Long-Term Water Pipe Condition Assessment: A Semiparametric Model Using Gaussian Process and Survival Analysis

Dilusha Weeraddana\textsuperscript{1}(✉), Harini Hapuarachchi\textsuperscript{2}, Lakshitha Kumarapperuma\textsuperscript{2}, Nguyen Lu Dang Khoa\textsuperscript{1}, and Chen Cai\textsuperscript{1}

\textsuperscript{1} Data61-CSIRO, 13, Garden Street, Eveleigh, NSW, Australia
{dilusha.weeraddana,khoa.nguyen,chen.cai}@data61.csiro.au
\textsuperscript{2} Monash University, Clayton, VC, Australia
{harinipranami,lakshitha.kumarapperuma}@ieee.org

Abstract. The maintenance and renewal of water mains demand substantial financial investments, and direct inspection of all water mains in a distribution system is extremely expensive. Therefore, a cost effective break mitigation technique such as a failure forecasting model that allows one to predict the water mains failure likelihood, would reduce the negative social impact and the cost to serve. We introduce a semiparametric Bayesian model for pipeline failure forecasting. The model is centred on a nonparametric Gaussian Process Regression (GPR), and uses a parametric survival model to capture the long-term survival probability using domain knowledge. The parametric element in our model allows the inclusion of survival probability, while the nonparametric part allows us to handle covariates and to employ incomplete prior knowledge about pipe failures. We apply our model to the proactive maintenance problem using a real dataset from a water utility in Australia. The results demonstrate that, our model performs better than competing models such as Support Vector Regression, Poisson regression, Weibull, Gradient Boosting, and GPR, leading to substantial savings on reactive repairs and maintenance. Our water pipeline failure prediction models have been deployed in three states across Australia, and are being monitored by each water authority.

Keywords: Pipe failure prediction · Semiparametric · Gaussian process regression · Survival analysis

1 Introduction

1.1 The Water Pipeline Failure Issue in Australia

The consequences of water main breakages can be extremely severe in terms of water supply disruption, high repair cost and compensation claims. The breakage of a water main could have far reaching consequences, including intangible costs that can have a significant effect upon the wider community. Not only could such an event have immediate disastrous impacts to the surrounding area, there would also be long-term socio-economic impact on power, water and other supplies to suburbs. Figure 1 (a) shows a sinkhole which has swallowed cars (and three houses have been flooded) after a water
main burst in Port Melbourne. This shows potential consequences stemming from a burst pipeline. The burst failures in the water main network in a region in Australia is shown in Fig. 1 (b). Across the entire region under our study, an average number of 400 pipe failures occur each year, causing water supply disruptions and myriads of property and environmental damages. Even while this manuscript was being drafted, two water main bursts have severely affected up to 60 houses in Sydney, Australia, within just 5 days apart [1]. A recent report from the Australian Infrastructure Statistics Yearbook indicates that the maintenance cost for water assets is over 1.4 billion Australian dollars in Australia [3].

![Image of consequences of pipe failure](https://www.sbs.com.au/news/sinkhole-swallows-cars-in-port-melbourne,
2014)

**Fig. 1.** (a) An example of consequences of pipe failure. Photo taken in Melbourne, Australia. (b) Water main network and failures in a given region in Australia, generated using QGIS.

### 1.2 Proactive Water Main Maintenance Using Machine Learning

Failure mitigation and water asset renewal programs balance the consequence of water main failure and the need to minimize cost to customers. Generally, water authorities develop risk based approaches in order (1) to avoid critical water pipe failures by prioritizing renewal of specific pipes; but also (2) to avoid replacing any pipe that is still in good working condition. To effectively achieve these objectives, machine learning techniques can be used with the advancement of Artificial Intelligence. Furthermore, the prediction estimation is expected to be long-term for the structural deterioration of water mains to effectively plan the renewal of water distribution systems and to develop risk based investment decisions for capital interventions.

