water e-journal
Online Journal of the Australian Water Association
ISSN 2206-1991
Volume 2 No 4 2017
https://doi.org/10.21139/wej.2017.035

PREDICTING SEWER CHOKES THROUGH MACHINE LEARNING

ADVANCED ANALYTICS TO PREDICT THE NUMBER, LOCATION AND TYPE OF CHOKES IN SEWER ASSETS

B Cameron, M McGowan, C Mitchell, J Winder, R Kerr, M Zhang

ABSTRACT

In 2015, sewer chokes cost Sydney Water and its customers over $10 million. Sewer chokes are blockages typically caused by external factors such as tree roots, fats and grease, and foreign items in the pipes such as wet wipes. Chokes may lead to sewage overflows through designated sewer system overflow points, or uncontrolled overflows onto public or private property, or rarely inside houses.

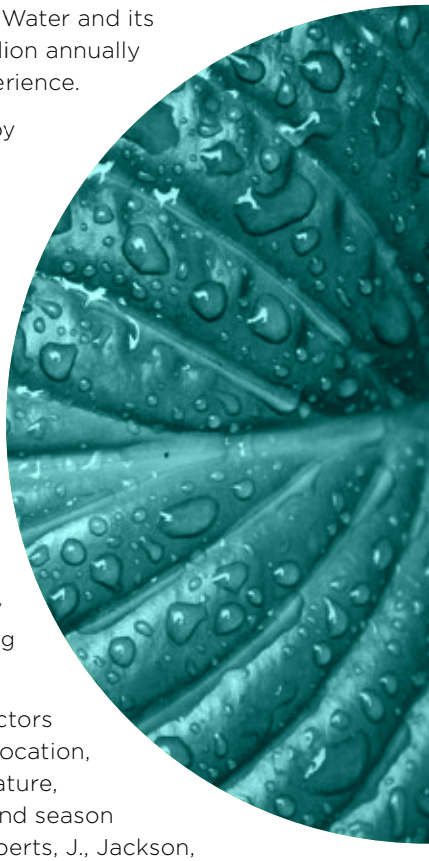
The likelihood of any particular main choking may be influenced by many environmental, social and structural factors. Traditional forecasting and prediction techniques are currently not capable of accurately predicting sewer chokes in the Sydney Water network. This limits the extent of the network where preventative maintenance strategies can be economically employed.

Sydney Water, through working with CSIRO's Australian Digital and Data Innovations Group, Data61, has used machine learning techniques to analyse factors that may contribute to sewer chokes and has developed a pilot model to predict the likelihood of future chokes in every sewer main asset. This may enable Sydney Water to shift from the reactive approach towards a more proactive approach to sewer choke management.

Using data analytics can assist Sydney Water to improve our ability to predict future events, such as sewer chokes. This will support Sydney Water's corporate strategy by enabling the business to deliver world class service to our customers, the community and the city.

INTRODUCTION

Sydney Water manages about 25,000km of sewer pipeline that transports wastewater from customers' homes to 28 wastewater and recycled water treatment plants. When chokes occur in these pipes, sewage can escape through designated emergency relief structures, other structures and through surface fittings. Sewage can overflow directly to waterways on streets, public land or private property or even inside homes. Reacting to sewer chokes costs Sydney Water and its customers in excess of $10 million annually and impacts on customer experience.



Sewer chokes can be caused by many different environmental, behavioural and structural factors - tree roots, fats, oils and grease (FOG), pipe and junction deformities, and foreign items in the pipes such as wet wipes.

Observations of extracted sewer chokes suggests that blockages are often caused by a combination of factors.

Tree roots are of particular concern as they are recorded as the cause of 80% of Sydney Water's sewer blockages during dry climatic conditions.

Environmental and physical factors such as tree species, age and location, soil type and porosity, temperature, availability of water, weather and season contribute to root growth (Roberts, J., Jackson, N., and Smith, M., 2006).

Wet wipes are an ongoing problem, with 75% of

Machine learning for asset maintenance

Sydney Water's chokes involving flushed wet wipes. The ongoing 'Wipes in Pipes' media program and the accumulation of FOG is driven by customer behaviour.

Pipeline characteristics have also been recognised as a significant factor to sewer chokes - pipe type, size, age, depth, location, defects and junctions all influence performance.

The complex interaction between these environmental, physical and customer factors makes predicting the location and likelihood of future chokes a significant challenge.

Currently, Sydney Water's maintenance strategy has three main elements:

1. Reactive response to chokes.
2. Corrective inspection of sewers that have choked 3 times in 5 years.
3. Preventive inspection of sewers likely to block and overflow to waterways (Sinclair Knight Merz, 2005).

The preventive part of the program is informed through a number of factors including tree canopy location, depth, age, diameter and flow rate.

In recent years, the number of sewer chokes has also been forecast using the correlation of choke occurrence and the Southern Oscillation Index (McGowan, M., Mitchell, C., 2015). However, since 2015, this correlation has been unable to predict accurately the overall number of chokes in Sydney Water's area of operations (Figure 1).

A research project was initiated in 2016 to improve the methodology used to predict sewer chokes and consider how predictions could inform a preventative maintenance program.

If successful, the outcome would be more efficient allocation of funds and resources for sewer choke management and reduced customer and environmental impacts from sewer chokes. It would also inform business strategies used to manage dry weather overflows.

