A Multivariate Clustering Approach for Infrastructure Failure Predictions

Simon Luo*[§], Victor W. Chu[†], Jianlong Zhou*, Fang Chen*[‡], Raymond K. Wong[‡] and Weidong Huang[§]
*DATA61, CSIRO, Eveleigh, NSW, Australia
[†]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[‡]School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia
[§]HITLAB Australia, School of Engineering and ICT, University of Tasmania, Launceston, TAS, Australia
Email: simon.luo,jianlong.zhou,fang.chen@data61.csiro.au, wchu@ntu.edu.sg, wong@cse.unsw.edu.au, tony.huang@utas.edu.au

***Abstract*—Infrastructure failures have severe consequences which often have a negative impact on the society and the economy. In this paper, we propose a machine learning model to assist in risk management to minimise the cost of infrastructure maintenance. Due to the vast volume and complexity of infrastructure datasets, such problem is often computationally expensive to compute. A Bayesian nonparametric approach has been selected for this problem, as it is highly scalable. We propose a two-stage approach to model failures, such as water pipe failures. The first stage uses an Infinite Gamma-Poisson Mixture Model to group water pipes with similar characteristics together based on the number of failures. The second stage uses the groups created in the first stage as an input to the Hierarchical Beta Process (HBP) to rank water pipes based on their probability of failure. The proposed method is applied to a metropolitan water supply network of a major city. The experiment results have shown that the proposed approach is able to adapt to the complexity of the large multivariate dataset and there is a double-digit improvement from the grouping created by domain experts.**

***Keywords*-Hierarchical Beta Process; Dirichlet Process; Infrastructure Failure Prediction;Water Pipe Failure Prediction; Clustering; Big Data; Sparse Data**

I. Introduction

Infrastructure provides the fundamental systems required for a city to function. Maintaining a city's infrastructure often requires a large amount of financial planning and risk management. Machine learning models are often developed to assist companies managing infrastructure by modelling the failure characteristics of infrastructure networks such as water supply networks, sewage, electric grids and telecommunications. Due to the vast size of these networks, there is a huge amount of data collected from the intrinsic attributes of the infrastructure network and the external environmental factors. The models developed must be able to model large datasets as these infrastructures can span thousands of kilometres. This paper proposes a Bayesian nonparametric approach to predict the failure characteristics of large infrastructure networks. Bayesian nonparametric approaches in general are highly scalable models which can adapt to the complex structure of the dataset.

A water supply network has been selected as a case study for the proposed approach. Water supply network is a key infrastructure that is responsible for distributing water resources. Water pipe failures lead to disastrous consequences which often have a significant impact on the economy and society. Statistical models have been developed to assist in the risk management for the water pipe network. As it is impractical to manually inspect the entire water pipe network, the model is developed to identify water pipes with a higher risk of failure and prioritise them for manual inspection. The pipes which are not considered to be a high risk will only be renewed reactively.

There have been two main areas to study water pipe failure prediction. These include physical modelling and statistical models. Physical modelling provides a failure prediction by modelling the deterioration process of the water pipes by using factors such as corrosion status index, pipe-soil interaction and hydraulic characteristics modelling. However, for big datasets, physical modellings often have limitations as it may be impractical to collect all physical factors.

Machine learning models are much more adaptable to large datasets as they model the failure behaviour using historical water pipe failure records [1]. The models consider both intrinsic water pipe features and external environmental factors together to make the best prediction. The model assumes a similar failure pattern which have appeared in the past are likely to reappear again in the future. We propose a Bayesian nonparametric approach to model the failure patterns of the water supply network dataset. Bayesian nonparametric models have the ability to scale to large datasets as the parameters of the model grow as the size of the dataset increases. The proposed approach consists of two stages. The first stage uses the Infinite Gamma-Poisson Mixture Model to generate an index for groups of pipes with similar failure characteristics. The second stage uses the groups generated in the first stage as an input to the Hierarchical Beta Process (HBP) [2].

It should be noted that the dataset for water pipe failure is extremely large and consists of three large datasets, the water pipe network data, the environmental factors and the water pipe failure data. The water pipe network data contains all the water pipe intrinsic features such as the material, the year it is laid and diameter of the pipe. The environmental factors include the external factors which may have an effect on the pipe failure, such as soil corrosiveness and tree canopy coverage. The water pipe failure data is a time series data which shows records of the date when a water pipe failure had occurred. Given the size of the dataset, it is difficult for traditional modelling techniques to make accurate water pipe failure predictions.

