Towards Effective Prioritizing Water Pipe Replacement and Rehabilitation*

Junchi Yan[†§], Yu Wang[§], Ke Zhou[‡], Jin Huang[§],
Chunhua Tian[§], Hongyuan Zha[‡], Weishan Dong[§]
[†] Shanghai Jiao Tong University
[§] IBM Research - China
[‡] Georgia Institute of Technology
{yanjc,yuwangbj,huangjsh,chtian,dongweis}@cn.ibm.com, {kzhou,zha}@cc.gatech.edu

Abstract

Water pipe failures can not only have a great impact on people's daily life but also cause significant waste of water which is an essential and precious resource to human beings. As a result, preventative maintenance for water pipes, particularly in urban-scale networks, is of great importance for a sustainable society. To achieve effective replacement and rehabilitation, failure prediction aims to proactively find those 'most-likely-to-fail' pipes becomes vital and has been attracting more attention from both academia and industry, especially from the civil engineering field. This paper presents an already-deployed industrial computational system for pipe failure prediction. As an alternative to risk matrix methods often depending on ad-hoc domain heuristics, learning based methods are adopted using the attributes with respect to physical, environmental, operational conditions and etc. Further challenge arises in practice when lacking of profile attributes. A dive into the failure records shows that the failure event sequences typically exhibit temporal clustering patterns, which motivates us to use the stochastic process to tackle the failure prediction task. Specifically, the failure sequence is formulated as a self-exciting stochastic process which is, to our best knowledge, a novel formulation for pipe failure prediction. And we show that it outperforms a baseline assuming the failure risk grows linearly with aging. Broad new problems and research points for the machine learning community are also introduced for future work.

1 Introduction

Clean water, distributed through a complex and growing network of water pipes, is essential for people's daily life. In fact, the large-scale urban water pipe networks are in fast growth to meet the increasing demand arising from the fast developing urban areas. However, the structural deterioration has presented great challenges to worldwide water utilities, posing a critical threat to not only the daily life but also a sustainable society since water is an essential and precious resource of human beings. For example, it is estimated that more than $32 billion cubic meters of treated water physically leak annually through distributed network worldwide [Kingdom *et al.*, 2006]. And the corresponding total annual cost is around $ 14 billion [Kingdom *et al.*, 2006]. As reported in [Carter and Rush, 2012], the New York City has spent $54.6 million to manage its break pipes from 1997 to 2011. As few municipalities can afford to systematically inspect all of their pipes in their water network due to scale of the water network and the difficulty of inspections — most pipes are laid underground, the task of proactively pinpointing those 'most-likely-to-fail' pipes, which enables cost-effective replacement and rehabilitant, becomes an important problem for a sustainable society.

Formally, given 1) a prediction time point or window, 2) the pipe-specific failure sequence and 3) the associated attributes, one aims to predict the failure likelihood on a single pipe-wise level. It leads to a binary supervised learning problem if the failed pipes are regarded as positive samples and non-failure ones negative; and if one further considers the fact that negative samples are censored, survival analysis models are assumed to better capture this subtle difference of the censored data from other type data during learning. This paper further formulates the problem into a stochastic process.

The pipe failure problem has been an issue of concerning for municipal engineers since the early studies [Arnold, 1960; Clark, 1960; Niemeyer, 1960; Remus, 1960]. Traditionally, subject matter experts (SME) devise certain business rules to decide which pipes are risky and should receive top fixing priority. While such an intensive domain knowledge driven methodology usually involves ad-hoc rule definition and is not conveniently scalable when new attributes arrive. Existing pipe integrity management methodologies mostly focus on oil and gas, and many risk factor framework have been proposed in these areas such as the report from the Pipeline Research Council International and the manual by [Muhlbauer, 2004], where 9 categories are classified in the former and 4 in the latter. Compared with oil/gas transmission pipelines, urban water distribution network usually has less data available in fine-granularity. A practical way is to categorize the factors into physical indicators, load, corrosion, weather and historical failure record.

*The work is partially supported by NSF IIS-1116886, NSF IIS-1049694, NSFC 61129001/F010403 and the 111 Project (B07022).

