PREDICATIVE ANALYTICS TOOLKIT FOR H_2S ESTIMATION AND SEWER CORROSION

B. Li [1], X. Fan [1], J. Zhang [1], Y. Wang [1], F. Chen [1], S. Kodagoda [2], T. Wells [3],
L. Vorreiter [4], D. Vitanage [4], G. Iori [4], D. Cunningham [4] and T. Chen [4]
1. Data61, CSIRO, 13 Garden Street, Eveleigh NSW 2015, Australia
2. University of Technology Sydney, 15 Broadway, Ultimo NSW 2007, Australia
3 University of Newcastle, Callaghan NSW 2308, Australia
4. Sydney Water, 1 Smith Street, Parramatta NSW 2150, Australia

ABSTRACT

This paper presents a predictive analytics toolkit, which is based on the emerging spatiotemporal data analysis techniques, for the estimation of hydrogen sulphide (H_2S) gas distribution and prediction of sewer concrete corrosion level. The toolkit is an easy-to-use desktop application with a user-friendly interface for querying and producing output results on GIS. The inputs to the toolkit are the sewer network geometry, monitored factors, and hydraulic information; the outputs of the toolkit are spatiotemporal estimates of H_2S gas concentration and concrete corrosion levels on the entire sewer network with uncertainties of the predictions. The toolkit is also able to integrate experts' domain knowledge or existing physical model's results as prior knowledge into the analytics model. The final outcomes of the toolkit can be used to prioritise high risk areas, recommend chemical dosing locations, and suggest deployment of sensors. A simulation of H_2S and corrosion level prediction on a subsystem of the sewer network in the greater Sydney area is reported to demonstrate the capability of the toolkit.

INTRODUCTION

Sewer corrosion is a serious problem in wastewater systems worldwide, particularly in warm climate countries such as Australia. Therefore predicting sewer corrosion is a critical task for water utilities around the globe in order to improve efficiency and save costs in chemical dosing, sewer pipe rehabilitation and sensor deployment. As sewer corrosion occurs in the presence of gaseous hydrogen sulphide (H_2S) generated from sulphur compounds in the sewage, a new and reliable toolkit is being developed in this work which enables spatiotemporal estimation of H_2S. Based on the H_2S estimation, the toolkit could further predict sewer corrosion level over the entire sewer network.

However reliable prediction of sewer corrosion has often been hampered by insufficient observations for accurate modelling– A problem commonly referred to as "sparsity" in data analytics. Therefore, analytical modelling of spatiotemporal H_2S distribution over the entire sewer network is nontrivial. Increasing the H_2S monitoring stations is also not feasible due to cost and accessibility. Therefore, in this work an attempt was made to use emerging data analytics techniques to estimate the spatiotemporal distribution of H_2S with limited number of observations. The model does not only estimate the H_2S quantity but also estimates the uncertainty associated with the prediction, which is an important measure in decision making. These H_2S quantities will be used in the overall data driven corrosion model. The final outcome of the prediction model includes the corrosion levels on the entire sewer network and uncertainties of the predictions, which can be used to prioritise high risk areas, recommend chemical dosing locations, and suggest deployment of sensors.

The predictive analytics toolkit being developed has the following features:

- The toolkit is a desktop application with a user-friendly interface for inputting queries and outputting results on GIS. For those utilities that do not have GIS, spreadsheets/look-up tables with the results for the sewer assets can be outputted. The toolkit can be easily used by utility staffs involved in asset management, sewer operation and planning. No special skills are required for a user to operate the toolkit, except for the general knowledge in sewer corrosion to collect the data for input and read the output results.

- The toolkit is able to perform spatiotemporal factor (e.g. H_2S and temperature) estimation on the entire sewer network, based on H_2S data collected from a limited number of monitoring sites. Based on the predicted spatiotemporal factors and observed corrosion levels, the toolkit is also able to further predict corrosion levels on the entire network. Both H_2S and corrosion level predictions are associated with uncertainties of prediction (or confidence).