It is also worth noting that the water pipe data is very sparse. There have been very few failures recorded during its observation period, therefore making traditional data mining techniques unable to make accurate predictions as they have a tendency to overfit the dataset. By grouping the water pipes with similar failure characteristics together, data can be shared between similar pipes for training to improve the performance of the model.

This paper builds on previous research in Li et al. [1] on Hierarchical Beta Process (HBP). HBP requires the water pipe grouping to be predetermined by domain experts based on the failure rate. However, it may be difficult to create such grouping for large multivariate datasets such as the water pipe network. Our contribution in this paper includes developing a multi-stage approach for water pipe failure prediction, by using the Infinite Gamma-Poisson Mixture Model to assign a group index to each pipe. The proposed method has been applied to three metropolitan areas. The performance of the model has shown to produce double-digit improvement compared to the recommendation provided by domain experts.

The following sections presented in this paper include, section II, a discussion of the related work. This is then followed by section III, outlines the proposed approach, where the Infinite Gamma-Poisson Mixture Model and the Hierarchical Beta Process is introduced. Then section IV, presents the application to water pipe failure prediction and finally section V draws the conclusion to this paper.

II. Related Work

The related work covers two main areas, the first outlines the related work in water pipe failure prediction in section II-A, the second explains the related work required to develop the proposed approach in section II-B on Dirichlet process and mixture models and section II-C on block models.

A. Water Pipe Failure Prediction

There have been a number of models proposed to assist in risk management for water pipe failure prediction. In the earliest of work, physical modelling which considers a variety of physical factors such as corrosion, the deterioration process of the water pipes and the pipe-soil interaction have been modelled to find the relationship between the pipes age and the pipe's failure rate. There have been a number of time models proposed with comparable performance, some examples include the time-exponential model [3], time-power model [4] and the time-linear model [5].

Later on, multivariate probabilistic models were developed, which made predictions based on intrinsic pipe features such as the material used to construct the pipe, the diameter of the pipe and the year the pipe is laid. One such approach is a semiparametric model known as the Cox's proportional hazards model. The Cox's model considers a combination of the time and the pipe attributes to make a failure prediction. Another example is the Weibull models and its variants [6], [7]. The Weibull model is a multivariate technique, it uses a Weibull distribution or a Weibull process to model the failure behaviour of the water pipes.

More recently, ranking based approaches have been proposed to assist in risk management for water pipe networks [8]. Rather than developing a model to predict the probability for each water pipe failing, each water pipe is assigned a rank based on the probability of failure. The ranking based approach has been shown to have superior performance compared to the time dependent model, Cox's model and Weibull model in [1]. The method proposed in this paper proposes a ranking based approach to model the failure behaviour using the Hierarchical Beta Process (HBP) proposed in [1].

B. Dirichlet Process and Mixture Models

Dirichlet processes are a family of stochastic processes which are often used as a prior for clustering [9]. Dirichlet processes are highly flexible as they do not require a fixed number of clusters specified beforehand, but rather learn the number of clusters based on the dataset. This makes the Dirichlet process highly scalable to the size of the dataset. Given these properties, Dirichlet processes are often found in the foundations of many Bayesian Nonparametric Mixture Models. Other applications of Dirichlet processes include document analysis [10], musical similarity analysis [11] and DNA sequence analysis [12]. The Infinite Gamma-Poisson Mixture Model uses the Dirichlet process to assign group index with similar failure characteristics.

C. Block Models

Stochastic block models are a generative model used to create groups for subsets with similar characteristics [13]. Bayesian nonparametric block models are an extension to stochastic block models to allow the model to scale to the size of the dataset. An example of a Bayesian nonparametric block model is the Infinite Relational Model (IRM) [14]. The IRM is an unsupervised learning technique which is used to discover systems of related concepts. IRM algorithm assumes that each entity belongs to a cluster, and then simultaneously discovers the clusters while clustering the features using the observed value. It does this by constructing multiple independent Dirichlet processes for grouping the components. The advantage of IRM over other clustering techniques is that it does not require a fixed number of clusters in advance, as the number of clusters grow as the number as more data is encountered. Applications of block models include finding system relations [14], link prediction algorithms [15] and clustering categorical datasets [16].