Another research line adopts the methodologies from statistic field to analyze the failure problem. One can refer to [Kleiner and Rajani, 2001] for an early review. Many early work focus on descriptive analysis towards pipe failures. [Shamir and Howard, 1979] calculate the average number of failures on a unit year and unit pipe length. And the spatial and temporal patterns of water distribution pipe failure in the City of Winnipeg are examined in [Goulter and Kazem, 1998]. On the other hand, predictive modeling is also investigated. [Kleiner and Rajani, 1999] addresses the problem of forecasting the aggregated number of pipe failures for the network, which is key to beforehand planing. [Pelletier *et al.*, 2003] performs survival analysis to predict the evolution of the annual number of pipe breaks and to estimate the impact of different replacement scenarios in real case studies. [Tian *et al.*, 2011] uses Cox survival analysis as a pilot study for pipe failure prediction. In a recent work, a rank boosting algorithm is adopted by [Wang *et al.*, 2013] to rank the pipe break risks. However, to our best knowledge, no previous work has presented an already-deployed system and being continuously generating the environmental and business benefits to the concerned metropolitan. Moreover, from the machine learning perspective, we originally propose to formulate the pipe failure events into a self-exciting stochastic process model.

The main contributions of this paper are in five folds:

1) We implement a web based 'environmental computing' system addressing the real-world large scale urban pipe network maintenance supporting system, and it has been deployed in a real commercial environment, generating continuous impacts and benefits for a better sustainable society;

2) We evaluate the system's prediction performances *quantitatively* against the traditional subject matter expert driven risk matrix method on the real-world large scale dataset. And our study further discovers some specific limitations of the risk matrix method, and verify the viability of data-driven approach towards pipe analysis;

3) Attributes closely relevant to failure are identified by computational modeling, leading the agencies to better understand the root cause of pipe failure, which in turn improves and streamlines the pipe construction and maintenance;

4) Perhaps more interestingly, the temporal clustering patterns of failure time stamps are unveiled and mathematically modeled by a self-exciting stochastic model, which outperforms a linear aging baseline. And it can work flexibly in the absence of enough attribute information; while this is often the case in practice. We are not aware of any prior work formulating the self-exciting point process for pipe failure prediction problem, nor even for asset management literatures;

5) Our novel formulation for the pipe failure prediction identifies the new problems for the machine learning community beyond conventional classifiers, leaving a broad room to boost the performance based on the this seminal work.

2 System overview

In this section, we first describe the system's main use cases, and then the studied dataset as well as the modules that comprise the J2EE web system are briefed. There are mainly three scenarios for the system, as summarized: 1) Help operator find the 'likely-to-fail' pipes; 2) Help decision maker allocate the maintenance budget; 3) Help agency reveal important factors leading to failure.

2.1 Dataset description

The used dataset is a real-world urban network from one costal metropolis, which consists of over 600,000 pipes in fresh water system and nearly 100,000 for salt. Most of the pipes are laid after 1950 and the average age is over ten years. For most samples, there is a number of profile attributes associated with each pipe. Basically, the attributes can be categorized into pipe physical indicators, load, corrosion and weather etc. In addition, as time goes by, the pipe failure event will be processed and recorded by the agency on a routinely basis. As a result, each pipe is assumed to be associated with profiling attributes and the failure records.

However, still a portion of pipes whose attributes are recorded incompletely, or not well documented, resulting in only knowing their failure records and the year when they are laid. This fact calls for new formulation beyond the off-the-shelf supervised learning methods.

2.2 Functional modules

The presented system is mainly comprised of four modules: a) data cleansing; b) pattern display; c) model training; d) risk prediction. It is also equipped with a GIS component to highlight the high-risky pipes on the map, and allows the user to directly manipulate the pipes from this interface instead of using a table-style list. An overview is shown in Fig.1.

Data preprocessing: The raw data are from two datasets: a) 'profile' dataset: pipe-specific profile attribute information, such as material type, length, completion year, location etc; and b) 'working order' dataset: historical pipe failure records, location and the completion year. In order to repopulate the failure information, the correspondence matching between the two datasets is performed. Specifically, the 'pipe id' serves as the foreign key for joining, and in the cases when the 'pipe id' is missing in one dataset, location is used to find the most likely pipe in the other.

Pattern display: To give a visual perception, the cleansed data is plot regarding with various aspects including water type (fresh or salt), district, material, diameter range, and length range etc. Different visualization chart is adopted such as histogram, bar, and bubble to plot certain distributions.

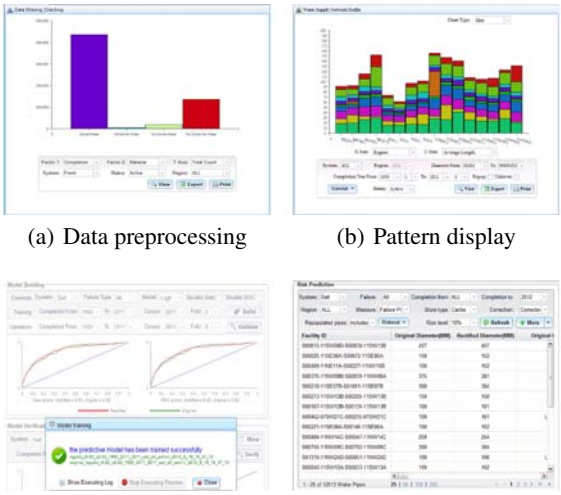
Model training: Learning based classifier and stochastic process model, together with risk matrix approach are evaluated and integrated in the system for pipe failure prediction.