### 1.3 Related Work

The water pipe breakage forecasting has been studied over the past few decades using a variety of methods and frameworks, ranging from physical models [6, 10] to machine learning models [14]. Nevertheless, physical models for the failure forecasting process require extensive information and data, and include uncertainties that can accumulate errors during the modeling process. As an alternative, machine learning models that try to forecast pipeline failures based on historical data are far less expensive, and therefore have attracted a lot of interest from water authorities recently [14].
In 1979, Shamir et al. developed a statistical forecasting technique to study how the number of breaks would change with time if the pipes were not replaced [24]. In that study, the authors used a Poisson model based on the age of the pipes. Most recently, tree-based machine learning techniques have been used to predict water pipe failure rates in VIC (Victoria) [26] and (Queensland) [16], Australia and Syracuse, USA [14]. However, tree based models are generally not capable to provide accurate predictions for extremely sparse components [15], such as a critical water main. Furthermore, prediction of water main breaks has been studied using survival-based methods, such as Poisson regression and Weibull model [2]. Survival analysis focuses on the study of time-to-event data, usually called survival times. These models are usually limited by their fixed model structure and their inability to adaptively adjust the model to the complexity of the problem. There is also a vast literature of estimating hazard function in survival analysis using Bayesian nonparametric techniques [11]. Even though these algorithms are flexible, it cannot incorporate expert knowledge which is fundamental for applications like pipe failure prediction [7]. The work reported in [8] defines a Gaussian process prior over the hazard function which models the hazard function as the multiplication of a parametric baseline hazard and a nonparametric part. In our proposed method, we model survival function on a set of clusters (based on pipe laid years) to predict the future pipe breakage. This parametric component is then embedded into the prior and likelihood of a Gaussian Process.

Although there is significant existing literature, there still exist open questions regarding intricate relationships among the major factors causing pipe failure, and their long-term effect on the life-time of a pipe. Thus, prediction of water main breaks becomes a complicated task due to their low failure rate and high cost of inspection, which have led to sparse historical data. As such, the underlying function which governs the pipeline failure rates is fairly sophisticated, making it difficult to obtain a suitable model by looking only at the input data. Therefore, a nonparametric prediction method like GPR is more suitable [19]. Such methods allow the data to speak for itself and require minimal assumptions on the data. Nevertheless, regression can face issues when dataset is very sparse. When we consider the datasets obtained from Australian water utilities, the general 10–12 year observation period was relatively short compared to the life cycle of water pipes (which could be more than 100 years), so that most (about 99%) pipes do not fail or fail just once during the observation period, making the data sparse (despite the low failure rate, the failure of a critical water main could lead to heavy financial losses). Trying to learn a flexible model with a few data points may result in overfitting where the model mistakes artefacts of the specific available samples as actual properties of the underlying distribution [4,25]. When forecasting water main failures we encounter this issue, with the sparse incident data. Our proposed approach aims to solve limitations of these commonly used machine learning techniques.

1.4 Our Contribution

In this paper we propose a semiparametric Bayesian model using survival analysis. The proposed model is simple yet effective for real-world industrial applications. Our approach models survival analysis on a set of clusters to predict the future pipe breakage. Then the resulting information is embedded into the prior and likelihood of the GPR.
The main aims of this work are: (1) For sparse pipe failure data, develop an efficient algorithm based on survival analysis and GPR, (2) Apply our algorithm to a real-world dataset obtained from Australian water utilities for water pipe condition assessment, (3) Compare ensuing results of the proposed algorithm with other machine learning techniques used in the industry and literature, (4) Interpret the outcomes in terms usable by the water utilities. In addition, as our model is based on GPR which is a special class of nonparametric Bayesian probabilistic modelling, it provides a distribution instead of a point forecast. Therefore, we will have a point forecast (the mean of the distribution) as well as the uncertainty which can be obtained from the variance of the distribution. As a result, for each pipe failure probability that is predicted with a GPR, we are given the perceived uncertainty of that prediction.

Currently, our predictive data analytic models are deployed in few states in Australia for short-term prediction purposes. Each of these Australian water utilities are monitoring the number and the location of water main failures to validate our model. They also use our model in their internal financial modelling, risk distribution assessment planning and also to assist in the development of condition assessment programs. Our results indicate that proposed method opens up a new avenue for robust pipe failure prediction, and we plan to incorporate this model to the future implementations.

Fig. 2. Schematics for (a) long-term failure prediction problem, (b) end to end process automated within Docker engine for the ease of use for the end users. User only needs to provide the input data files. System will go through several steps ranging from feature extraction, data cleaning, data matching, and pattern recognition to failure forecasting.