- The toolkit is able to integrate experts' domain knowledge or physical model into the analytics model. The adopted data analytics technique is a Bayesian nonparametric model which provides a way to regularise the prediction with domain knowledge. In particular, the analytics model can use the predictions of the physical model (Wells & Melchers, 2016) as the prior knowledge that imposes restriction on the range of the prediction.

- The output of the predictive analytics, the spatiotemporal H_2S estimation and corrosion level prediction are used to prioritise high-risk areas, adjust chemical dosing profiles, and optimise sensor deployment. All these functions are also enabled in the toolkit, supported by the background data analytics model.

OVERVIEW OF THE TOOLKIT

The toolkit is the outcome of the collaborative project between Data61, University of Technology Sydney, University of Newcastle and Sydney Water, aiming to look at the applicability of data analytics to develop a new and reliable toolkit and enable more useful features for corrosion and odour management. The work is also built on the current knowledge from the Corrosion and Odour (SCORe) research project jointly funded by the Australian government and major water utilities in Australia.

The toolkit is a desktop application with a user-friendly interface for querying and producing output results on GIS. The input of the toolkit include the sewer network system (GIS), monitored/sampled factors, hydraulic information, and it can also incorporate existing corrosion model's results as prior knowledge. The toolkit (overview is given in Figure 1) will use data analytics techniques to enable: (1) Spatiotemporal Corrosion Prediction over the entire sewer network; (2) H_2S (and other parameters) Estimation; (3) Smart Chemical Dosing Optimisation; and (4) Optimal Sensor Deployment. In the following, we will give an illustration for these functional modules.

(1) Spatiotemporal H_2S Estimation

We apply the toolkit on a subsystem of sewer in Sydney. Figure 2 illustrates the sewer network and a number of different observation sites on the sewer network. There are 17 H_2S observation sites at a monitoring frequency of 15 minutes from Jan 2011 to Dec 2015. The data analytics model is to estimate the spatiotemporal dynamics of H_2S on the entire network over time and visualise it via animation on the map. Figure 3 illustrates a frame of the animation which plots the H_2S distribution on the network at 01:15:00, 15-Sep-2015.

(2) Corrosion Prediction

Based on the estimated H2S and other monitored or estimated factors, the toolkit can integrate physical model (or experts' domain knowledge) to predict corrosion levels with uncertainty on the entire sewer network. Figure 4 provides an illustration of corrosion prediction, where three corrosion levels (High, Medium, and Low) are denoted in three colours while the prediction uncertainty is denoted in thickness.