III. Proposed Approach

The proposed method is a two-stage approach. The first step uses the Infinite Gamma-Poisson Mixture Model to create groups and the second step uses the Hierarchical Beta Process to rank the pipes based on the probability of failure. This section will introduce the Hierarchical Beta Process first in section III-A to provide an understanding of the requirements of the grouping. This is then followed by the grouping algorithm, the Infinite Gamma-Poisson Mixture Model in section III-B.

A. Hierarchical Beta Process

The Hierarchical Beta Process (HBP) is the model selected to make the water pipe failure prediction. This section first explains the beta process then its extension to the beta-Bernoulli process, followed by the hierarchical modelling.

1) Beta Process: The beta process was first developed for applications in survival analysis [17]. It was later generated for general propose by Thibaux and Jordan [2]. The beta process $B \sim \text{BP}\left(c, B_0\right)$ can be defined as a Lévy process with a positive random measure B on space Ω. The Lévy process depends on two parameters, the concentration function c and the base measure H_0. For the special case where $c\left(\omega\right)$ and $H_0\left(\omega\right)$ are constant, they are called the concentration and the mean parameters respectively. The Lévy measure for a the beta process for a disjoint infinitesimal partition of Ω can be generated by equation (1).

$$\begin{aligned} H\left(B_k\right) &\sim \text{Beta}\left(cH_0\left(B_k\right), c\left(1 - H_0\left(B_k\right)\right)\right), \\ H\left(\omega\right) &= \sum_l \pi_l\left(\delta_{w_l}\right), \\ \pi_l &\sim \text{Beta}\left(cq_l, c\left(1 - q_l\right)\right), \end{aligned} \tag{1}$$

As the beta process is defined over a general space Ω, the beta process can be used as prior distribution for Bayesian nonparametric models.

2) beta-Bernoulli Process: The observation of the water pipe failure can be modelled using a Bernoulli process $X_j \sim \text{BeP}\left(H\right)$ with the measure H, where j represents the index for each draw. A draw from a Bernoulli process can be represented by the stick breaking process shown in equation (2), where δ_{ω_i} corresponds to the same atom location of H.

$$\begin{aligned} X_j\left(\omega\right) &= \sum_l x_{l,j}\delta_{\omega_l}\left(\omega\right), \\ x_{l,j} &\sim \text{Bernoulli}\left(\pi_l\right), \end{aligned} \tag{2}$$

The beta process is a conjugate prior of the Bernoulli process and can be solved highly efficiently. The posterior distribution of the beta process and Bernoulli process can be thought of as the probability of failure based on the individual observations and can be modelled by a beta process with modified parameters as shown in equation (3). The full derivation can be found in Hjort [17].

$$B|X_{1,\dots,n} \sim \text{BP}\left(c + n, \frac{c}{c+n}B_0 + \frac{1}{c+n}\sum_{i=1}^{n} X_i\right), \tag{3}$$

3) Hierarchical Beta Process: The beta-Bernoulli process is capable of using the observation to model the probability of each pipe failing. However, as the water pipe failure data is very sparse, only very few water pipe failures have been observed during the life span of the water pipe. Therefore considering each pipe individually is impractical as the model will have a tendency to overfit the dataset. The beta-Bernoulli process is also unable to consider other attributes which may effect the pipe failure such as the environmental factors and intrinsic pipe features.

To overcome this issue a hierarchical model has been constructed by attaching a beta process to the beta-Bernoulli process to create the Hierarchical Beta Process (HBP). The HBP groups pipes with similar characteristics together, to allow information of the pipes to be shared to make better failure predictions. The algebraic form and graphical model of HBP is shown in equation (4) and Figure 1a, where c_k and q_k are the concentration parameters and the mean parameters for the group k respectively.

$$\begin{aligned} q_k &\sim \text{Beta}\left(c_0q_0, c_0\left(1 - q_0\right)\right), & k &\in [1, \dots, K] \\ \pi_l &\sim \text{Beta}\left(c_kq_k, c_l\left(1 - q_k\right)\right), & l &\in [1, \dots, L] \\ x_{l,j} &\sim \text{Bernoulli}\left(\pi_l\right), & j &\in [1, \dots, m_l] \end{aligned} \tag{4}$$

The failure probability of each pipe can be calculated by inferring π. For each pipe l, the pipe belongs to a group k which are predefined groups defined by domain experts. However, defining these groups for large multivariate datasets is quite difficult. Our solution to this problem is to propose a model in section III-B known as the Infinite Gamma-Poisson Mixture Model to provide a grouping for pipes with similar characteristics.