2 Our Approach

We observed a fixed time period for all mains (e.g. 10 years), thus the observation period for older mains (where the age is greater than 10 years) on the age axis can be represented by Pipe 1 in Fig. 2 (a), and a younger water main can be represented by Pipe 4. One challenge is to predict whether the Pipe 4 will fail on age $t^*_1$ to $t^*_2$ in the future. Another challenge is to forecast failure likelihoods of older pipes a few years into the future (e.g. predict whether the Pipe 1 will fail at the age of $t^*_n$). For the first case, an intuitive method can be used to determine the failure rate at the age of $t^*_1$ and $t^*_2$ using the observed failure rates of Pipes 1–3. However, what is the relation between pipe $i$ and $j$, which we will denote as $k(i, j)$? Thus, we need to know how to transfer $k(3, 4)$ and $k(1, 3)$ into $k(1, 4)$. In more a complicated scenario, we can write $k(i, j)$ as $k_i(j, i)$.
to assume the variation of relationship with time. The intricate interplay among these underlying functions is sophisticated, making it difficult to obtain a suitable model by looking only at the input data such as pipeline features and corresponding failure rates. Therefore, a nonparametric prediction method such as GPR is more suitable [19] to handle this sort of prediction tasks. In the second scenario, when the prediction duration is very long, we do not have much data for the overlapped age period (or in the worst case scenario, there is no overlapping age period as in \( t^* \)). In order to tackle this challenge, we employed a survival based technique to model the age based relationship and embedded the results into the prior and likelihood of the GPR. This will be further described in the next section.

2.1 Proposed Method for Long-Term Pipe Failure Prediction

The framework of the proposed model is depicted in Fig. 2(b). The first step entails pre-processing pipe attribute data and pipe failure data obtained from the internal database of the water utility. In the next step, the influential and significant factors are investigated and a water main failure prediction model is developed using a semiparametric machine learning model. Then the performance of the model is evaluated. Finally, a long-term failure forecasting model is proposed, with an end-to-end runnable tool to automate the entire prediction process. The following subsections discuss the aforementioned processes.

Data Preparation and Factor Analysis. There are three main data sources used as the input to the analytical model:

1. Network data: this dataset describes water main information such as asset number, installation date, material, diameter, length, and location.
2. Work order data: this dataset contains the water main failure information such as asset number, failure date, location, and failure type (burst, fitting).
3. External data: this includes information in addition to assets, such as weather data from Bureau of Meteorology and census data from Australian Bureau of Statistics.

The above data should be sufficiently accurate for the intended use, so a data quality review has been undertaken based on three key characteristics: completeness, validation, and consistency (examination for invalid values). The quality review demonstrates that the data is sufficient and accurate for further analysis. Accordingly, this process allows to establish a comprehensive data file with complete information for each asset that can be used as an important input to further analysis. In addition, environmental and demographic factors need to be matched with the network data. Specifically, failure records and information are assigned to the corresponding assets based on the work order number, and environmental and demographic information are assigned to the assets based on the geographic locations. To quantify the underlying factors affecting pipe failures, in each of the factor in isolation, we calculate the mutual information between the ‘Pipe Failure’ parameter and each feature (we have selected a primary set of asset specific features). Pipe size (or diameter) shares the highest amount of mutual information with failures while pipe type has the least effect on failures.
In general, all predictors by themselves display very low levels of mutual information indicating that by themselves, they do not predict failures sufficiently well. However, as we shall show later in the experiment section, the six features (material, diameter, age, length, historical failures, laid year) in unison will provide us with a much more accurate prediction model of pipe failures.

**Survival Analysis.** In addition to collectively considering the aforementioned feature set, we intend to further improve the proposed GPR procedure by incorporating the estimated survival probability of each pipe in the prior and likelihood of the predictive posterior. For this purpose, we aim to populate a newly calculated survival probability estimate feature in both our training and test datasets. The Kaplan-Meier estimate is considered as one of the most widely used non-parametric estimators of the survival function $S(t)$. It has often been utilized to estimate the survival probabilities when the inputs include censored data [5, 20]. For censored subjects, the event occurred (or will occur) sometime after the date of last follow-up. However, it is not customary to ignore these subjects, as they provide some information about survival: it is known that they survived beyond a certain point, although the exact timing of the event is unknown [9]. We will first utilize the available training data to obtain the non-parametric Kaplan-Meier survival estimate as [13],