Risk prediction: Prediction results are displayed and can be exported in different formats including shapefile, office xlsx, and csv file for further processing and supports the decision making. Meanwhile, pipes scored with different risk levels can be highlighted as the hot spots in the map.

3 Methodology

3.1 Attribute selection

Examination of the attributes associated with the pipes will help to identify those factors that appear most susceptible to



(a) Data preprocessing (b) Pattern display

(c) Model training (d) Risk prediction

Figure 1: System modules demonstration.

failure. Once these attributes have been identified, hopefully it will lead to a means of reducing the failure rate in the future by improving the key conditions if feasible. Through a model variable correlation analysis and feature importance qualifying, we summarize the final selected attributes feeding to the computational model in Table 1. One should note that the importance ranking is based on the available pipe profile information under a supervised learning fashion, which may not be perfectly precise because some of the raw data are difficult to collect or estimate especially for the external factors like rainfall, highway impact etc. Some other relevant factors are excluded for certain reasons: e.g. 'bury depth' is excluded due to its strong correlation with 'zoneImp' and 'joint number' is strongly correlated to 'length'. 'Pressure' is also excluded due to unacceptable amount of missing values.

Table 1: Pipe-specific attributes considered in the system.

	factor	description
1	Material	material type, such as PE, GI, GIL etc.
2	Length	pipe length
3	Diameter	pipe diameter
4	Age	time since the completion date
5	ZoneImp	supply zone impact, categorical
6	Rainfall	nearby average rainfall volume
7	WCNo	water crossing number
8	HighImp	nearby highway impact, categorical
9	Excavate	road excavation impact, numerical

3.2 Binary classifiers using attribute information

The system is equipped with several predictive models and also allows easy extension for integrating more new models. Cox Model [Cox, 1972] (together with a new survival analysis algorithm: Multi-Task Logistic Regression (MTLR) [Yu and Baracos, 2011]), Artificial Neural Network (ANN), Logistic Regression (LGR) and Chaid Tree are available in the system which explore the massive labeled training data. However, there are still some limitations: From the practical perspective, one problem is for some pipes, especially for those constructed before 1970, of which the associated attribute information is largely incomplete, inconsistent, or unreliable. And collecting the relevant attributes information is often costive and difficult, such as estimating the water turbidity, the rainfall, and the soil type etc.; the other subtle but worth-noting issue is one cannot guarantee the current system has identified exhaustively all factors with respect to pipe failure, and sometimes training based on incomplete covariants may be misleading. From the theoretical perspective, conventional binary classifiers like ANN, Logistic Regression or Chaid tree cannot naturally explore and model the particular property for 'censored' samples compared with the Cox model and the MTLR models [Yu and Baracos, 2011] etc.

As the prediction score obtained from classifiers such as ANN, Chaid Tree is not a posterior likelihood, it makes the risk measurements from fresh and salt systems are incomparable since in many cases the system user customizes and uses different models for different systems. To address this issue, the parametric Sigmoid model [Platt, 1999; Lin *et al.*, 2007] is used in the system to calibrate the failure likelihood on an equal footing for cross-system risk ranking. Another advantage is knowing the likelihood naturally leads to the obtain of the failure number expectation for the next year. And it is informative for the decision maker to better determine and allocate the budget in advance.

3.3 Modeling temporally clustered failures using one-dimensional self-exciting process

Apart from the attribute-complete samples, there are still a few, whose attributes are missing in both fresh and salt systems. For these pipes, the above classifiers cannot be trained as no attribute data is available. On the other hand, it is appealing to suppose *pipes are prone to leak repeatedly shortly after a recent failure event (even the previous failure has been fixed up)* either inherently due to the damage brought by the previous failures, or externally lasting external impacts, such as road excavation. In another word, we conjecture the behind mechanism is not only due to events triggering the next in the causal sense, but also the clustering reflects the correlation of event occurrence due to unobserved variables such as geological factors and urban activities. Our inspection to the failure records verifies this idea and some examples are shown in Fig.2. In addition, such temporal clustering patterns are also observed and discussed in the previous literature [Goulter and Kazem, 1998], in the context of civil engineering study. And their empirical study further shows the temporal decay rate obeys an exponential distribution since the occurrence time instance. As a result, we are motivated to quantitatively model the failure behavior from the temporal perspective, by leveraging the historical event samples.