4) Inferencing Method: To solve equation (4), an approximation method proposed in [1] has been used for computational efficiency. The final approximation to calculate the parameter q_k is given by equation (5).

$$\begin{aligned} &P\left(q_k|c_k, \{z_l\} = k, \{y_{l,1,\dots,m}\}_{z_l=k}\right) \sim \\ &\text{Beta}\left(c_0q_0 + \sum_l s_l, c_0\left(1 - q_0\right)\sum_l\sum_t^{m-s_l-1}\frac{c_k}{c_k+t}\right) \end{aligned} \tag{5}$$

To calculate π, the value can be directly sampled from its

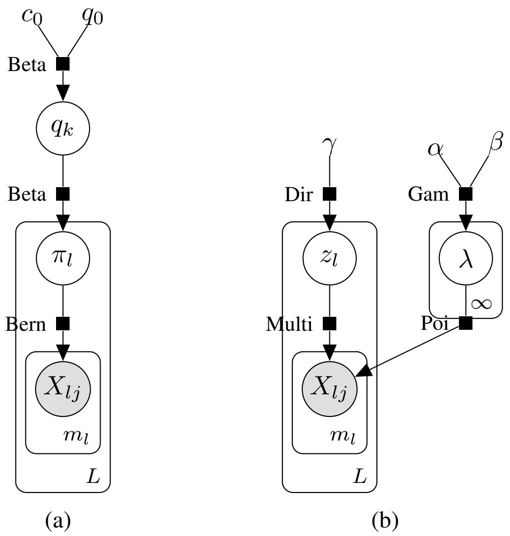


Figure 1: (a) The graphical model for Hierarchical Beta Process. (b) The graphical model for Infinite Gamma-Poisson Mixture Model.

conditional distribution shown in equation (6).

$$P\left(\pi_l | q_{z_l}, c_{z_l}, y_{l,1,\ldots,m}\right) \sim \text{Beta}\left(c_{z_l} q_{z_l} + \sum_{j=1}^{m} y_{l,j}, c_{z_l}\left(1 - q_{z_l}\right) + m - \sum_{j=1}^{m} y_{l,j}\right) \tag{6}$$

B. Flexible Grouping Algorithm

The flexible grouping algorithm has many similarities with the IRM. However, instead of finding system relations between features, the algorithm is designed to group components with similar properties. The IRM can be seen as a variant of the Infinite beta-Bernoulli Mixture Model as it uses the beta-Bernoulli distribution as a base measure for the Dirichlet Process. Replacing the base measure with a gamma-Poisson distribution creates the Infinite Gamma-Poisson Mixture Model. The Infinite Gamma-Poisson Mixture Model groups features with similar observed values into the same cluster. The model assumes that the observations are a mixture of random counts which follow a Poisson distribution.

To define the model, suppose that the observed data x with n observations. Let z_i be the vector of cluster assignment for x_i. The joint distribution of the generative model is given by equation (7). The equation assumes that each data point is conditionally independent from the cluster assignment.

$$P\left(x_1, \ldots, x_N, z_1^{(d)}, \ldots, z_K^{(d)}\right) = \prod_{i=1}^{N} P\left(x_i^{(d)} | z_1^{(d)}, \ldots, z_N^{(d)}\right) \prod_{j=1}^{K} P\left(z_j^{(d)}\right) \tag{7}$$

To complete the generative model in equation (7), the equations for the prior and the likelihood functions are described in the section III-B1 and section III-B2 respectively.