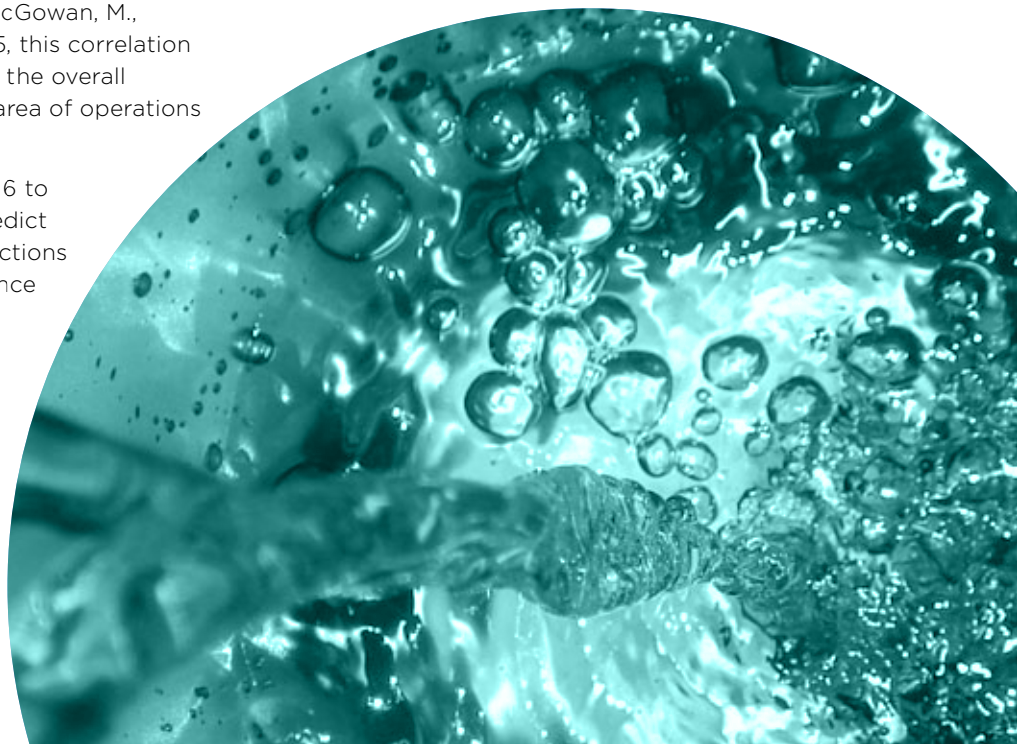
To inform a preventative maintenance program, high resolution predictions – location, rate of choke occurrence and likelihood at an individual pipe level, are required. This is in contrast to other models, which aim to predict overall numbers of chokes (Franks, S., 1999), or location of chokes within a limited area of pipes (Bailey, J., et al. 2016). To achieve this, detailed data of all possible factors involved in sewer chokes would be evaluated for their ability to predict chokes ("mutual information score") and incorporated into a model.

METHODOLOGY

A machine learning approach, along with some of its component statistical methodologies, were piloted to predict the location and occurrence of sewer chokes.

Sydney Water has data on approximately 300,000 blockages occurring in their wastewater system between 2001 and 2016. Each of these are geo-located, linked to a specific asset and contain details such as the main cause of the choke and the costs associated with the work order.

A machine learning approach seemed suitable to pilot given the large volume of data available and the complex interactions that lead to each individual choke - pipe characteristics (e.g. material and length), its surrounding environment (e.g. weather, topography and tree coverage) and customer behaviour (e.g. wet wipe disposal, FOG).



Demographic, building approval and customer property related data were used as a proxy in modelling the impact of customer behaviour on chokes. 7,942 data groups from the 2011 census data were matched to choke rates for SA2 (statistical area level 2) regions.

These data groups cover a wide range of information such as gender, age, birthplace, education, occupation, and many others. These factors were tested for their ability to provide informative indicators depicting common customer behaviours, some of which show high correlation with certain types of chokes, e.g. grease-caused chokes.

The pilot predictive model was developed in four stages:

1. Factor analysis – identifying the relative importance of each feature (potential factor present in dataset) to contribute to accurate predictions. This was achieved using a Mutual Information (MI) score for each feature, where the higher the MI the more important the feature is in the model. This analysis showed different models would be required for the four most prevalent choke types listed in the dataset.
2. Model development and training – identifying and 'learning' the patterns in the data which lead to chokes. These latent blockage patterns were then incorporated into four different algorithms, one for each main choke type. A 'training set' of the data for the period 2001 to 2014 was used to build a predictive model, as depicted in Figure 2. The model output listed the likelihood of each individual pipe to experience a choke. Refer to Equation 1 for the algorithm used.
3. Validation – The model was then 'back tested' to predict chokes in 2015. Pipes that were identified, by the model, to have the highest likelihood for chokes were ranked and the results were compared against actual choke records.
4. Prediction – Potential chokes, their location and likelihood, were then predicted for 2016.

Algorithm Adopted

For each group, the blockage likelihood of a group member was then depicted by an intensity function:

$$\lambda=\mu(t)+\gamma\sum_{tk<t}g(t-t_k)+f(\theta_1,\theta_2,\ldots,\theta_m)$$

The output of the function is the predicted blockage possibility for the considered pipe. The function makes the prediction based on:

1. the pipe's intrinsic blockage possibility over time, determined using a Weibull process
2. the influences of the pipe's previous blockages , and
3. influences of the pipe's physical and environmental attributes e.g. age, material, diameter and length of pipe, soil moisture, tree coverage percentage and demographic attributes.

Other data-driven and machine learning-based approaches related to this study can be found in:

- Guo, T., Zhang, B., Wang, Y., et al., 2016;
- Lin, P., Zhang, B., Guo. T, et al., 2016;
- Lin, P., Zhang, B., Wang, Y., et al., 2015; and
- Li, Z., Zhang, B., Wang, Y., et al., 2014.

RESULTS

Factor Analysis

Factor analysis considering choke records and other relevant data identified the relative importance of each possible contributing factor of the four most prevalent choke types.

For example, the most influential factors related to chokes caused by tree roots in the dataset were tree canopy coverage, length, depth, soil moisture, soil type, material and laid year, in addition tothe relationship between each of these factors and their level of exclusivity to each other.

Table 1. Factors considered in factor analysis

Pipe data		
Material	Length	Age
Size	Joint type	Laid Year
Asset number	Depth	Location
Flow depth		
Choke data		
WO number	Choke date	Location
Asset number	Cause	
Lot coverage	Trade waste catchments	
Climate & soil data		
Rainfall	Temperature	Evaporation
Soil moisture	Soil type	
Tree data		
Tree coverage		
Other data		
Census data	LGA coverage	Zoning
Sewer SCAMP		Lot coverage
Trade waste catchments		Building cover
Catchment coefficient		

Figure 3 shows there is a relationship between pipe depth and tree coverage factors. This demonstrates that not all factors are exclusive of each other i.e. depth of pipe and tree canopy coverage. Therefore, they cannot be considered in isolation when developing a sewer choke prediction model.