A series of investigations to pipe failure in New York and Philadelphia [O'Day, 1982; Ciottoni, 1983] conclude that age is not an influential factor, and our empirical study also suggests a simple assumption: the background rate of failure rate is relatively time-stationary as shown in Fig.5 for failure age

distribution. Based on these assumptions, we formulate the time-stationary and temporal-clustered failures into a computationally efficient (with some mathematical approximation) self-exciting stochastic process.

Formally, given a sequence of failure events with time stamp $t_1, t_2, ..., t_n$, we investigate a specific class of point processes termed as Hawkes Process [Hawkes, 1971], commonly used in earthquake analysis [Ogata, 1988; 1998], and recently used in the criminological literature [Johnson, 2008], which is modeled by a conditional intensity function:

$$\lambda = \mu + \alpha \sum_{t_k<t} g(t - t_k).$$

In line with the observation made by [Goulter and Kazem, 1998], an exponential kernel is used:

$$\lambda = \mu + \alpha w \sum_{t_k<t} e^{-w(t-t_k)}.$$

where μ is the background rate controlling the intrinsic failure event influence time window, and α is the amplitude of the influence, while w^{-1} bears the physical implication for the average waiting time until a new failure comes. And according to [Rubin, 1972], the likelihood is given by:

$$L = \prod_{i=1}^{n} \lambda(t_i) exp\{-\int_0^T \lambda(s)ds\}.$$

To solve this problem, maximum likelihood estimation is performed in an Expectation-Maximization fashion to infer the model's parameters. While [Lewis *et al.*, 2011] address the kernel form self-exciting terms specifically, here we begin the derivation in the context of general Hawkes process where $g(x)$ is not specified. Let $t_0 = 0$ and $t_{n+1} = T$, first one can obtain the log likelihood function in the form:

$$\begin{aligned}
&\log L(\mu, \alpha | t_1, t_2, ..., t_n) \\
&= \sum_{i=1}^{n} \log \lambda(t_i) - \int_0^T \lambda(t)dt \\
&= \sum_{i=1}^{n} \log \lambda(t_i) - \int_0^T (\mu + \alpha \sum_{t_j<t} g(t - t_j)dt) \\
&= \sum_{i=1}^{n} \log \lambda(t_i) - \left\{ \mu T + \sum_{i=0}^{n} \int_{t_i}^{t_{i+1}} \alpha \sum_{t_j<t} g(t - t_j)dt) \right\} \\
&= \sum_{i=1}^{n} \log \lambda(t_i) - \left\{ \mu T + \sum_{i=0}^{n} \sum_{j=1}^{i} \alpha(G(t_{i+1} - t_j) - G(t_i - t_j)) \right\} \\
&= \sum_{i=1}^{n} \log \lambda(t_i) - \left\{ \mu T + \sum_{j=1}^{n} \sum_{i=j}^{n} \alpha(G(t_{i+1} - t_j) - G(t_i - t_j)) \right\} \\
&= \sum_{i=1}^{n} \log \lambda(t_i) - \left\{ \mu T + \sum_{j=1}^{n} \alpha(G(T - t_j) - G(0)) \right\}.
\end{aligned}$$

Observing log function is concave and $G(0)$=0, one can further derive the lower bound for the objective function:

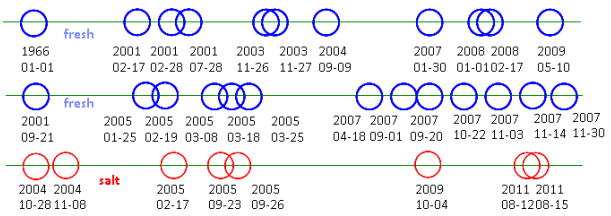


Figure 2: Illustration of the temporal cluster pattern of failures (the dates correspond to the failure time stamp; top two lines are pipes from fresh system, the bottom from salt).