1) Generating Clusters: To generate the clusters a probability distribution is assigned to each partition to allow the clusters to grow. This distribution is chosen from a Dirichlet process. There are many different perspectives on the Dirichlet process, the most common including the stick breaking process, the Chinese Restaurant Process (CRP) and limiting the number of clusters k to infinity in the Dirichlet Process Mixture Model (DPMM). This particular approach uses the CRP perspective on the Dirichlet Process [18]. The CRP is a "Rich gets richer" model, as the clusters attract new members in proportion to its size. But there is also a probability of forming a new cluster which is given by the concentration parameter γ. The CRP can be expressed as equation (8), where n_a is the number of objects already assigned to the cluster.

$$P\left(z_i = a | z_1, \ldots, z_{i-1}\right) = \begin{cases} \frac{n_a}{N-1+\gamma}, & n_a > 0 \\ \frac{\gamma}{N-1+\gamma}, & a \text{ is a new cluster} \end{cases} \tag{8}$$

2) Creating Clusters from Dataset: Clusters can be generated by fitting a distribution to the dataset. The Poisson distribution is chosen to model the total number of failures observed in the dataset. The Gamma distribution is selected to be the base distribution for the Poisson distribution as it is a conjugate prior to the Poisson distribution. The Gamma-Poisson distribution models returns a higher probability for data points with similar value, therefore creating clusters for data points with similar values. The Gamma-Poisson distribution can be combined with the CRP by using a DPMM, hence the complete generative model of the Infinite Gamma-Poisson Mixture Model is expressed as equations (9). The graphical model can also be seen in Figure 1b.

$$\begin{aligned} z_n^{(d)} &\sim \text{CRP}\left(\gamma\right), \\ \lambda_{k^{(1)},\ldots,k^{(D)}} &\sim \text{Gamma}\left(\alpha, \beta\right), \\ x_n &\sim \text{Poisson}\left(\lambda_{z_n^{(1)},\ldots,z_n^{(D)}}\right), \end{aligned} \tag{9}$$

Where, $d \in [1 \ldots D]$, $n \in [1 \ldots N]$ and $k^{(d)} \in \left[1, \ldots, K^{(d)}\right]$. The parameter λ can be seen as the average value over the each cluster. The average value of each cluster λ is used as an input parameter for the Poisson distribution which models the distribution of values for each cluster.

3) Inferencing Algorithm: As the solution for the joint distribution in equation (7) is not tractable analytically, a Markov Chain Monte Carlo (MCMC) approach is taken to approximate the solution [19]. As all conditional distributions can be calculated analytically, a Gibbs sampling method is the chosen inferencing algorithm. Gibbs sampling allows variables to be repeatedly updated one-by-one until the solution has reached convergence.

The conditional probability of the likelihood function $P(x_i|z)$ in equation (7) can be computed analytically as conjugate priors have been used on λ. The derivation of the analytical solution is shown in equation (10), where $\bar{X}$ is the mean value of the dataset X and N is the number of data points.

$$\begin{aligned} P(x_i|z) &= \int P(X|z,\lambda)\, P(\lambda)\, d\lambda \\ &= \int \prod_{i=1}^{N} P(x_i|z,\lambda)\, P(\lambda)\, d\lambda \\ &= \frac{\beta^{\alpha}\Gamma\left(\bar{X}N+\alpha\right)}{\Gamma(\alpha)\,(N+\beta)^{\bar{X}N+\alpha}} \left[\frac{1}{\prod_{i=1}^{N} x_i!}\right] \end{aligned} \tag{10}$$

The conditional probability to assign a data point to each cluster for each Gibbs step is given by equation (11).

$$P(z=a|x_i) \propto P(x_i|z=a)\, P(z_i=a|z_1,\ldots,z_{i-1}) \tag{11}$$

The conditional probability distribution for the cluster assignment is calculated for each data point. The cluster assignment is picked using a multinomial. This is repeated until the solution for the cluster assignment has converged. The full Gibbs sampling algorithm is outlined in Figure 2.

```
Input: dataset $X$, hyper-parameter $\gamma$, number of iterations $T$
Output: Group index $Z$
Start Gibbs sampling
for $t = 1, \ldots, T$ do
    Iterate through all data points
    for $n = 1, \ldots, N$ do
        Remove current data point from cluster
        Draw new value using multinomial from equation (11)
        Assign data point to new cluster
    end for
end for
```

Figure 2: The Gibbs sampling algorithm for the Infinite Gamma-Poisson Mixture Model.