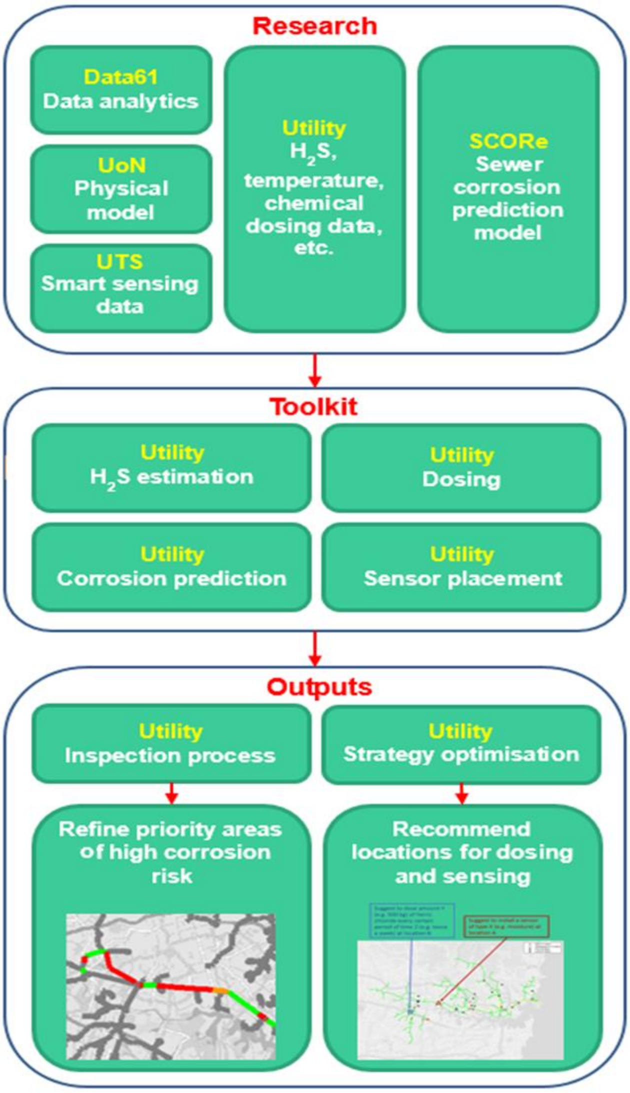

Figure 1: Overview of the toolkit.

(3) Smart Dosing

The estimation of H_2S and predicted corrosion levels can also enable the features for dosing strategy. Given a budget, the locations and amounts of chemical dosing can be optimised according to the H_2S concentration, sewer corrosion level, and hydraulic information on the sewer network. For example, in Figure 5, the toolkit suggests to dose certain amount of chemical every certain time period for Location B on the sewer network.

(4) Monitoring

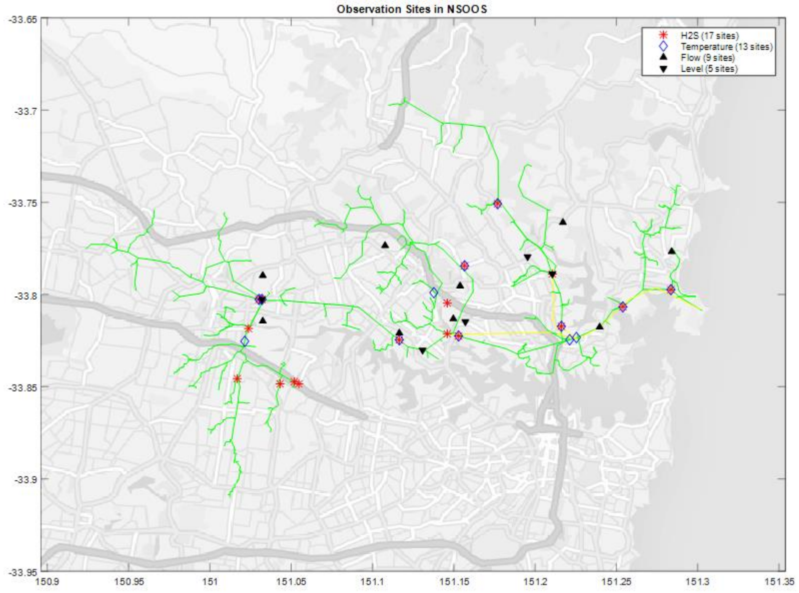

Figure 2: A subsystem of sewer network in Sydney.

Strategic deployment of sensors can maximise the monitoring capability on the sewer network. New sensors can be installed at locations with high uncertainty of H_2S estimation obtained from spatiotemporal factor prediction and corrosion prediction phases. For example, in Figure 5, the toolkit suggests Location A on the sewer network to install a certain type of sensor.

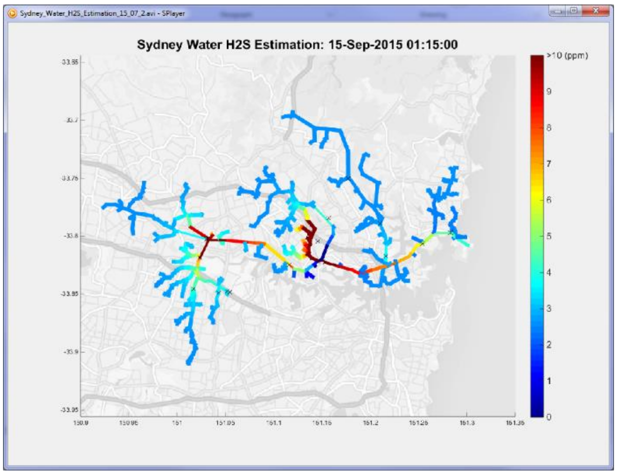
Sydney_Water_H2S_Estimation_15_07_2.avi - SPlayer
Sydney Water H2S Estimation: 15-Sep-2015 01:15:00
>10 (ppm)
9
8
7
6
5
4
3
2
1
0

Figure 3: Spatiotemporal estimation of H2S in the entire network over time visualised via animation. The plot illustrates a frame of the video which plots the H2S distribution on the network at 01:15:00, 15-Sep-2015.