$$\begin{aligned}
&\log L(\mu, \alpha | t_1, t_2, ..., t_n) \\
&= \sum_{i=1}^{n} \log \left\{ \mu + \sum_{j=1}^{i-1} g(t_i - t_j) \right\} - \left\{ \mu T + \sum_{j=1}^{n} \alpha G(T - t_j) \right\} \\
&\geq \sum_{i=1}^{n} \left\{ p_{ii} log \frac{\mu}{p_{ii}} + \sum_{j=1}^{i-1} p_{ij} \log \frac{g(t_i - t_j)}{p_{ij}} \right\} - \mu T - \sum_{j=1}^{n} \alpha G(T - t_j).
\end{aligned}$$

where in the E-step:

$$\begin{aligned}
p_{ij}^{k+1} &= \frac{\alpha^k g(t_i - t_j)}{\mu^k + \sum_{j=1}^{i-1} \alpha g(t_i - t_j)}, j = 1, ..., i-1. \\
p_{ii}^{k+1} &= \frac{\mu^k}{\mu^k + \sum_{j=1}^{i-1} \alpha g(t_i - t_j)}.
\end{aligned}$$

Note that the above choices of p_{ij} and p_{ii} make the above lower bound tight when $\mu = \mu^k$ and $\alpha = \alpha^k$, which ensures that log-likelihood increases monotonically during the iterations. And in the M-step, the partial derivatives in terms of the objective function are given as:

$$\begin{aligned}
\frac{\partial L}{\partial \mu} &= \sum_{i=1}^{n} \frac{p_{ii}}{\mu} - T = 0. \\
\frac{\partial L}{\partial \alpha} &= \sum_{i>j} \frac{p_{ij}}{\alpha} - \sum_{j=1}^{n} G(T - t_j) = 0.
\end{aligned}$$

Thus the updating μ and α in the kth iteration are:

$$\mu^{k+1} = \frac{1}{N} \sum_{i=1}^{n} p_{ii}^k, \quad \alpha^{k+1} = \frac{\sum_{i>j} p_{ij}^k}{\sum_{j=1}^{n} G(T - t_j)}.$$

Particularly, given an exponential kernel as $g(t - t_j) = we^{-w(t-t_j)}$, we use the approximation $e^{-w(T-t_i)} \approx 0$ when $wT \gg 1$ as suggested in [Lewis *et al.*, 2011] which shows the scale parameter w and α can be obtained by:

$$w^{k+1} = \frac{\sum_{i>j} p_{ij}^k}{\sum_{i>j} (t_i - t_j) p_{ij}^k}, \quad \alpha^{k+1} = \frac{1}{n} \sum_{i>j} p_{ij}^k$$

4 Experimental results

4.1 Results on attribute-complete data

The area under the Receiver Operating Characteristic (ROC) curve [Fawcett, 2006], aka. the Area Under Curve (AUC), is

used to assess the predictive models in this paper. ROC has become a standard metric in the area[Provost *et al.*, 1997]. We evaluate the performance of all testing methods regardng with overall performance as the prediction year varies from 2001 to 2011. Table 2 gives a rough statistics about the well recorded failure counts along the years. Also, the risk matrix method is quantitatively compared. Risk matrix method is popular for pipe integrity management. The main idea is to assign score to each pipe by considering the impact of the risk factors: 1) material; 2) age; 3) failure record; and 4) surrounding condition. For the samples with comeplete attributes, we evaluate three classifiers and two survival analysis models because we are interested in verifying whether the survial analysis models that are able to inherently handle censored data would outperform the other type data based approaches. For all tests, cross-validation is performed and the final AUC is obtained by averaging the hold-out testing results. In particular, for samples of fresh system, the cross-fold is set to ten; while for salt, cross-fold is five due to its smaller size.

Table 2: Recorded failure/total number of pipes (proportion).

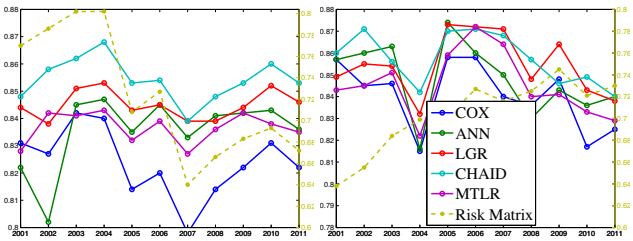
Year	Fresh	Salt
2001	1081/146354(0.74)	489/37707(1.30)
2002	1032/163268(0.63)	501/42079(1.19)
2003	1178/183738(0.64)	589/46025(1.28)
2004	1121/208749(0.54)	526/50757(1.04)
2005	1749/235055(0.74)	626/55991(1.12)
2006	1319/259045(0.51)	486/60742(0.80)
2007	2941/283275(1.04)	675/64314(1.05)
2008	3413/298863(1.14)	704/68320(1.03)
2009	2557/326125(0.78)	721/74812(0.96)
2010	2537/370247(0.69)	655/82581(0.79)
2011	2687/405416(0.66)	696/87279(0.80)

Table 3: Mean AUC and percentage of true failures covered by the top 2% of the prediction ranking list. Prediction year = 2011, hold-out fold = 10 for fresh system, and 5 for salt.