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - f_i}{n_i},$$

(1)

where $f_i$ denotes number of failure events at time $t$, $t_i$ denotes a time when at least one event occurred and $n_i$ is the number of subjects at risk of failure just prior to time $t$. Using this model, we can successfully obtain the survival probability estimates for all pipe ages spanned by the training dataset. We would need to extrapolate the estimated survival probability function to traverse beyond the training age span, for the purpose of testing the final GPR model. As the Kaplan-Meier model is not suitable for the purpose of extrapolation, we sought a parametric model which closely aligned with the Kaplan-Meier estimate along the training age span. As we shall show later in our results section, the Weibull survival probability estimate [23] fulfills this requirement. The Weibull survival probability is a continuous function featured by two parameters, namely, the scale parameter and the shape parameter. Due to these two parameters functioning simultaneously, the Weibull model provides a parametric distribution with additional flexibility compared to models with single parameters such as the exponential distribution [17]. The Weibull survival probability estimate is obtainable as [22],

$$\hat{S}_w(t) = p(T > t) = \exp \left[ -\left( \frac{t}{\alpha} \right)^\beta \right]$$

(2)

where $\alpha > 0$ and $\beta > 0$ denote the scale and shape parameters. We aim to use this model, fed by the training dataset, for the purpose of estimating the survival probabilities in the training age span as well as for the aforementioned purpose of extrapolation. The proposed procedure is outlined in the Algorithm 1.
Long-Term Failure Forecasting Using Nonparametric Gaussian Process. At the next stage of our model, we assume that the function $f$ evaluated at a point $x$ is random and drawn from a multivariate Gaussian distribution [19], where $i$ and $j$ represent different assets. The survival probabilities calculated in the previous section are embedded in $x$ (along with other pipe specific features), and used to generate the prior and likelihood of the GPR. The response variable contains binary values, 0 or 1 to indicate the presence of a failure event.

\[
\begin{bmatrix}
  f(x_i) \\
  f(x_j) \\
  \vdots
\end{bmatrix} \sim \mathcal{N}
\begin{pmatrix}
  \begin{bmatrix}
    m(x_i) \\
    m(x_j) \\
    \vdots
  \end{bmatrix},
  \begin{bmatrix}
    k(x_i, x_i) & k(x_i, x_j) & \cdots \\
    k(x_j, x_i) & k(x_j, x_j) & \cdots \\
    \vdots & \vdots & \ddots
  \end{bmatrix}
\end{pmatrix}
\]

where $m$ and $k$ are the mean and covariance function, respectively. Since the function $f$ is random, the optimum $f$ will be inferred via the Bayesian inference. Therefore, we compute the posterior over all the random functions. As the likelihood and prior are Gaussian, the posterior over functions is also another Gaussian given by,

\[
p(f|x, y) \sim \mathcal{N}(f|\mu, \xi)
\]

where $f$ is a vector containing all the random functions evaluated at training data input vector, $x$ is the feature vector and $y$ denotes the training response variable (our case is binary indicating whether the pipe was failed or not).

Next, we can compute the predictive posterior over all the random functions considering all the pipes.

\[
p(y_\ast|X_\ast, X, y) = \int p(y_\ast|X_\ast, f, X)p(f|X, y)df
\]

The predictive distribution is again Gaussian, with a mean $\mu_\ast$, and covariance $\xi_\ast$.

\[
p(y_\ast|X_\ast, X, y) \sim \mathcal{N}(y_\ast|\mu_\ast, \xi_\ast)
\]

$\mu_\ast, \xi_\ast$ can be obtained from:

\[
\mu_T = K(X_T, X)[K(X, X) + \sigma_n^2 I]^{-1}y
\]

\[
\xi_T^2 = K(X_T, X_T) - K(X_T, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, X_T) + \sigma_n^2 I
\]

We formed the following Kernel to model the inter-pipe relationship between pipes, $i$ and $j$. The GP priors with this kernel expect to see functions which vary smoothly across many length scales.

