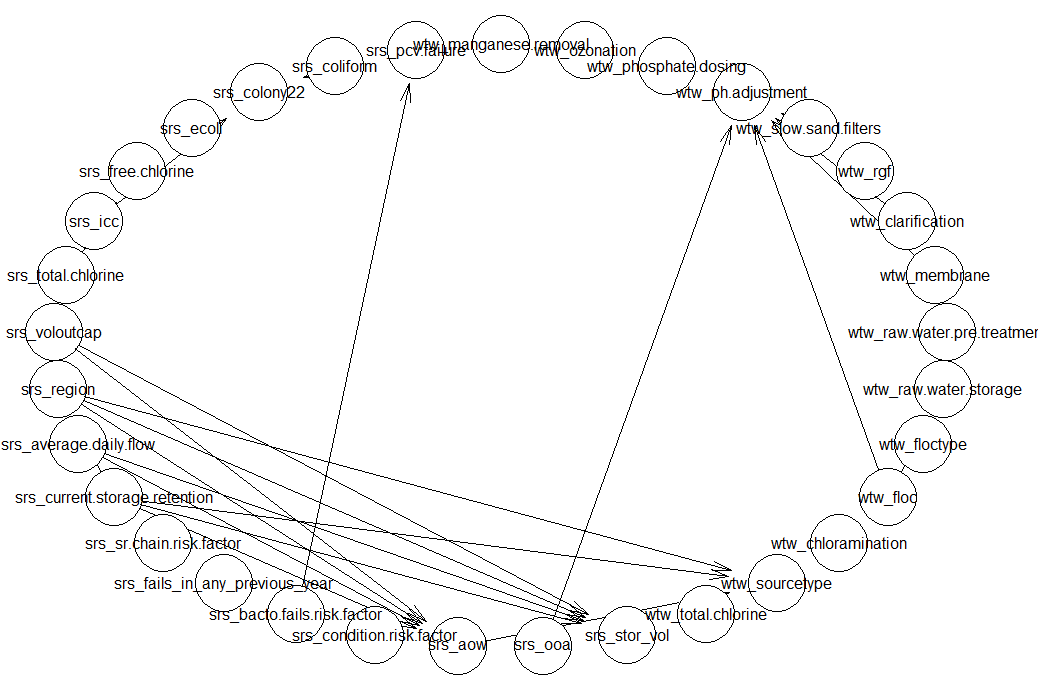
Fast Incremental Association



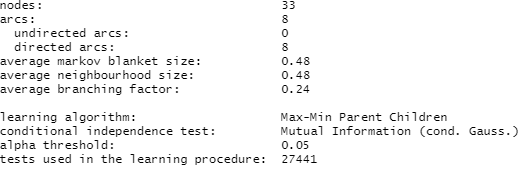


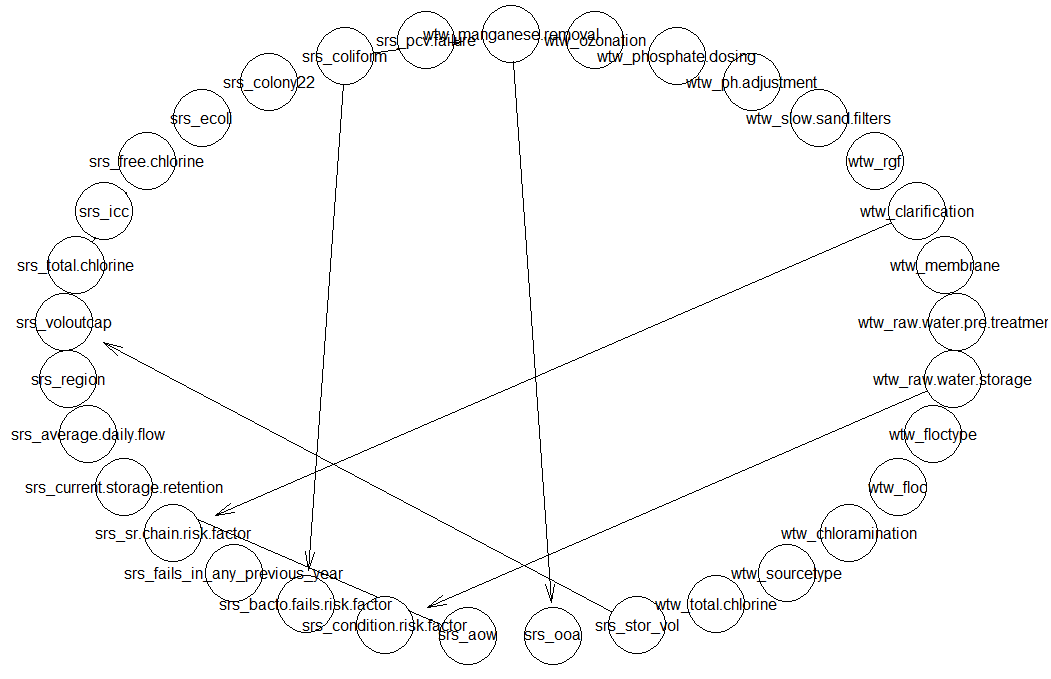
Interleaved Incremental Association





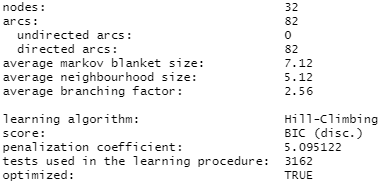
Max-Min Parents and Children

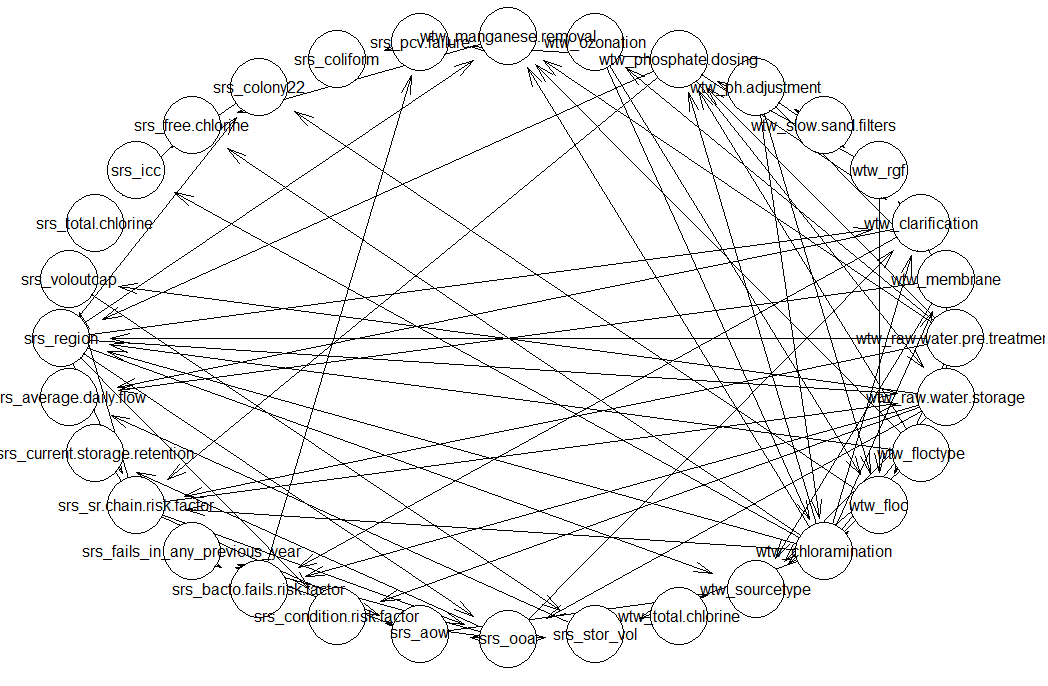