	Chiad	ANN	LGR	COX	MTLR
AUC-fresh	0.853	0.836	0.846	0.822	0.835
Cover-fresh	23.1%	22.8%	22.7%	21.8%	22.0%
AUC-salt	0.840	0.840	0.838	0.825	0.829
Cover-salt	21.3%	21.3%	21.2%	20.5%	20.8%

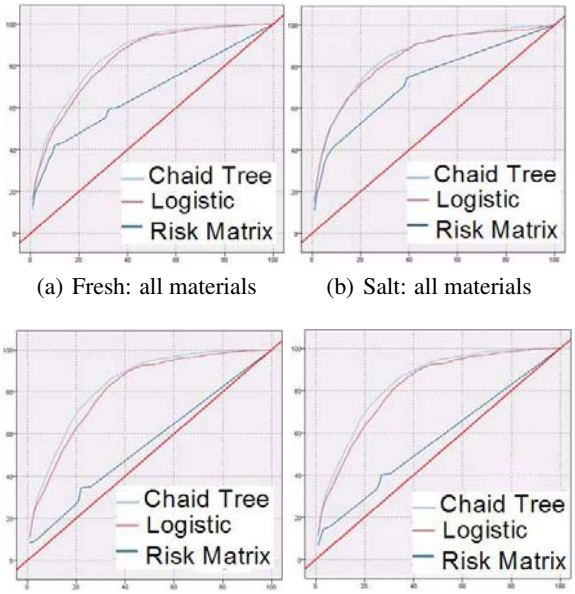
4.2 Results on attributes-missing dataset using Hawkes Process model

We are interested in estimating the self-exciting term in the Hawkes process model thus those with zero failure record are excluded due to they can only influence the unconditional background failure rate μ. Thus basically we will select the pipes that has failed for at least once. As a case study, we collect 676 pipes whose attributes are missing/incomplete or in low quality for fresh system, and 200 for salt. The completion date of these pipes are known[1].



(a) Fresh system (b) Salt system

Figure 3: Model sensitivity test on attribute-complete samples across year 2001 to 2011. The auc measure for risk matrix method is on the right side, others on the left.



(a) Fresh: all materials (b) Salt: all materials

(c) Fresh: exclude GI & GIL (d) Salt: exclude GI & GIL

Figure 4: Evaluation between data-driven and knowledge-based risk matrix methods using attribute-complete samples; prediction year = 2011. The diagonal in red is baseline.

We use a concrete example to illustrate the experiment design: given a pipe with failures at: $t_1, t_2, ..., t_n$, the first n-1 events are used as the training samples, and we have many samples for training the same model; while the last event at t_n is used to build the reference ground truth for model testing. Specifically, the pipe is regarded as a positive testing sample if its last failure so far falls within a certain period after the failure at t_{n-1}, i.e. $t_n \in (t_{n-1}, t_{n-1} + T]$. A practical T is

[1]In fact, as shown in the EM training algorithm derivation for the *time-stationary* Hawkes Process model, the model does not sensitively rely on the observation window T. Thus a rough estimation of T is enough. And for those pipes of which the completion date is unknown, the model is also workable as long as the failure records are complete and accurate, while this is often the case because the pipe failure will receive immediate processing and recording.

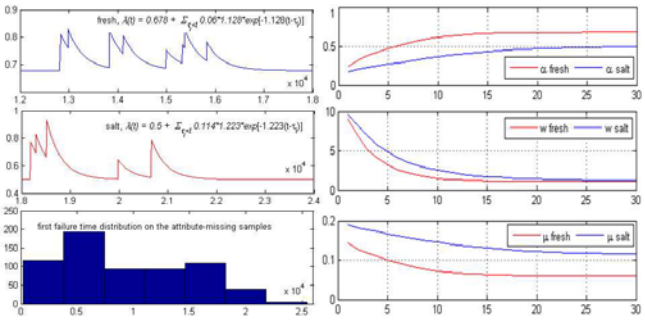


Figure 5: Left: failure rate fitting using the learned model on the first and the third pipe as exemplified in Fig.2 and the pipe age distribution of the first failure event on the attribute-missing samples; Right: Hawkes model learning convergence behavior. One shall be conscious for the first failure pipe age distribution, the decrease after 18000 days is probably due to the fact observation window is always censored. The average age of the sampled pipes in the plot is 12175 days and the standard deviation is 5706.9 days. It suggests our time-stationary assumption is not violated to a large extent.

set to a half year in our test. The prediction score for each pipe is calculated by the failure rate integral within the period: $\int_{t_{n-1}}^{t_{n-1}+T} \lambda(t)dt$ where the parameters are learned using the training samples. Based on this protocol, we train and test the Hawkes models for fresh and salt systems respectively. The ROC are presented in Table 4. And the iterative model parameter convergence and the fitting examples are plot in Fig.5. Note the linear aging model is compared as a baseline where the failure rate is modeled as $\lambda = Constant \times age$.