METHODOLOGY

The core module underpinning the toolkit is an analytics model based on Bayesian nonparametric method for spatiotemporal estimation. A typical Bayesian model is in the form of “Prediction = Prior Knowledge × Data Likelihood”, where “Prior Knowledge” provides a hypothesis space to the model such that the model is not only driven by the data (in terms of “Data Likelihood”) when data are sufficient, but does not deviate too far from the domain expert’s hypothesis when data are insufficient. Through Bayesian modelling, we can thus (1) integrate domain experts’ knowledge or existing H_2S simulation results as prior knowledge and (2) predict the H_2S estimation result as a posterior distribution, whose variance can be viewed as the uncertainty of the prediction. In the following, we will give an introduction to the data analytics model for spatiotemporal H_2S and corrosion level prediction on the entire network.

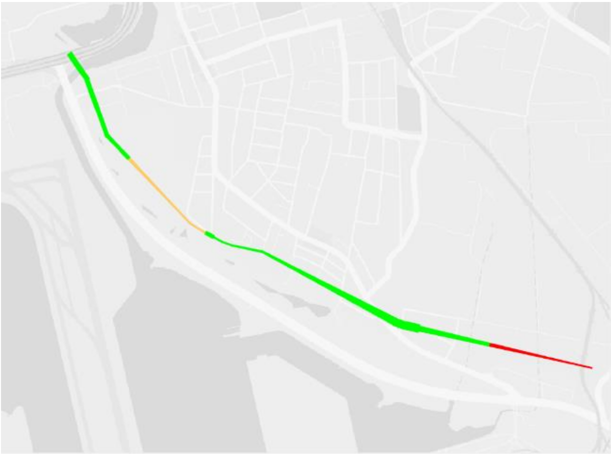

Figure 4: Illustration of corrosion prediction, where three corrosion levels (High, Medium, and Low) are denoted in

red, orange, and green. The prediction uncertainty is denoted by the thickness (higher thickneses are represented by higher uncertainties).

Spatiotemporal H_2S Estimation

We consider a spatiotemporal analytics model which is able to estimate H_2S concentrations on the entire sewer network over time. For each time stamp, H_2S concentration of any point (corresponding to a sewer asset) on the sewer network can be estimated as a weighted combination of H_2S of all the observed assets:

$$H_2S(U) = w_{A\to U}H_2S(A) + w_{B\to U}H_2S(B) + w_{C\to U}H_2S(C) + \cdots \tag{1}$$

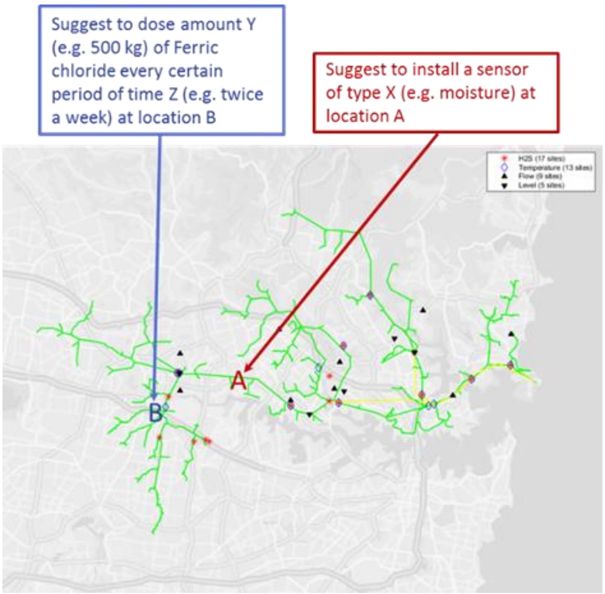
Suggest to dose amount Y (e.g. 500 kg) of Ferric chloride every certain period of time Z (e.g. twice a week) at location B
Suggest to install a sensor of type X (e.g. moisture) at location A
H2S (17 sites)
Temperature (13 sites)
Flow (9 sites)
Level (5 sites)
A
B