### Score-Based Models

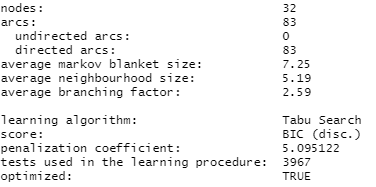
Hill-Climbing Algorithm

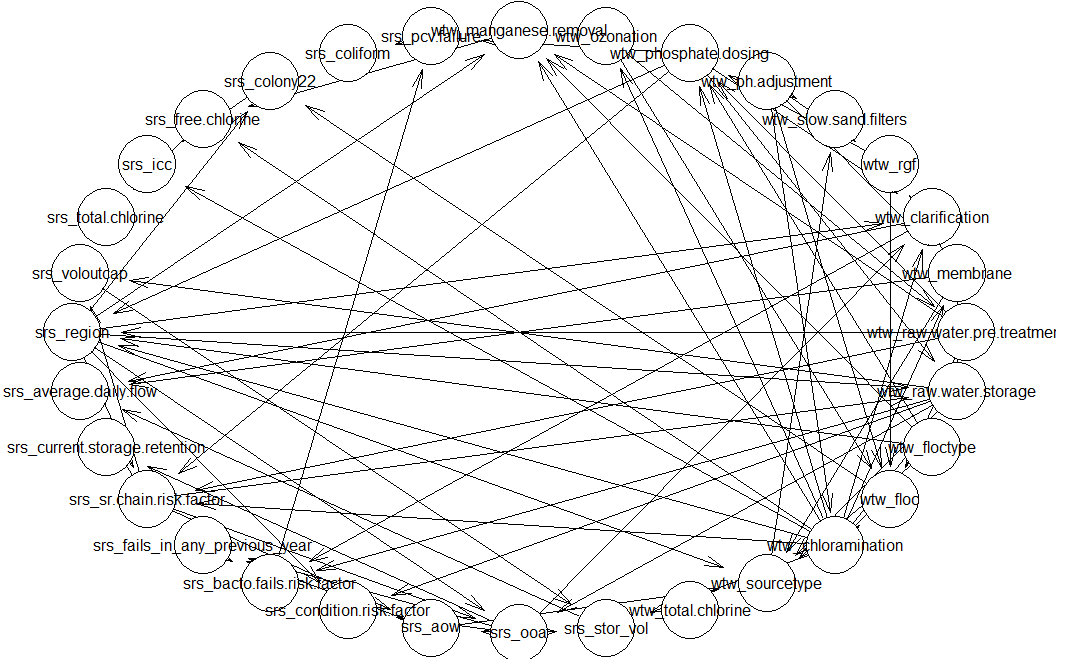




The Hill-Climbing algorithm produced a model with a BIC score of -224647.7

Tabu Search Algorithm





The Tabu Search algorithm produced a model with a score of -224626.8, which is slightly better than the hill climbing algorithm