Models for other choke types - debris, soft and grease chokes, also considered demographic and related property characteristic data .eg. Census data, building approval and property type. For example, lot coverage type (property characteristic data) has been considered as an important factor for chokes caused by debris. There are 31 lot types from lot coverage data (such as school, commercial, industrial, etc.) and it was found that super lot and mixed development have the highest correlation with debris choke, as shown in Figure 4. In terms of grease and soft choke-caused chokes, human factors, such as living and eating habits, take a great contribution. We obtained the census data from ABS which contains 7,942 census factors for suburbs, e.g. education, occupation, race, etc. Based on our analysis, education background (Persons did not go to school age 45-54 years) and occupation (Manufacturing occupation, machinery operators and drivers) are the most related factors to grease-caused chokes compared to other census factors, as shown in Figure 5. While for soft chokes, family background (Persons lone parent age 15-24 years) and occupation (Males manufacturing age 15-19 years) are most related as shown in Figure 5. Only the most correlated factors from ABS census data were selected and lot coverage data for different choke types, as shown in Figure 6. Figure 5 shows that the range of mutual information scores is lower than the one in Figure 7. The total number of chokes caused by debris, grease and soft chokes is much less than the number of root chokes, and therefore causes the choke distribution for those choke types to have a more random distribution. Even so, we can still discover the relative correlation between factors and chokes from these figures.

Figures 4 to 7 show that tree canopy coverage still has a strong correlation with other choke types besides root chokes. It is because one choke can be caused by a combination of different factors, and only one choke type had been recorded.

Table 1 lists all the factors that were considered in the factor analysis.

Analysis of demographic and other property information showed correlations between both building approvals and property type to debris chokes.

Different property types also have significant relationships with debris and grease chokes. Debris, soft and FOG chokes have similar choke patterns among different catchment areas.

Table 2. Factors considered in each predictive model

Choke Type Model			
Tree root	**Soft**	**Debris**	**Grease**
Depth	Depth	Depth	Depth
Length	Length	Length	Length
Tree Canopy Coverage	Tree Canopy Coverage	Tree Canopy Coverage	Tree Canopy Coverage
Laid Year	Laid Year	Laid Year	Laid Year
Moisture	Family Background	Soil Type	Education Background
Soil Type	Occupation	Super Lot	Occupation
Material	Soil Type	Material	Soil Type
Joint Type	Joint Type	Mixed Development	Material
Size	Material	Joint Type	Joint Type
	Size	Size	Size

MODEL RESULTS

When considering pipes with the highest probability of chokes (top 10%) the model correctly identified 43% of pipes with chokes in 2015. This was more than four times the number identified by a random selection of pipes (Figure 8).

Figure 9 shows the difference in success rate between the different types of choke prediction models, and the model overall.

As shown in Figures 10 to 12, the curve of the number of tree root chokes with 6-month lag translation fits well with climate and soil factors. As a result, three climate factors (TempMax, Evaporation and Negative WRel2End (soil moisture at depth 0.2-1.5m) with 6-month lag translation) were used for training a model to predict monthly tree root caused chokes. Data from the beginning of July 2007 to the end of December 2013 was used and tested from the beginning of January 2014 to the beginning of December 2015 as shown in Figure 9.

(The weather data is only available from early January 2007, including TempMax, TempMin, WRel1End, WRel2End, Soil Evaporation, etc.). This is forecasting gross numbers of chokes rather than forecasting for specific sites or regions.

DISCUSSION

The project showed that a machine learning model could be built to predict the occurrence of sewer chokes if a significant amount of relevant data is available. The ability of the pilot model to inform a real preventative maintenance program or overflow maintenance strategy is limited as additional factors need to be considered – actual cost, disruption to the customer, timing, and methods and reliability of inspections and cleaning. These will be considered during further validation of the model with field investigations in 2017.

Once validation is completed, this program may:

- reduce the social, reputational and environmental impact of sewer chokes,
- enable more efficient allocation of funds, staffing and other resources,
- improve compliance with health and environmental regulation,
- reduce disruptions caused by sewer overflows, and
- reduce rebates and property damage costs.

Evaluation of the preventative maintenance program should include disruption to the customer caused by inspecting sewers via assets running through private property, versus maintaining assets by exception. Responding to sewer chokes is also relatively cheap. This means that targeting higher cost, higher consequence chokes is likely to improve efficiency and performance significantly.

Though the primary purpose of the model is to predict chokes rather than identifying the causes of chokes, it was still possible to infer some interesting correlations. The model suggests that the primary factors for chokes are outside of Sydney Water's control e.g. tree location, weather, customer behaviour, and historically used pipe materials.

All major water utilities in Australia are regularly benchmarked against each other for choke performance. Consideration should be given to normalising performance data for all utilities by taking account of the different inherent factors that each utility operates within.

For example, the high percentage of tree root chokes in the Sydney Water servicing area may have implications for benchmarking and normalising performance. Given the similarly high proportion of tree canopy coverage over sewer assets, as well as the clear link between weather and tree root growth, and subsequently choke rate, these links are important when assessing the performance of the utility in this area.

Further research and discussion may also include how performance can be normalised against weather factors such as temperature, soil moisture and evaporation.

This is especially pertinent when considering that Sydney Water is regulated annually on the number of individual and repeat uncontrolled dry weather overflows, and number of overflows reaching waterways; limits which are fixed for four years.

These factors should also be considered when developing environmental strategies, such as greening Sydney. While this is a very positive initiative, care should be taken to prevent additional burden on sewer assets. This can be done by reducing dependence on certain tree species and avoiding planting in close proximity to assets, especially in areas where the consequence of a sewer choke may be higher, such as near waterways. This will also reduce the environmental cost of sewer chokes by preventing overflows.