Figure 5: Illustration of online smart chemical dosing (Location B) and senor deployment (Location A) recommendations.

where $H_2S(U)$ denotes predicted H_2S at any unknown point on the sewer network (e.g., green dots in Figure 6) and $H_2S(A)$, $H_2S(B)$, $H_2S(C), \ldots$ denote those points with observed H_2S (e.g., the three red dots in Figure 6). It is worth noting that in Eq.(1) the weights $w_{A\to U}$, $w_{B\to U}$, $w_{C\to U}, \ldots$ are learned automatically through using a Bayesian method and the resulting weights have the following properties:

- Weight $w_{A\to U}$ is inversely proportional to the geodesic distance from point A to point U.

- Weight $w_{A\to U}$ is impacted by flow direction, if flow runs downstream from A to U, then $w_{A\to U} > w_{U\to A}$.

Gaussian Process based Analytics Model

The spatiotemporal estimation problem introduced above is essentially a regression problem in data analytics. To best estimate the values of unknown points given some observed ones, we adopt a Bayesian nonparametric model, named Gaussian Process (GP) (Rasmussen, 2004), to achieve this goal. Gaussian process computes posterior predictive distributions for unknown points based on the known data at the observed points. The final solution of GP has the same form of Eq.(1). Each weight in Eq.(1) is a function of a covariance matrix. A covariance matrix measures the pairwise similarity between different sewer assets. There are various ways to define the similarity between a pair sewer assets, denoted as A and B for convenience. Geodesic distance is used in this paper. In specific, if the geodesic distance between A and B is large, the similarity between them is small, and vice versa. This property implies that if A is far away from B, it will have small impact on B. In addition, we also incorporate the flow direction into the similarity measure. For example, if the flow runs from A to B, the similarity between A and B is larger than the similarity between B and A. An intuitive interpretation is that the upstream A could affect B more than B affects A.

Corrosion Level Prediction

As identified in the study of (Wells & Melchers, 2016), H_2S, temperature, and humidity are critical driving factors for sewer corrosion; the relationship between the driving factors and the corrosion rate has been quantitatively analysed. Since temperature is also a driving factor, we also estimate the temperature on the entire sewer network using the same approach to H_2S estimation. Once H_2S, temperature, and other factors have been estimated on the entire sewer network based on the approach introduced above, it enables us to further predict the corrosion level on the entire network.

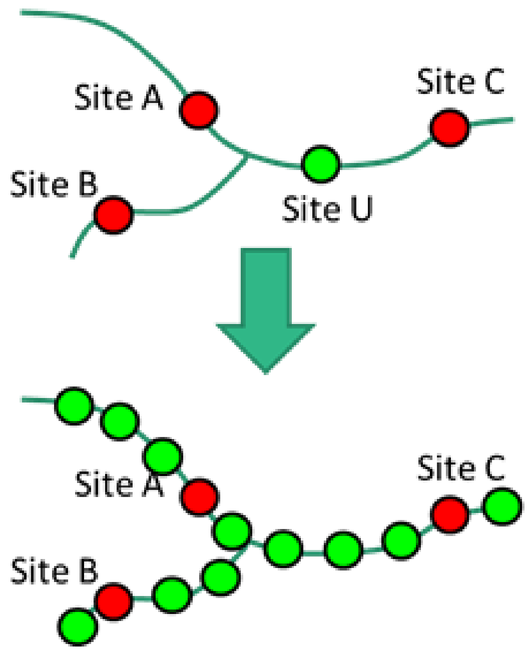

Figure 6: H2S estimation on the sewer network. H2S of any point (green dot) on the network can be estimated as a weighted combination of H2S on the observed points (red dots).