4) Grouping for Multivariate Datasets: The Infinite Gamma-Poisson Mixture Model shares many similarities with IRM, many of the properties which are applies to IRM also applies to Infinite Gamma-Poisson Mixture Model. IRM considers each dimension independently, then merges the relations together by considering the number of unique combinations. The same property can be applied to the Infinite Gamma-Poisson Mixture Model. The group index can be calculate for each dimension and later merged together by considering the number of unique combinations. This makes the Infinite Gamma-Poisson Mixture Model highly scalable to the size of the dataset.

Due to the high dimensional dataset, we consider the following toy example as a scenario to visualise group index generated by the Infinite Gamma-Poison Mixture Model. A water pipe network data, with the laid year and the diameter are features that have been identified to be factors which may effect the failure rate. The failure rate along with each of the feature combinations are tabulated in Table I.

Table I: Toy dataset to demonstrate the grouping generated by the Infinite Gamma-Possion Mixture Model.

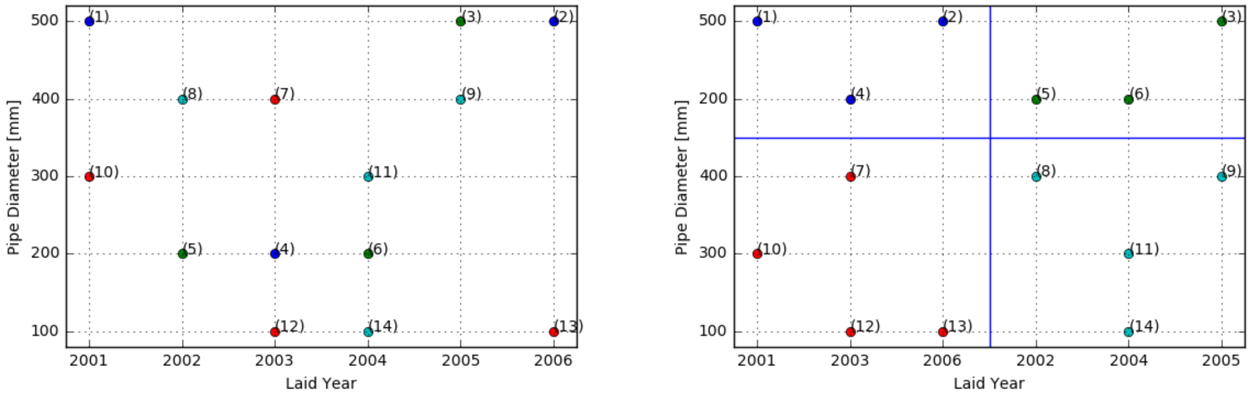
Unique Group ID	No. Pipes	Laid Year	Size[mm]	No. Failure
1	100	2001	500	2
2	100	2006	500	3
3	100	2005	500	10
4	100	2003	200	2
5	100	2002	200	11
6	100	2004	200	10
7	100	2003	400	7
8	100	2002	400	16
9	100	2005	400	15
10	100	2001	300	6
11	100	2004	300	17
12	100	2003	100	8
13	100	2006	100	6
14	100	2004	100	16

The Infinite Gamma-Poisson Mixture Model considers each of the features individually. That is, the model proposes a group index for the laid year and the size independently. Then the unique combinations for each group index for each feature is assigned as the final group index for each group. The grouping generated is tabulated in Table II. It can be seen in the final grouping, that pipes with similar failure counts are placed within the same group. For instance, group 1 has failure counts ranging from 2-3, group 2 has failure counts ranging from 10-11, group 3 has failure counts ranging from 6-8 and group 4 has failure counts ranging from 15-16.