Table 4: ROC evaluation on attribute-missing samples.

	fresh system	salt system
Hawkes self-exciting	0.676	0.612
Linear aging baseline	0.487	0.536

4.3 Discussion

Some observations are made as listed in the following:

Risk matrix method is prone to focuses on the most predominant aspects and deteriorate largely on a subtle classification boundary: As observed from Fig.3, all data-driven methods are consistently superior than the empiric-driven risk matrix approach for any given prediction year. And Fig.4 gives a more subtle example: by excluding the samples with material type 'GI' or 'GIL'[2], the performances deteriorates severely with respect to both salt and fresh systems. This is reasonable considering these two types are relatively vulnerable than other types. Once such samples are excluded, the risk matrix method become weaker because the classifying boundary is probably a nonlinear combination of attributes. Our results are in line with [Cox, 2008] discussing about limitations of the risk matrix approach.

Binary classifiers, which do not consider the labels in the training set are from censored observation, achieve comparable performance on the censored data: [Yu and Baracos, 2011] point out that the binary classifier cannot fully capture the characters of the censored data. While in our testing, we found both the classical Cox model and the state-of-the-art MTLR show no significant superiority against the binary classifiers. In our analysis, this may be due to on one hand, the implicit nonlinear feature transformation as performed by ANN, Chaid can boost the performance; on the other hand, the labeling noise or unobserved labels brought by the censored observation can be addressed by the classifiers to some extent. Considering [Yu and Baracos, 2011]'s heavier training overload, in our application, the Chaid tree and Logistic Regression are recommended for their interpretability and efficiency especially when the training data grows fast and more attributes arrive incurring the high-dimensionality issue.

Formulating the failure event sequence into a self-exciting stochastic process model is beneficial against a simple linearly aging baseline or random guess: We also observe that the self-exciting Hawkes Process model achieves an acceptable result, and the temporal cluster grouping is visually demonstrated in our plot. Perhaps more importantly, our method gives a rigorous description about the particular temporal patterns of pipe failures, and we believe this methodology will also apply to many other asset management scenarios.

5 Conclusion and future work

An already-deployed pipe failure prediction system was detailed. We show that binary classifiers are comparable to the 'censored' survival analysis models and the imbalanced dataset is not a big issue from our empirical study. Moreover, we propose another way of looking at the pipe failures by formulating it as a self-exciting stochastic process, where the attributes are not a must as in other models.

This work opens up a new space for the machine learning community to address the pipe failure prediction problem and asset management applications, which is different from conventional learning paradigms and worth future studies: 1) Spatial-temporal modeling is appealing since the pipe failures exhibit somewhat spatial grouping as pointed by [Goulter and Kazem, 1998]; 2) this paper tackles the different types of failure indiscriminately since the failure type information is currently unavailable, while taking the type into account will lead to a *multivariate* point Hawkes process worth further investigation; 3) the possible underlying failure correlation among the nearby pipes can also be modeled by a *multi-dimensional* Hawkes process; 4) it is challenging but interesting to explore if one can repopulate the missing-attributes from the temporal patterns of failures, one relevant work is [Stomakhin *et al.*, 2011]. It is also appealing to study if the graph matching method [Tian et al., 2012] can be applied to model the topological pattern of pipe networks, which is assumed to have the connection with pipe failures.

[2]GI: Galvanized Iron; GIL: Lined Galvanized Iron

References

[Arnold, 1960] G. E. Arnold. Experience with main breaks in four large cities-philadelphia. *Journal of the American Water Works Association*, 53(8):1041–1044., 1960.

[Carter and Rush, 2012] S. Carter and P. V. Rush. Filtration avoidance annual report. *Department of Environmental Protection of New York City, Tech. Rep.*, 2012.

[Ciottoni, 1983] A. S. Ciottoni. Computerized data management in determining causes of water main breaks: The philadelphia case study. In *International Symposium on Urban Hydrology, Hydraulics and Sediment Control*, Lexington, KY, 1983. University of Kentucky.