The Gaussian Process based analytics model is again exploited for corrosion level prediction. But this time more factors should be considered than those analytics models for predicting individual factors (e.g. H_2S or temperature). As introduced previously, the covariance matrix in the GP for H_2S estimation only considers the geodesic distance between points A and B. This could be reasonable and sufficient for H_2S or temperature estimation since both the factors should change smoothly along the sewer network. However, this assumption may not be valid when applied to corrosion level prediction, because two close assets may have different corrosion levels due to various factors. Considering that corrosion rate is highly affected by H_2S and temperature as reported in (Wells & Melchers, 2016), both H_2S and temperature are incorporated into the corrosion level prediction model. GP allows for incorporating a variety of factors in a convenient manner by deriving a combined covariance matrix, which is defined as a linear combination of three individual covariance matrices: the pairwise covariance of A and B for geodesic distance, H_2S, and temperature, respectively. The three coefficients for linear combination are learned automatically form the training data.

Uncertainty of H_2S Prediction

Gaussian Process is a Bayesian model, which means that each prediction at the unknown point is a posterior (Gaussian) distribution with mean and variance. Mean can thus be used as prediction result while variance can be used as prediction uncertainty – A smaller (or larger) variance indicates lower (or higher) uncertainty. Uncertainty is proportional to the geodesic distances from the predicted point to the observation points (see Figure 7).

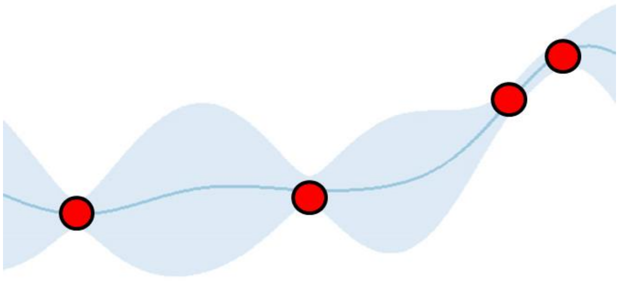

Figure 7: Illustration of prediction uncertainty of Gaussian process based analytics model. The curve denotes the mean value of the prediction and the bandwidth denotes the uncertainty. The farther the prediction point away from the observation points (red dots), the more uncertain the prediction result is.

ADVANCED SENSOR DEVELOPMENT

Surface temperature and surface moisture content provide pivotal living conditions for bacteria who is responsible for sewer concrete corrosion. A sensing system was designed for measuring the above two quantities. An infrared (IR) radiometer was used to sense surface temperature in a non-contact way. It measures the infrared radiation emitted from the surface of interest which is relating to the temperature of the surface. As the resistivity of concrete is changing with the presence of different levels of moisture content, a resistance measuring device was utilized to measure the surface moisture. Data loggers, enclosures, power systems and cabling systems were designed and developed for continuous operation of the sensing module in harsh sewer conditions. The sensing system was lab tested and deployed in a sewer belonging to the Sydney Water for more than three months.

RESULTS

Data for the Evaluation

We test the proposed corrosion prediction method in a subsystem of sewer in Sydney. Figure 2 illustrates the sewer network and a number of observation sites on the sewer network. There are 17 observation sites for monitoring H_2S and temperature at a sampling frequency of 15 minutes from Jan 2011 to Dec 2015. To predict the corrosion risk levels, we first estimate monthly H_2S and temperature distributions on the entire network over five years using the method introduced above in “Spatiotemporal H_2S Estimation”. The estimated monthly H_2S and temperature data along with the sewer geometry can then be used as the input of the analytics model introduced above in “Corrosion Level Prediction”.

Because there is no detail concrete loss data for each pipe as ground truth (i.e. observations) for calibrating the model, we adopt the Structure Grade (1~5) extracted from traverse reports as a surrogate for the corrosion level. The training data used for the evaluation is obtained from traverse reports in two periods of time: 2007-2009 and 2010-2015. In each period, a set of sewer pipes are examined and their Structure Grades were recorded. There are 17 sewer pipes which were investigated in both the periods and we can use the Structure Grades examined to calculate the corrosion rate for each sewer pipe as the ground truth for calibrating the model (see Figure 8).