Demographic factors such as age, location and property type have been shown to be indicative factors for chokes not flagged as tree root chokes. While preventative cleaning of sewers would be productive in preventing these types of sewer chokes, changing customer behaviours can be considered as a way to prevent future chokes.

Factors with shorter temporal fluctuations, such as the seasonal changes in tree root growth could be considered when developing maintenance strategies, as research suggests that root cutting at certain times of the year may affect future tree root growth (Roberts, Jackson and Smith, 2006).

The use of a high number of data sets also highlighted the need for accurate, reliable records. For example, some choke records were discounted as they could not be linked to an asset, and whilst census data contributed to the model, the data is from 2011. Also, despite records indicating a particular cause for a choke, there may be multiple factors contributing to the choke which are not currently captured. This additional information may improve the ability to predict the occurrence of future chokes.

While this project used recently developed, higher resolution soil moisture data with relative success, other data that may have been useful was not available e.g. tree species could be an indicator of tree root intrusion into sewer assets. While individual tree species data is not available, techniques to improve fast identification, such as hyperspectral imaging, are becoming more reliable and readily available. Other data such as soil temperature and more accurate asset data may also improve the accuracy of the model. New datasets should be considered in future versions of the model.

The participation of subject matter experts was also vital to the success of the analytical model. This helped to focus the data selection, explain and fix anomalies and data errors, and prevent causal links between unrelated factors.

CONCLUSION

The project showed that machine learning and factor analysis techniques can be applied to improve the prediction of sewer chokes. However, despite the model providing a 33% improvement in the ability to predict sewer chokes, its use to inform any future preventative maintenance programs will need to be evaluated as part of a strategy review. Benefits of using the results from the model will need to be considered alongside the current low cost reactive response approach to sewer chokes and the disruptive nature of sewer inspections on customers.

Over the coming years, the model will be validated by assessing the condition of pipes that are flagged to have a high probability of chokes in the model. This data will help refine the modelling approach over time and determine the effectiveness of the model prediction.

The project highlighted the potential need for improved or new data that is currently not readily available, such as tree species coverage, ground temperature and accurate digitised asset data.

The ability to predict the number, location and type of chokes likely to occur may be able to inform future resourcing, business and operational planning.

The effectiveness of the predictive model will help inform the review of Sydney Water's Dry Weather Overflow Management Strategy.

THE AUTHORS

Bronwyn Cameron originally joined Sydney Water in 2012 as an undergraduate Co-op Scholar. She returned to Sydney Water on the Graduate Program in 2014. During her 4 years at Sydney Water, she has worked in civil delivery, network operations and strategic analytics. She currently works in Service Planning developing asset strategies. Bronwyn holds a First Class Honours degree in Civil Engineering from UNSW and a Master in Water Resources.

Mark McGowan is a Civil Engineer with more than 25 years' experience in the water industry. Mark started his career with the NSW Public Works Department and joined Sydney Water in 1992. Mark has held roles in Operations, Maintenance and Asset Management mainly relating to Wastewater Networks. Mark has a keen interest in leveraging data and information to better manage water and wastewater infrastructure

Craig Mitchell is a Civil Engineer (Hons) with over 27 years' experience working for Sydney Water. Craig has held various roles in operations, planning and asset management positions since 1990.

Craig is currently working in the Asset Reliability team as a Service Planning Lead for civil assets. He enjoys resolving issues in an evidence based, logical, systematic approach. Provides in-depth and expert analysis for asset risk, life cycle cost and customer/regulatory requirements.

Judith Winder has over 20 years water industry experience and has been employed by Sydney Water since 1998 in various Divisions. She has worked primarily in strategy to minimise environmental and human health impacts from wastewater discharges and in biosolids research and development to support beneficial reuse to land. She is currently the Service Planning Lead for biosolids and other residuals. Judith holds a Bachelor of Applied Science (Environmental), two UTS prizes in aquatic and terrestrial ecology and a Master of Science (by thesis).

Rod Kerr has over 27 years of water industry experience within the New South Wales Environment Protection Authority and now Sydney Water. He has worked in various scientific and management roles. He has developed and implemented award winning wastewater management strategies. He is currently the Service Planning Lead for wastewater. His specialties are Sewer Asset Management, Environmental and Public Health Protection, Media and Government Relations.

Dr. Matt Zhang is a senior research scientist and team leader at Data61 CSIRO. He received his Ph.D. degree in computer science from University of New South Wales. His research interests focus on machine learning, data mining and data-driven smart urban planning. He has extensive experience in utilizing advanced data analytics techniques to help industry improve productivity. He led and contributed to the collaborative projects with 30 water utilities globally for providing predictive asset maintenance solutions. He also serves as a reviewer for prestigious international academic journals, such as IEEE Trans. Knowledge and Data Engineering (TKDE) and IEEE Trans. Image Processing (TIP). He was the finalist of the young water professional of the year in AWA NSW 2016.

REFERENCES

Roberts, J., Jackson, N., and Smith, M., 2006. Tree Roots in the Built Environment. Centre for Ecology and Hydrology, Natural Environment Research Council, Great Britain.

Sinclair Knight Merz, 2005. Sydney Water Sewerage Network Choke Management Strategy, Sydney Water.

McGowan, M., Mitchell, C., 2015. Using tree canopy mapping to improve efficiency of wastewater network CCTV programs. OzWater'15, Sydney Water, NSW, Australia.

Franks, S., 1999. Disaggregation of environmental factors affecting sewer pipe failures, Journal of infrastructure Systems 5.4.

Bailey, J., et al., 2016. Developing Decision Tree Models to Create a Predictive Blockage Likelihood Model for Real-World Wastewater Networks, procedia Engineering 154.

Guo, T., Zhang, B., Wang, Y., et al., 2016. Data-driven reticulation water main failure prediction. OzWater'17, CSIRO, Data61, NSW, Australia.

Lin, P., Zhang, B., Guo. T, et al., 2016. Interaction Point Processes via Infinite Branching Model. The 13th AAAI Conference on Artificial Intelligence, CSIRO, Data61, NSW, Australia.