[Clark, 1960] E. J. Clark. Experience with main breaks in four large cities-new york. *Journal of the American Water Works Association*, 53(8):1045–1048., 1960.

[Cox, 1972] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[Cox, 2008] L.A. Jr. Cox. What's wrong with risk matrices? *Risk Analysis*, 28(2), 2008.

[Fawcett, 2006] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters - Special issue: ROC analysis in pattern recognition*, 27(8):861–874, June 2006.

[Freund *et al.*, 2003] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 2003.

[Goulter and Kazem, 1998] I. C. Goulter and A. Kazem. Spatial and temporal groupings of water main pipe breakage in winnipeg. *Can. J. Civ. Eng.*, 15:91–97, 1998.

[Hawkes, 1971] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 1971.

[Johnson, 2008] S. Johnson. Repeat burglary victimisation: a tale of two theories. *IEEE Trans. Automatic Control*, 4:215–240, 2008.

[Kingdom *et al.*, 2006] B. Kingdom, R. Liemberger, and P. Marin. The challenge of reducing non-revenue water (nrw) in developing countries : how the private sector can help : a look at performance-based service contracting. In *Water Supply and Sanitation Sector Board discussion paper series*, Washington, DC, USA, 2006. World Bank.

[Kleiner and Rajani, 1999] Y. Kleiner and B. Rajani. Using limited data to assess future needs. *Journal of the American Water Works Association*, 91(7):47–61, 1999.

[Kleiner and Rajani, 2001] Y. Kleiner and B. Rajani. Comprehensive review of structural deterioration of water mains: statistical models. *Urban Water*, 3(3), 2001.

[Lewis *et al.*, 2011] Erik Lewis, George Mohler, P. Jeffrey Brantingham, and Andrea Bertozzi. Self-exciting point process models of civilian deaths in iraq. *Security Journal*, 2011.

[Lin *et al.*, 2007] Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C. Weng. A note on platt's probabilistic outputs for support vector machines. *Machine Learning*, 2007.

[Muhlbauer, 2004] W. Kent Muhlbauer. *Pipeline Risk Management Manual (Third Edition)*. Elsevier Inc., 2004.

[Niemeyer, 1960] H. W. Niemeyer. Experience with main breaks in four large cities-indianapolis. *Journal of the American Water Works Association*, 53(8):1051–1058., 1960.

[O'Day, 1982] D. K. O'Day. Organizing and analyzing leak and break data for making main replacement decisions. *Journal of the American Water Works Association*, 74(11):589–596., 1982.

[Ogata, 1988] Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.*, 83(401):9–27, 1988.

[Ogata, 1998] Y. Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50:379–402, 1998.

[Pelletier *et al.*, 2003] G. Pelletier, A. Mailhot, and J.-P. Villeneuve. Modeling water pipe breaksłthree case studies. *Journal of Water Resources Planning and Management*, 129(2):115–123, 2003.

[Platt, 1999] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[Provost *et al.*, 1997] Foster Provost, Tom Fawcett, and Ron Kohavi. The case against accuracy estimation for comparing induction algorithms. In *International Conference on Machine Learning*, 1997.

[Remus, 1960] G. J. Remus. Experience with main breaks in four large cities-detroit. *Journal of the American Water Works Association*, 53(8):1048–1051., 1960.

[Rubin, 1972] I. Rubin. Regular point processes and their detection. *IEEE Trans. on Information Theory*, IT-18(5):547–557, 1972.

[Shamir and Howard, 1979] U. Shamir and C. Howard. An analytical approach to scheduling pipe replacement. *Journal of the American Water Works Association*, 71(5):248–258, 1979.

[Stomakhin *et al.*, 2011] Alexey Stomakhin, Martin B. Short, and Andrea L. Bertozzi. Reconstruction of missing data in social networks based on temporal patterns of interactions. *Inverse Problems*, 27, 2011.

[Tian *et al.*, 2011] Chunhua Tian, Jing Xiao, Jin Huang, and Felipe Albertao. Pipe failure prediction. In *Proceedings of SOLI*, Beijing, China, 2011. IEEE.

[Tian *et al.*, 2012] Y. Tian, J. Yan, H. Zhang, Y. Zhang, X. Yang, and H. Zha. On the convergence of graph matching: Graduated assignment revisited. In *ECCV,* 2012.

[Wang *et al.*, 2013] Rui Wang, Weishan Dong, Yu Wang, Ke Tang, and Xin Yao. Pipe break prediction: A data mining method. In *ICDE*, 2013.

[Yu and Baracos, 2011] Chun-Nam Yu and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *NIPS*, 2011.