Evaluation Method

We adopt the leave-one-out (LOO) evaluation, which is a popular evaluation in data analytics: We have 17 ground truth sewer pipes, which we have known their corrosion rates. At each time, we hide one pipe for evaluation and use the remaining 16 sewer pipes for training the analytics model (i.e., calibrating the model). The model trained based on the 16 sewer pipes is used to predict the corrosion rate of the hidden one. The evaluation is conducted on each of the 17 sewer pipes and the final performance is averaged over the 17 individual prediction results.

As the direct output of the analytics model is the corrosion rate (CR). To obtain the corrosion risk level, that is, the predicted Structure Grade (SG), at a certain time, we need to use the following equation:

$$\widehat{SG}(t) = SG(t_0) + CR * (\mathrm{t} - \mathrm{t}_0) \tag{4}$$

where $SG(t_0)$ denotes the known Structure Grade at time t_0 while $\widehat{SG}(t)$ denotes the predicted Structure grade at time t. In out setting, the time t_0 refers to the year in which the first traverse report was conducted and $(\mathrm{t} - \mathrm{t}_0)$ is the time difference between the two traverse reports.

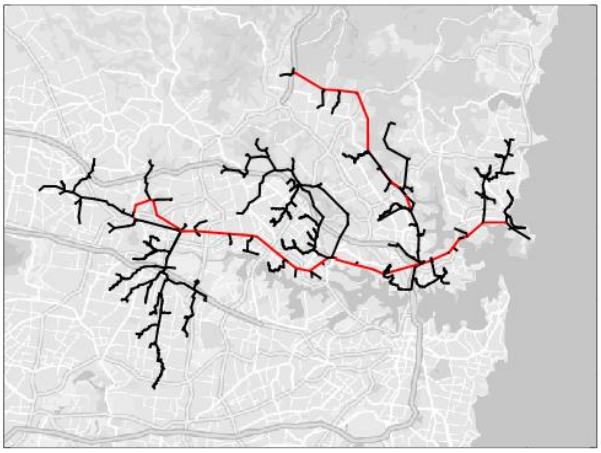

Figure 8: Segments highlighted in red represent sewer pipes with traverse report.

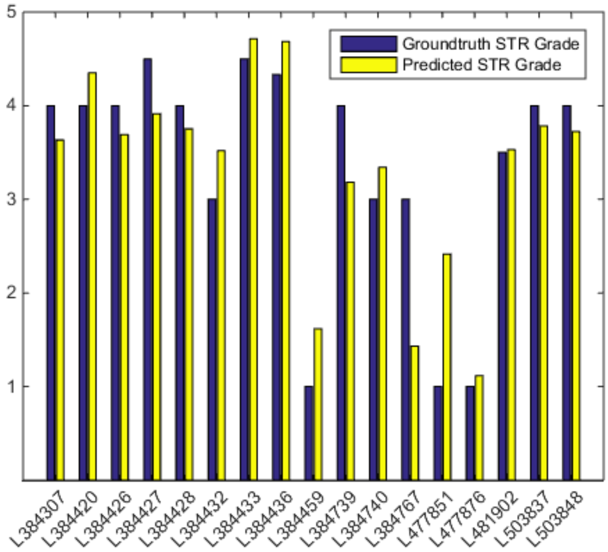

Figure 9: Corrosion risk level (Structure Grade in 1~5) prediction results for the 17 sewer pipes with ground truth Structure Grades.

Evaluation Results

Figure 9 plots the Structure Grade prediction (in yellow) for the 17 sewer pipes which have the ground truth (in blue). We can thus evaluate the prediction performance by comparing the predicted Structure Grade $\widehat{SG}(t)$ and the ground truth Structure Grade $SG(t)$. We can see that in most cases the prediction error $\left|SG(t) - \widehat{SG}(t)\right|$ is only 0.2 or less; only three prediction errors are more than 0.5.