Table II: The group assignment generated by the Infinite Gamma-Poisson Mixture Model

Laid Year		Pipe Size		Combined	
ID	Index	ID	Index	Failures	Index
1	1	1	1	2	1
2	1	2	1	3	1
3	2	3	1	10	2
4	1	4	1	2	1
5	2	5	1	11	2
6	2	6	1	10	2
7	1	7	2	7	3
8	2	8	2	16	4
9	2	9	2	15	4
10	1	10	2	6	3
11	2	11	2	17	4
12	1	12	2	8	3
13	1	13	2	6	3
14	2	14	2	16	4

A visual representation of this table is shown in Figure 3. Figure 3a shows that the grouping generated by the Infinite Gamma-Poisson Mixture Model is complex and cannot be



(a)

(b)

Figure 3: The results from the Infinite Gamma-Poisson Mixture Model for the toy dataset. The colour of the dot represents a group with similar failure characteristics. (a) Shows the results with axis in numerical order. (b) Shows the results with the axis reordered so that the groups with similar failure characteristics are together.

separated by a simple boundary. However, these groups generated are not random, to better visualise how the grouping is generated by the Infinite Gamma-Poisson Mixture Model, the axes in Figure 3b are reordered so that groups of pipes with similar number of failures are grouped together. After reordering the axes, a vertical line and a horizontal line can be drawn to separate each group.

As each feature is treated independently, the complexity does not increase as more features are added to the model. The model is also able to group in complexity as the number of data points increases, with these two properties, this makes the group assignment generated by Infinite Gamma-Poisson Mixture Model highly scalable to the size of the dataset.

IV. Application To Water Pipe Failure Prediction

This section applies the proposed approach to water pipe failure prediction. This section first outlines the data collected in section IV-A. This is then followed by the prediction results which are discussed in section IV-B.

A. Data Collection

Three metropolitan areas are selected for the experiment. These regions have been selected to represent a range of metropolitan areas: region A represents a high population density area, region B represents a medium population density area and region C represents a low population density area. The details of these regions are tabulated in Table III.

Table III: Details of the metropolitan areas for each water supply network

Region	Area[km^2]	Pop.	Pop. Density[per km^2]
Region A	25	205 000	8 200
Region B	85	230 000	2 706
Region C	685	210 000	307

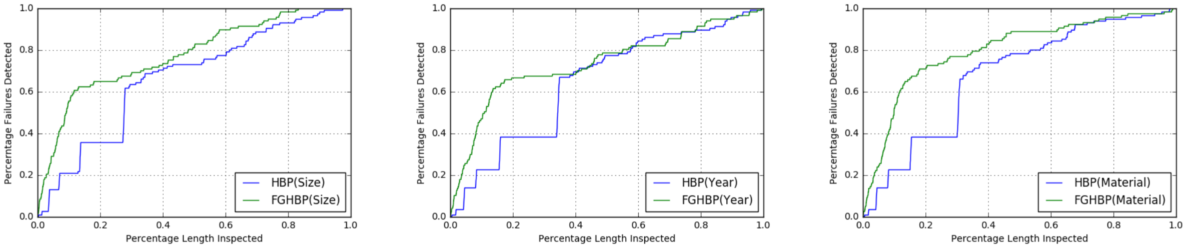
For the dataset collected for the three metropolitan areas, the pipes are laid between years 1884 - 2011. These pipes are often split into two main categories, reticulation water main (RWM) and critical water main (CWM). The categories of these pipes are defined using the pipe diameter. Pipes which have a diameter less than 300 millimetres are categorised as RWM, while pipes with a diameter greater or equal to 300 millimetres are categorised to be CWM. The ratio of these pipes for each region is summarised in Table IV. As the role of RWM and CWM is different, it is expected that the failure behaviour of RWM and CWM is also different. Therefore RWM and CWM are considered separately. This particular study will focus on the RWM.

Table IV: Summary of pipe network and pipe failure data

Region	Type	No. Pipes	No. Failures	Total Length[m]
Region A	RWM	12 866	1 586	571 966
	CWM	1 945	131	116 938
Region B	RWM	9 687	2 702	388 262
	CWM	1 000	98	127 060
Region C	RWM	15 951	4 249	1 005 346
	CWM	2 050	172	209 495