\[
k(x_i, x_j) = \eta^2 + \frac{\sum_{m=1}^{m=n}[x_i(f_m) - x_j(f_m)]^2 + [x_i(s) - x_j(s)]^2}{2\alpha l^2/\eta^2}
\]

Here, $\eta$ determines the average distance of the function away from its mean, the length-scale $l$ determines the length of the ‘wiggles’ and $\alpha$ determines the relative weighting of large-scale and small-scale variations. $x_i(f_q)$ denotes the $q^{th}$ feature of
the pipe \( i \), and \( x(s) \) represents the survival probability calculated in the previous step using Weibull method. Intuitively, if two pipes, \( i \) and \( j \) are similar, then the survival probabilities of those pipes (also their respective functions, \( f(x_i) \) and \( f(x_j) \)) should also be similar, which explains why the function generated by a GPR is smooth. Our assumption of this similarity or smoothness is encoded by the kernel function \( k(x_i, x_j) \).

Algorithm 1. Obtaining extrapolated survival estimates

1: Define features
   \[ X_{\text{train}} = \{x_1, x_2, \ldots, x_t, \ldots, x_B\} \]
   \[ X_{\text{test}} = \{x_{n+1}, x_{n+2}, \ldots, x_{n+m}\} \]
   Ages = \( \{t_1, t_2, \ldots, t_i\} \) where \( t_i = x_{i, \text{age}} \)
   \[ y = \{y_1, y_2, \ldots, y_t\} \] where \( y_t = 1 \) denotes a failure

2: Define equations for the Weibull model
   Probability density: \( f_{\alpha, \beta}(t) = \frac{\beta}{\alpha} \left( \frac{t}{\alpha} \right)^{\beta - 1} \) \( e^{-\left( \frac{t}{\alpha} \right)^\beta} \)
   Cumulative density:
   \[ L(x_1, \ldots, x_n, \alpha, \beta) = \prod_{i=1}^n f_{\alpha, \beta}(t_i) \]
   Log likelihood:
   \[ l(x_1, \ldots, x_n, \alpha, \beta) = \sum_{i=1}^n \log (f_{\alpha, \beta}(t_i)) \]

3: Find maximum likelihood estimators
   Numerically solve for \((\alpha, \beta) = (\hat{\alpha}, \hat{\beta})\) where,
   \[ l(x_1, x_2, \ldots, x_n, \alpha, \beta) = \sup \{\sum_{i=1}^n \log [f_{\alpha, \beta}(t_i)]\} \]
   \[ \frac{\partial}{\partial \alpha} l(x_1, \ldots, x_n, \alpha, \beta) = \frac{\partial}{\partial \alpha} \sum_{i=1}^n \log [f_{\alpha, \beta}(t_i)] = 0 \]
   \[ \frac{\partial}{\partial \beta} l(x_1, \ldots, x_n, \alpha, \beta) = \frac{\partial}{\partial \beta} \sum_{i=1}^n \log [f_{\alpha, \beta}(t_i)] = 0 \]

4: Populate new feature in training and testing data
   for \( t \) in range \([\min(X_{\text{train}}, \text{age}) : \max(X_{\text{test}}, \text{age})]\) do
   for all \( x_i \) where \( i \in [1, n+m] \) and \( x_{i, \text{age}} = t \) do
     Set, \( x_i, \text{survival prob} = \hat{S}(t) = f_{\alpha, \beta}(t) \)

Table 1. Data statistics for each laid year group

| LY | TrFC | TrFR | TeFC | TeFR |
|----|------|------|------|------|
| 1970 | 56   | 38.81| 42   | 40.75|
| 1971 | 66   | 49.99| 44   | 46.66|
| 1972 | 123  | 38.39| 84   | 36.71|
| 1973 | 127  | 56.47| 90   | 56.03|
| 1974 | 54   | 23.16| 60   | 36.03|
| 1975 | 163  | 56.14| 96   | 46.29|
| 1976 | 155  | 52.00| 119  | 55.89|
| 1977 | 58   | 27.42| 43   | 28.46|
| 1978 | 108  | 36.55| 87   | 41.22|
| 1979 | 53   | 26.25| 47   | 32.59|
| 1980 | 57   | 33.53| 30   | 24.7 |
| 1981 | 46   | 21.73| 40   | 26.46|
| 1982 | 57   | 18.18| 36   | 16.07|
| 1983 | 39   | 14.91| 36   | 19.27|
| 1984 | 106  | 33.85| 88   | 39.34|
| 1985 | 43   | 18.95| 37   | 22.83|
| 1986 | 68   | 19.64| 57   | 23.05|
| 1987 | 35   | 19.09| 32   | 24.43|
| 1988 | 57   | 18.13| 59   | 26.28|
| 1989 | 61   | 18.52| 49   | 20.83|
| 1990 | 53   | 22.64| 38   | 22.72|

LY=laid year, TrFC = training set failure count, TrFR = training set failure rate, TeFC = testing set failure count, TeFR = testing set failure rate.