Lin, P., Zhang, B., Wang, Y., et al., 2015. Data Driven Water Pipe Failure Prediction: A Bayesian Nonparametric Approach. CIKM, CSIRO, Data61, NSW, Australia.

Li, Z., Zhang, B., Wang, Y., et al., 2014. Water pipe condition assessment: a hierarchical beta process approach for sparse incident data. ICML 2014, Beijing, China.

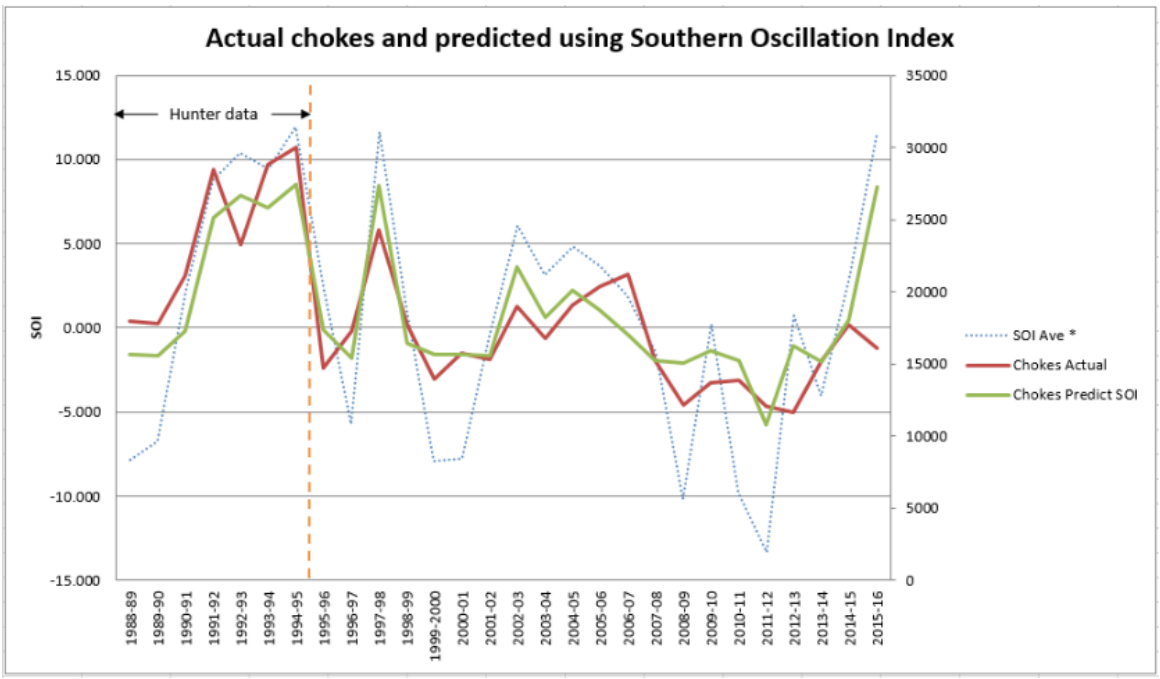


Figure 1 . Overall number of chokes (actual and predicted) for Sydney Water

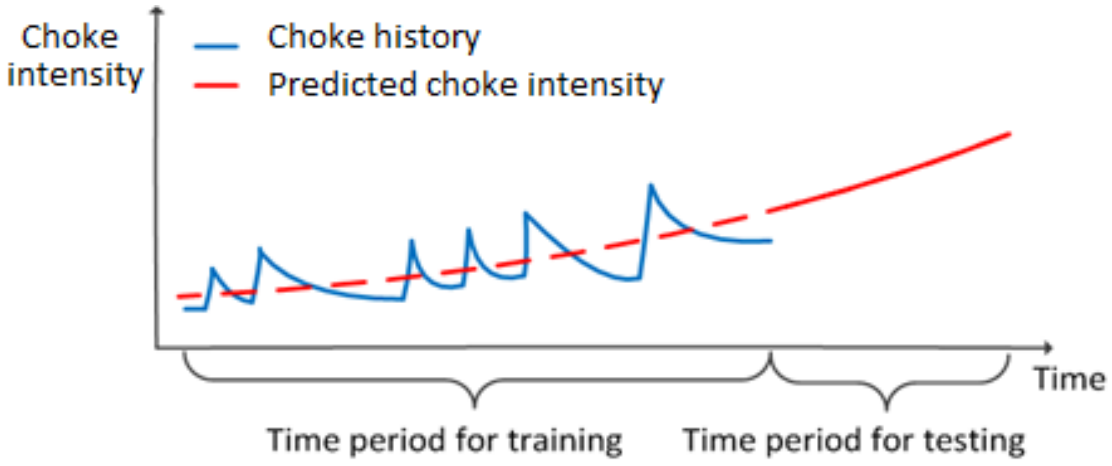


Figure 2. Illustration of modelling training and model testing (validation)