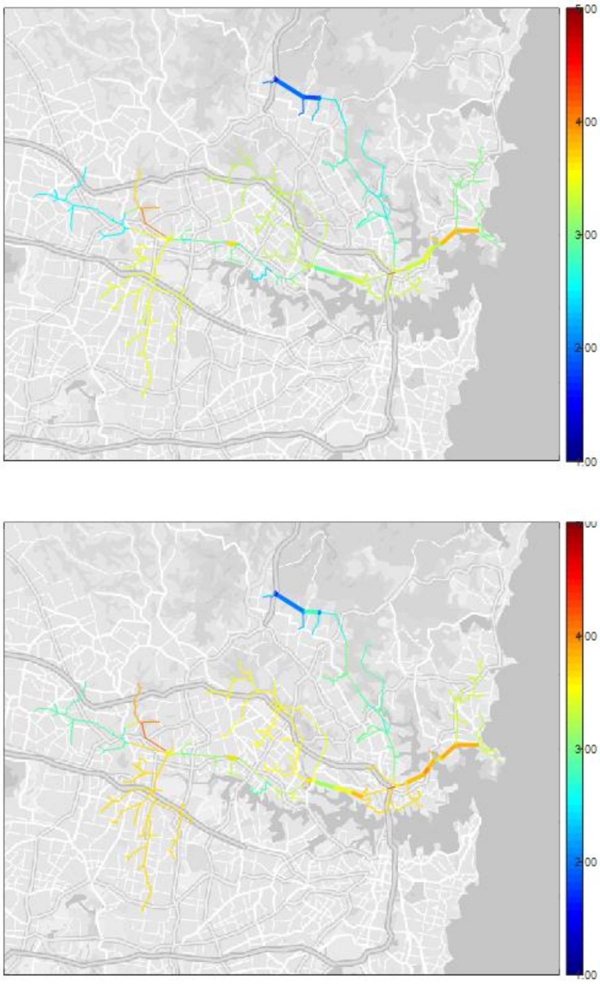

Figure 10: Corrosion risk level (Structure Grade in 1~5) prediction results on the entire sewer network over time from Jan 2011 to Dec 2015. Here we only show the result in the first month (upper panel) and the last month (bottom panel) of the five years.

As mentioned in the methodology section, the output of our sewer corrosion prediction model is relating to the prediction on the entire sewer network. In other words, one can query the Structure Grade of any asset on the network. Our model predicted the Structure Grade results on the entire network over the five years, and the prediction results in the first month (Jan 2011) and the last month (Dec 2015) of the five years are illustrated in Figure 10.

CONCLUSION

As current physical models have limited calibration sites, they cannot be generalised to entire sewer networks with variable conditions. By integrating the advantage of the physical model for monitoring sites, this work introduces a data analytics toolkit for spatiotemporal prediction over the entire sewer network, which provides a powerful complementary method for H_2S/corrosion prediction without monitored parameters. The prediction results with uncertainty can help to prioritise high risk areas, recommend chemical dosing locations, and suggest deployment of sensors.

The proposed data analytics model for corrosion risk prediction was evaluated in a Sydney sewer subsystem. The data collected consisted of a limited number of H_2S and temperature monitoring sites, the sewer geometry data, and the Structure Grade data extracted from traverse reports. The evaluation results show that the proposed analytics model can predict Structure Grade (1~5) with less than 10% error considering its range.

The implementation stage will see the Toolkit become a desktop application with a user friendly interface. Without the need for specialised training, asset management staff will be able to input specific queries relating to asset corrosion and have the choice of both GIS and non-GIS (e.g. spreadsheets, look-up tables) as output formats.

REFERENCES

Wells, T. & Melchers, R. 2016. Concrete Sewer Pipe Corrosion – Findings from an Australian Field Study. Ozwater 2016. Melbourne, Australia.

Rasmussen, C.E. 2004. Gaussian Processes in Machine Learning. Advanced Lectures on Machine Learning. Lecture Notes in Computer Science. 3176.