A Bayesian network model for predicting the sewer blockages at a large utility in the southeast

Barbara Podgorska
Mentored by Patrick Heritier Robbins

Introduction

Approximately 11.4 to 37.9 million cubic meters of untreated wastewater are released annually in the United States due to sanitary sewer overflows (Itaquy et al., 2017). Discharge from sanitary sewers contains pathogens and hazardous substances which pose threats to the environment and public health, making it crucial for utilities to respond to and clean up overflows quickly. Cleaning overflows is expensive; therefore, it is essential that strategies are implemented to prevent them.

The Environmental Protection Agency determined that blockages are the leading cause of sanitary sewer overflows, accounting for 43% (EPA, 2022). The purpose of this project was to create a Bayesian network model which assessed the risk of a blockage in a sewer pipe, allowing utilities to mitigate the threat of sanitary sewer overflows due to blockages.

To predict the probability of blockages, this project used the tree augmented naïve (TAN) Bayesian network algorithm to create a machine learning model. Like a traditional Bayesian network, TAN Bayesian networks use the rules of conditional probability, but with the addition of a tree structure where each feature of the model is dependent on the classification variable and only one other feature. TAN Bayesian networks have been proven be more accurate due to their unique structure (Friedman et al., 1997).

Materials and Methods

Geographical information systems (GIS) and pipe cleaning data were obtained for a utility in the southeast. GIS data was extracted using the QGIS software. The data was cleaned in Excel and joined through Python using the Pandas library. Features were chosen based on known impacts on blockages and their availability in the data.

After cleaning was completed, 27,310 pipes were retained from the data with the following features: operational area, subbasin, length, diameter, slope, material, and years since the asset was last cleaned. Each pipe was classified into one of the following risk categories: none, mild, moderate, or severe. The pipe's category was determined by adding together the severity of each previous blockage.

Following the categorization of pipes, the data was split into 80% training and 20% testing data. The training data was imported into GeNie, a software created for Bayesian networking. Next, the software determined the model's parameters and began to train the model. The model was initially trained with a sample size of 50. Model adjustments ultimately led to a sample size of 100. Background knowledge was added to the model which allowed users to set a predetermined model structure. Background knowledge allows the model to prioritize and thus increase the weight of variables. In this model, the time which a pipe was last cleaned was in the first tier, whereas the remaining variables were in the second tier of background knowledge. This allowed for the last cleaning of a pipe to be prioritized by the model.

At first, multiple models were created using a Bayesian search algorithm, however, the models had accuracies lower than 50%. The algorithm was switched to the TAN Bayesian network to improve accuracy. Once models were trained, they were validated and improved upon until maximum accuracy was achieved.

Results

After training multiple Bayesian networks, the final model achieved an accuracy of 79.21% in its validation phase. The final structure of the network is seen in Figure 1. Through learning, the weights of influence between each variable were uncovered. The model determined that the pipe's subbasin most influenced the predicted rating followed by the year the pipe was last cleaned with weights of 0.263 and 0.203, respectively. Material was found to be the least influential with a weight of 0.068.

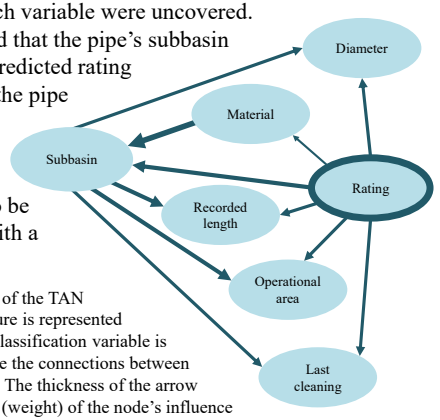
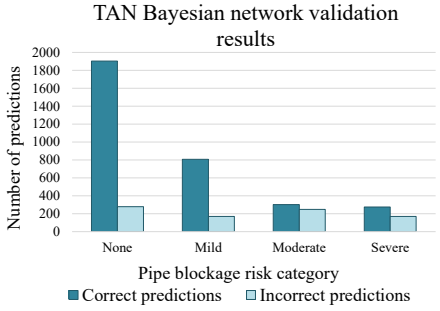


Figure 1 (right): A diagram of the TAN Bayesian model. Each feature is represented by a light blue circle. The classification variable is outlined in dark green as are the connections between the nodes shown as arrows. The thickness of the arrow corresponds to the strength (weight) of the node's influence in the model. Arrows pointing away from the classification node are a result of backpropagation.

Graph 1 (right): A bar graph displaying the number of correct and incorrect predictions the model made for each risk category. The model was 82.77% accurate for no risk of blockages, 54.82% for mild risk, 87.24% for moderate risk, and 61.94% for severe risk.



Results (continued)

		Predicted Risk			
	Classification	None	Mild	Moderate	Severe
Actual Risk	None	1904	254	12	12
	Mild	120	807	21	27
	Moderate	73	146	301	29
	Severe	37	113	19	275

Table 1 (above): A confusion matrix of the model's validation results displaying the actual risk of a pipe's blockage versus the models predicted risk of a pipe's blockage for each category of risk.

Conclusions

The purpose of this project was met, a TAN Bayesian network was successfully built and achieved an accuracy of 79.21%. This network could be used as a basis for utilities looking to prevent blockages through the implementation of machine learning.

Although the overall model was able to achieve a high level of accuracy, the moderate and severe risk classifications underperformed with approximately 55% accuracy and 62% accuracy, respectively. This discrepancy could be a result of the disproportionally small number of moderate and severe cases in the training data. As evidence of this, the model had a greater accuracy for the categories with more available training data, no risk at approximately 87% and mild risk at 83%. Likely for those same reasons, there were 150 instances of the model predicting a pipe to either have no risk or mild risk when the actual risk of a pipe blockage was severe (Table 1). Future studies could focus on creating models with greater data sets from multiple utilities to create a universal model.

Nevertheless, the structure of this model could still be implemented by utilities to successfully detect pipes that have a risk of blockage. By detecting these pipes utilities would be able to clean them before they result in overflow, resulting in costly repairs and threatening the environment and public health.

References

Environmental Protection Agency. (2022). *Sanitary Sewer Overflows (SSOs).* https://www.epa.gov/npdes/sanitary-sewer-overflows-ssos

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*, 131–163. https://doi.org/10.1023/A:1007465528199

Itaquy, B., Olufunso, O., & Giacomoni, M. H. (2017). Application of Multiobjective genetic algorithm to reduce wet weather sanitary sewer overflows and surcharge. *Journal of Sustainable Water in the Built Environment*, *3*(3). https://doi.org/10.1061/JSWBAY.0000826