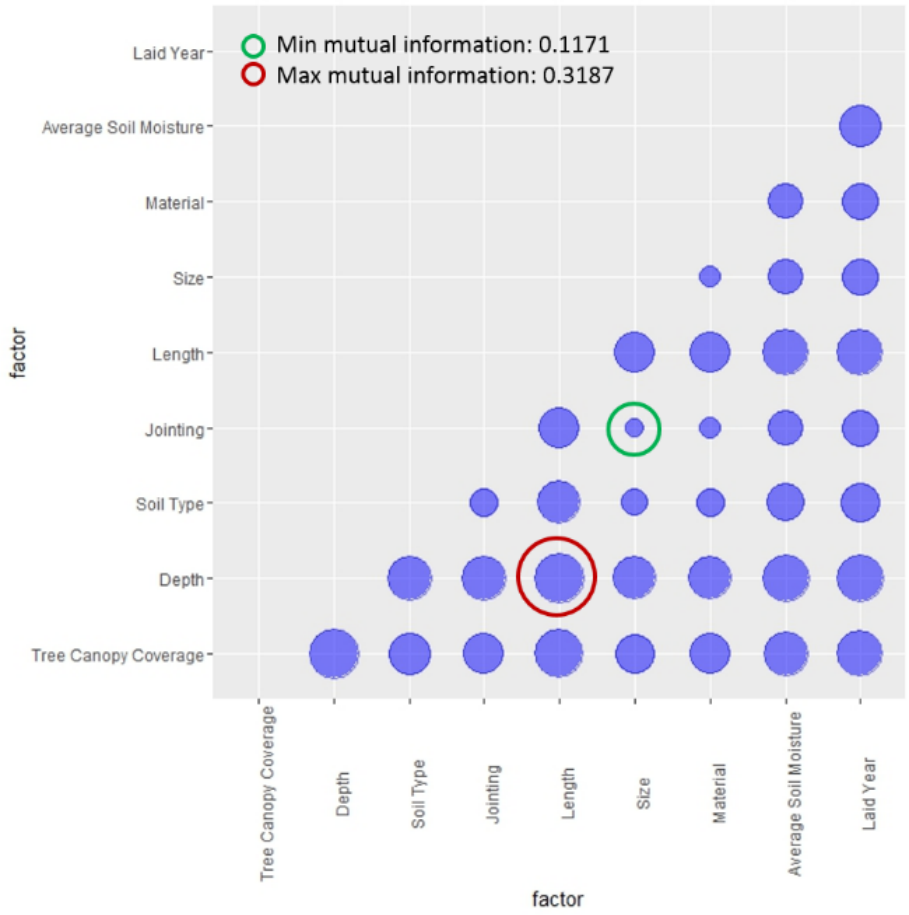


Figure 3. Mutual Information of factor combinations (importance of two factor combinations) related to tree root chokes. High mutual information indicates high importance.

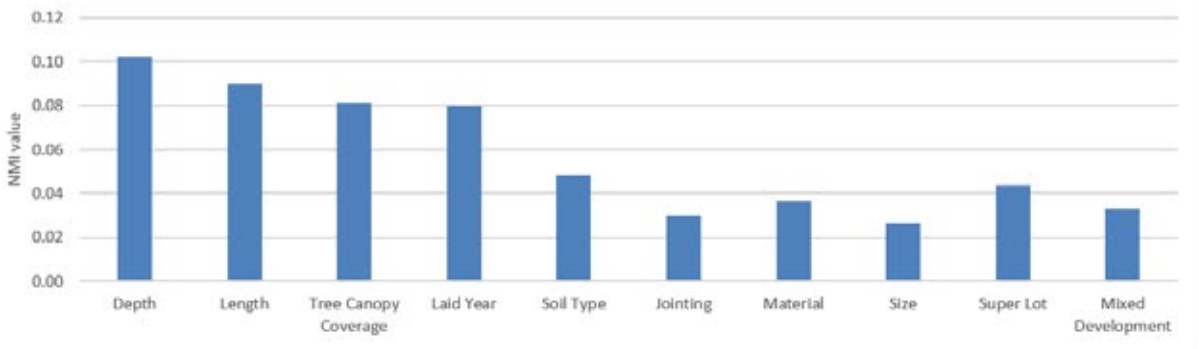


Figure 4. Mutual Information result of factors related to debris chokes (importance of factors). High mutual information indicates high importance.

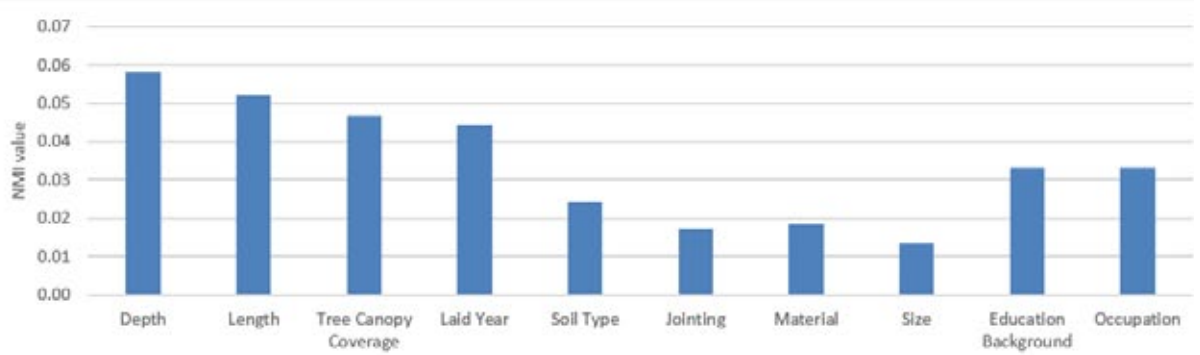


Figure 5. Mutual Information result of factors related to grease chokes (importance of factors). High mutual information indicates high importance.

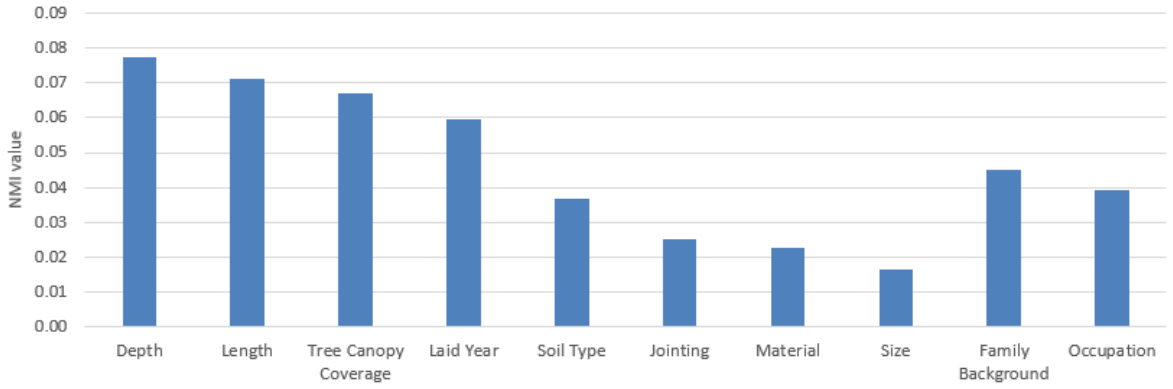


Figure 6. Mutual Information result of factors related to soft chokes (importance of factors). High mutual information indicates high importance.