3 Experimental Setup

3.1 Data Description and Preparation

The studied dataset from the water utility is comprised of water main data and historical work orders (failure records). There are 2800 work orders (failure records) collected from 2005 to 2016. We use the failure records from 2005 to 2011 as the training set, and the samples from year 2012–2016 are used as the testing set. The dataset is divided into 20 groups of water mains which were laid in the years from 1970 to 1990, covering 1888 km of pipeline network. Data statistics are presented in Table 1.

We compare our proposed semiparametric method against the following commonly used machine learning techniques: Gradient Boosting Regression (GBR) [14], Weibull [21], Support Vector Regression (SVR) [18], Poisson regression [27] and GPR [12].
3.2 Results and Discussion

**Estimation of Survival Probabilities:** In the result generation phase, we first decomposed two partitions from the total available dataset as the training and test datasets for the GPR. The training dataset was obtained to comprise the entries for pipes laid from 1970–1990 and observed from 2005–2011, where as the testing dataset captured the entries for the same laid years, observed from 2012–2016.

It could be observed that the training dataset comprised of entries for all pipe ages ranging from 15 to 41 and the testing dataset spanned the age range 22 to 46, as depicted by Fig. 3 (a). We then estimated the survival probabilities for pipes with ages 15 to 41 using the Kaplan-Meier survival estimate presented in Eq. (1). We used the full training dataset for this purpose, assuming that each failure could be treated as a first time failure (which allows to compute the age at failure starting from the actual laid date for the pipe) and obtained the Kaplan-Meier survival curve depicted in Fig. 3 (b). We then considered parametric survival models to extrapolate the survival probabilities beyond the maximum observed age in the inputs (training data), for the purpose of populating the survival probability feature in the testing data. It was evident that Weibull survival probability given by Eq. (2) closely aligns with the Kaplan-Meier survival estimate for the training age span, as depicted by Fig. 3 (b). Thus, we obtained the extrapolated Weibull survival probability estimates for the full age span (15 to 46) using the same assumptions, and only the training data as inputs. These results were used to populate the survival probability feature of both training and test datasets.

![Fig. 3.](image_url)

(a) Age histograms for training and testing data, (b) Survival probability estimates of pipes at different ages computed using Kaplan-Meier fitter (with training data as inputs) and Weibull fitter (extrapolated with training data as inputs), (c) Suburb level model verification. (d)-(e) Long pipeline failure rates have been calculated from year 2012–2016. The shaded area represents the uncertainty associated with each prediction. The model was trained from year 2005–2011.
Long-Term Failure Likelihood for Each Pipe: In pipeline failure forecasting, we are interested more in the predictive accuracy of the model and less about the parameter of the function. Therefore, instead of inferring the parameters to get the latent function of interest (as in parametric methods), here we infer the function \( f(x) \) given in Eq. (3) directly, as if the function \( f(x) \) represents the ‘parameters’ of our model. We have inferred the function \( f(x) \) from the data through the Bayesian inference to obtain a probability distribution for each pipeline failure. In order to quantify the prediction performance of our model, we extensively studied the AUC curves generated by each machine learning technique for each testing year from 2012–2016. The results are populated in the Table 2. The highest AUC is reported for the proposed semiparametric model, which shows that the proposed method outperforms all the other techniques for all the 5 testing years. In general, the proposed GPR technique tends to maintain the same accuracy as the predictions are made further into the future.