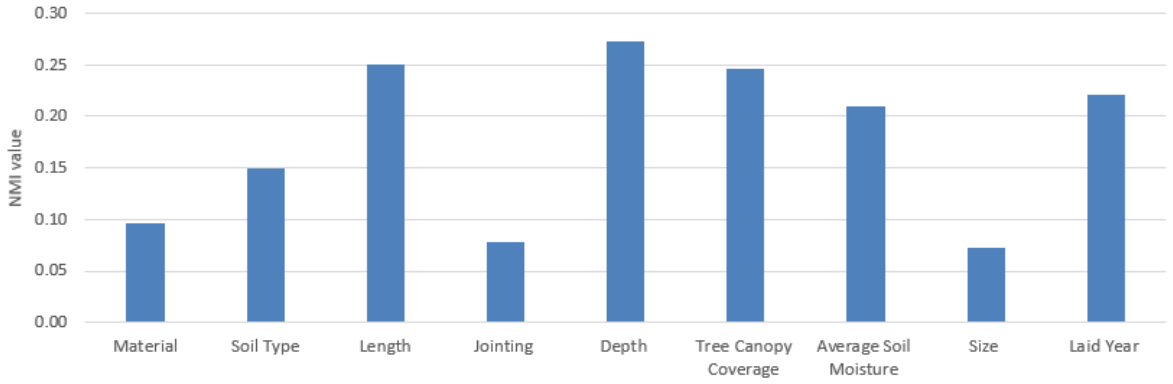


Figure 7. Mutual Information result of factors related to tree root chokes (importance of factors). High mutual information indicates high importance.

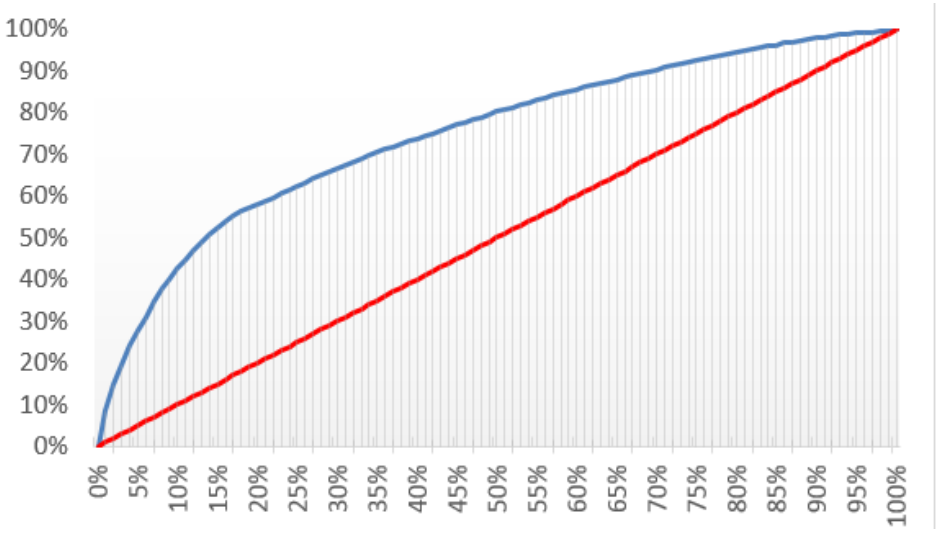


Figure 8. Prediction performance curves on the whole network.
The x-axis indicates the percentage of the inspected pipe length over the whole network length. The y-axis indicates the percentage of correctly detected blockages over the number of blockages in 2015.

The blue line represents the forecast performance of the model developed, the red line represents the likely performance from random maintenance. The current performance of Sydney Water is likely to be represented by a line lying between the red and the blue lines.

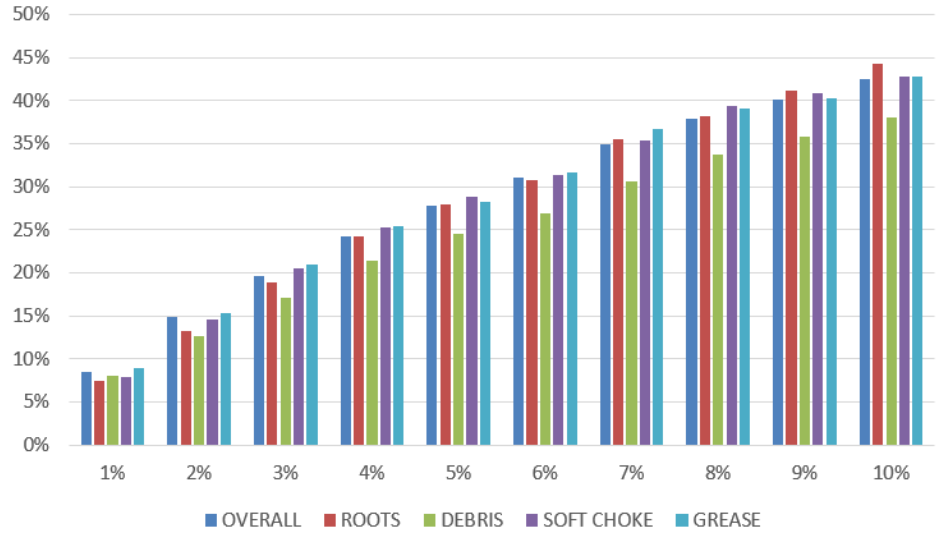


Figure 9. Choke prediction based on predicted choke probability (zoom in to top 10% of total network length) on 2015. The x-axis indicates the percentage of the inspected pipe length over the whole network length. The y-axis indicates the percentage of correctly detected blockages over the number of blockages in 2015.