We performed further analysis of predicted results using our model on the suburb level. For this study, the suburbs within a state are first sorted, based on the cumulative probability of all the pipes within the suburb for a given prediction year. This forms a suburb based dataset and the general procedure is then followed to obtain the suburb based ROC curve. In the suburb based analysis, we notice that by the aggregation of pipe failure probabilities predicted by proposed model across suburbs, we were able to predict a significant proportion of pipe failures. In Fig. 3 (c), we clearly see that when we inspect top 10 risky suburbs, 60%–70% of the pipe failures were recovered. Our analysis of critical suburbs also reveals quite interesting facts regarding pipe failures and their distribution across suburbs. As Fig. 3 (c) demonstrates, the proposed model performs well in precisely recalling suburbs with greater-than-average numbers of failures in all 5 years into the future. These results suggest that in a given year, the pipe failures are clustered in a few vulnerable suburbs. Prediction uncertainty for the long-term failure prediction has been obtained and illustrated with the mean prediction and actual failure rates as shown in Fig. 3 (d)-(e) for AC (Asbestos Cement), PVC (Plasticized Polyvinyl Chloride) separately. These long-term prediction curves show that actual failure rates align with the results generated by the prediction model and lie within the uncertainty interval. The mean failure prediction is ideal for modelling future behaviour of pipeline network benchmarking performance indices such as unplanned water interruption and water main breaks. In addition to the mean prediction, all the water utilities require uncertainty interval of the prediction. This is required to evaluate the impact and cost of more targeted water network levels of service to inform both short and long-term renewals budgets.

The ensuing results suggest that the semiparametric modelling approach has the potential to more accurately identify high risk pipes over the existing parametric or nonparametric models that are currently available in the water industry. In addition, we are able to obtain the uncertainty associated in the long-term forecasting. This will allow us to target our condition assessment activity so that the pipes in poor-condition are found with the same level of condition assessment activity. It could also be possible to avoid replacement of those pipes that still have a safe level of remaining service life, providing more effective use of our assets.
Table 2. AUC for each method from testing year 2012–2016

| Testing year | GBR model | SVR model | Weibull model | Poisson model | GPR model | Our model |
|--------------|-----------|-----------|---------------|---------------|-----------|-----------|
| 2012         | 0.68      | 0.51      | 0.50          | 0.59          | 0.67      | 0.70      |
| 2013         | 0.64      | 0.58      | 0.55          | 0.60          | 0.70      | 0.73      |
| 2014         | 0.63      | 0.51      | 0.49          | 0.60          | 0.62      | 0.67      |
| 2015         | 0.61      | 0.53      | 0.47          | 0.58          | 0.64      | 0.65      |
| 2016         | 0.60      | 0.53      | 0.53          | 0.60          | 0.63      | 0.67      |

We would also like to highlight that our water pipeline failure prediction models have been deployed in three states across Australia, where we have used numerous other features such as soil data, ground level data, weather data and various pressure data.

4 Conclusion

Pipeline maintenance and renewal programs balance level of service requirements and the need to minimize cost to customers. We have identified, tailoring Artificial Intelligence (AI) techniques to model long-term pipeline failure forecasting provides accurate insights into water main networks. This will essentially assist water authorities to carry-out proactive pipeline maintenance much effectively. We have presented a thorough survey of the landscape of semiparametric survival analysis as it pertains to predictions of survival rates and correspondingly decease rates of assets. We have used data from the water management authorities to validate the survival analysis technique we propose, to compare against other popular machine learning techniques that have been proven effective the in similar applications. We performed a thorough analysis of the performance of the techniques in making predictions over multiple years. The results show that the semiparametric Gaussian process has consistently shown to outperform the other techniques.

This is one of our ongoing predictive analytics projects, which is being carried-out in close collaboration with Australian water authorities. This work highlights Australia’s efforts in using AI to uplift the future living conditions of the society. Ultimately, we believe this work, at the intersection of Machine Learning and Asset Management, will lead to more effective and proactive infrastructure maintenance in the water industry.