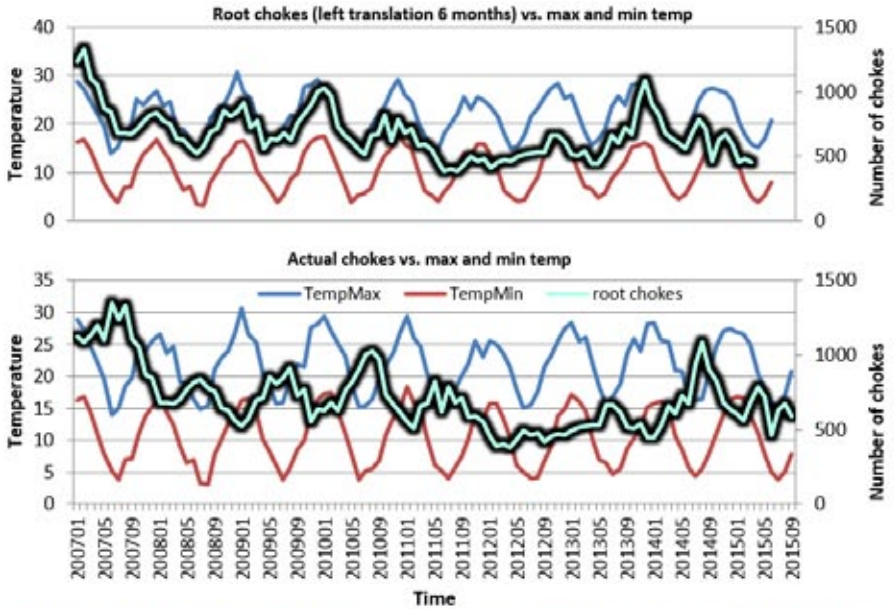


Figure 10. Root choke vs. Temperature. The correlation between the number of tree root chokes and the present maximum temperature is 0.3441. The correlation between the number of tree root chokes and the maximum temperature 6 months ago is 0.490519, which is much larger than the correlation with the present maximum temperature. It indicates that the high rate of root chokes in winters may be due to root growth in the previous summer seasons.

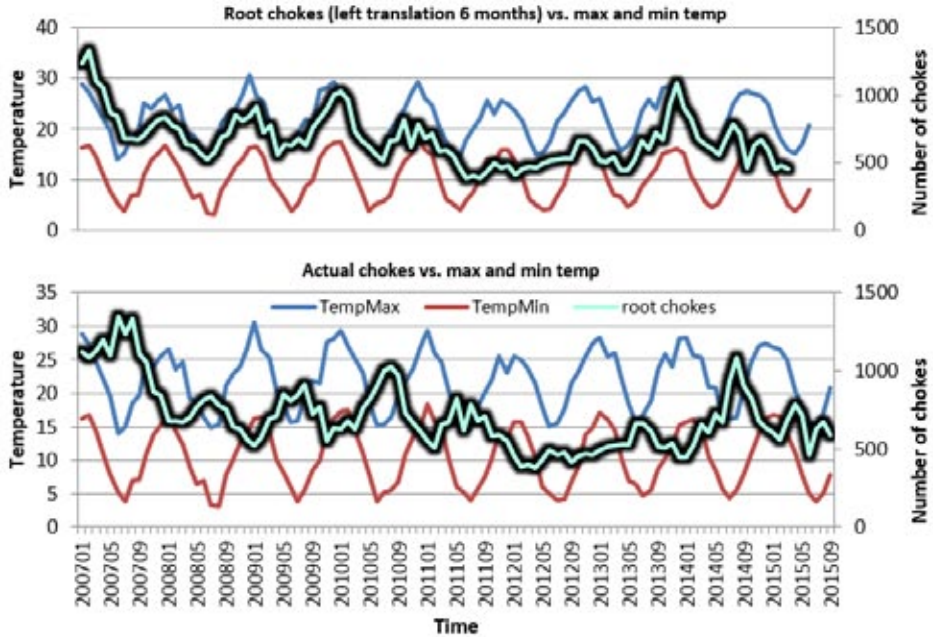


Figure 11. Root choke vs. Relative soil moisture, where WRel1End indicates the upper layer of relative soil moisture at end of aggregation period (to 0.2m) and WRel2End indicates the lower layer of relative soil moisture at end of aggregation period (0.2-1.5m).

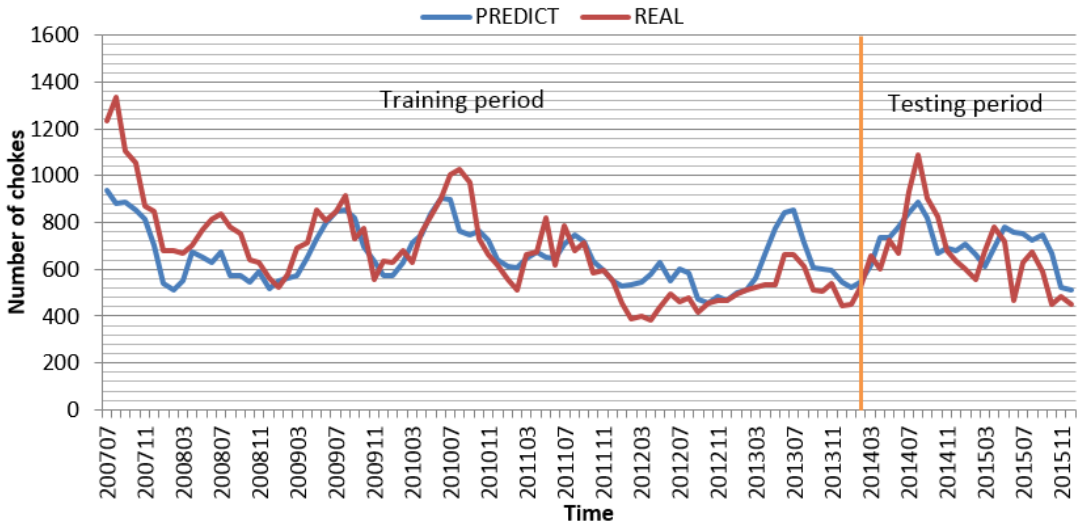


Figure 12. Actual root caused chokes and predicted using climate & soil factors. The mean absolute error (MAE) between the predicted and actual number of root chokes is 96.094.