References

1. 7News: Bus towed after being stuck in Sinkhole at Caringbah after water main bursts (2019). https://7news.com.au/news/traffic/bus-stuck-in-sinkhole-at-caringbah-after-water-main-bursts-c-434845
2. Asnaashari, A., McBean, E., Shahrour, I., Gharabaghi, B.: Prediction of watermain failure frequencies using multiple and poisson regression. Water Sci. Technol. Water Supply 9(1), 9–19 (2009)
3. BITRE: Australian infrastructure statistics report. In: Australian infrastructure statistics, Australian Government, Department of Infrastructure, Transport, Cities and Regional Development (2014)
4. Brahim-Belhouari, S., Bermak, A.: Gaussian process for nonstationary time series prediction. Comput. Stat. Data Anal. 47(4), 705–712 (2004)
5. Collett, D.: Modelling Survival Data in Medical Research. Chapman and Hall/CRC, Boca Raton (2015)
6. Cronin, D.S., Pick, R.J.: Prediction of the failure pressure for complex corrosion defects. Int. J. Press. Vessels Pip. 79(4), 279–287 (2002)
7. De Iorio, M., Johnson, W.O., Müller, P., Rosner, G.L.: Bayesian nonparametric nonproportional hazards survival modeling. Biometrics 65(3), 762–771 (2009)
8. Fernández, T., Rivera, N., Teh, Y.: Gaussian processes for survival analysis. In: Advances in Neural Information Processing Systems, pp. 5021–5029 (2016)
9. Goel, M.K., Khamma, P., Kishore, J.: Understanding survival analysis: Kaplan-Meier estimate. Int. J. Ayurveda Res. 1(4), 274 (2010)
10. Gould, S., Boulaire, F., Burn, S., Zhao, X.L., Kodikara, J.: Seasonal factors influencing the failure of buried water reticulation pipes. Water Sci. Technol. 63(11), 2692–2699 (2011)
11. Hjort, N., Holmes, C., Müller, P., Walker, S.: Bayesian Nonparametrics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2010)
12. Hong, S., Zhou, Z., Lu, C., Wang, B., Zhao, T.: Bearing remaining life prediction using gaussian process regression with composite kernel functions. J. VibroEng. 17(2), 695–704 (2015)
13. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. 53(282), 457–481 (1958)
14. Kumar, A., et al.: Using machine learning to assess the risk of and prevent water main breaks. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 472–480 (2018)
15. Langford, J., Li, L., Zhang, T.: Sparse online learning via truncated gradient. J. Mach. Learn. Res. 10, 777–801 (2009)
16. Liang, B., Weeraddana, D.: Pipeline failure data analytics and prediction. In: OzWater, pp. 25–33. Australian Water Association (2018)
17. Liu, X.: Survival Analysis: Models and Applications. Wiley, Hoboken (2012)
18. Mashford, J., De Silva, D., Marney, D., Burn, S.: An approach to leak detection in pipe networks using analysis of monitored pressure values by support vector machine. In: 2009 Third International Conference on Network and System Security, pp. 534–539. IEEE (2009)
19. Rasmussen, C.E.: Gaussian processes in machine learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (eds.) ML -2003. LNCS (LNAI), vol. 3176, pp. 63–71. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-28650-9_4
20. Rich, J.T., Neely, J.G., Paniello, R.C., Voelker, C.C., Nussenbaum, B., Wang, E.W.: A practical guide to understanding Kaplan-Meier curves. Otolaryngol.-Head Neck Surg. 143(3), 331–336 (2010)
21. Sahu, S.K., Dey, D.K., Aslanidou, H., Sinha, D.: A weibull regression model with gamma frailties for multivariate survival data. Lifetime Data Anal. 3(2), 123–137 (1997)
22. Scholz, F.: Inference for the Weibull distribution. In: Stat 498B Industrial Statistics, pp. 6–10 (2008)
23. Selvin, S.: Survival Analysis for Epidemiologic and Medical Research. Cambridge University Press, Cambridge (2008)
24. Shamir, U., Howard, C.D.: An analytic approach to scheduling pipe replacement. J. Am. Water Works Assoc. 71(5), 248–258 (1979)
25. Topa, H., Honkela, A.: Gaussian process modelling of multiple short time series. arXiv preprint arXiv:1210.2503 (2012)
26. Weeraddana, D., et al.: Utilizing machine learning to prevent water main breaks by understanding pipeline failure drivers. In: OzWater, pp. 25–33. Australian Water Association (2019)

27. Zhang, X., Zhong, S., Wu, Z., Li, Y.: Seasonal prediction of the typhoon genesis frequency over the western north pacific with a poisson regression model. Clim. Dyn. 51(11–12), 4585–4600 (2018)