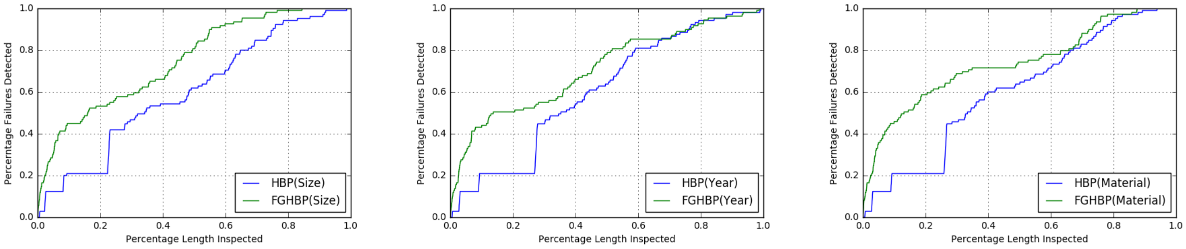
B. Prediction Results

The prediction results for HBP have used grouping which have been specified by domain experts. The grouping specified by domain experts attempts to group water pipes with similar failure rates together. The proposed approach is labelled as Flexible Grouping Hierarchical Beta Process (FGHBP). Domain experts have provided grouping approaches for 3 different features. These include the year the pipe is laid, the diameter of the pipe and the material



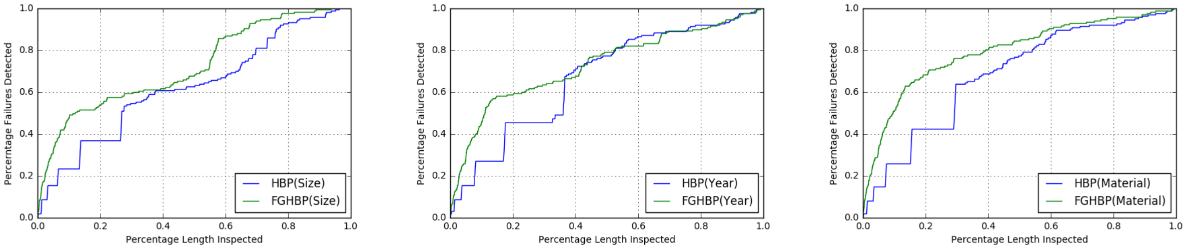
(a) (b) (c)

Figure 4: Comparisons of methods for Region A. (a) Pipe Diameter (size), (b) year the pipe is laid, (c) material used to manufacture the pipe.



(a) (b) (c)

Figure 5: Comparisons of methods for Region B. (a) Pipe Diameter (size), (b) year the pipe is laid, (c) material used to manufacture the pipe.



(a) (b) (c)

Figure 6: Comparisons of methods for Region C. (a) Pipe Diameter (size), (b) year the pipe is laid, (c) material used to manufacture the pipe.

Table V: A comparison of results for each region using Area Under Curve (AUC).

	Region A			Region B			Region C		
	Material	Year	Size	Material	Year	Size	Material	Year	Size
HBP	67.84%	64.81%	66.72%	60.43%	60.57%	60.34%	68.53%	67.14%	62.57%
FGHBP	79.87%	73.62%	77.66%	73.53%	71.08%	75.11%	79.34%	72.95%	72.26%

used to construct the pipe. All 3 grouping approaches have been compared with the proposed approach.

The performance of the model for each region is shown in Figure 4, Figure 5 and Figure 6. The horizontal axis represents the cumulative length of the inspected pipes as a percentage. The vertical axis shows the percentage of failure detected for each region in 2012. A model has a better performance if it can detect more failures at a particular length with respect to the percentage of length inspected. To provide a quantitative comparison between these results Area Under Curve (AUC) is calculated for each model. These results are shown in Table V.

The results have shown that the proposed approach have improve the performance of the failure prediction significantly. The HBP uses water pipes with similar failure behaviour to predict the likelihood of a water pipe to fail. The Infinite Gamma-Poisson Mixture Model is able to provide a higher quality grouping compared to the domain experts, thus leading to the higher performance in failure prediction across all regions.

V. Conclusion

This paper has proposed a two-stage approach for infrastructure failure prediction. Datasets containing infrastructure information are often extremely large and extremely sparse. The proposed approach uses a two-stage approach to tackle this problem. The first stage uses an Infinite Gamma Poisson Mixture Model to group water pipes with similar failure rates. The second stage uses the Hierarchical Beta Process rank for the water pipe failure prediction. As the models from both stages are built using a Bayesian nonparametric framework, the proposed approach is highly scalable to large datasets.

The proposed approach has been applied to water pipe failure prediction to demonstrate its ability to scale for large datasets. The results have shown that the proposed approach has an increase in performance compare to the grouping given by domain experts across all regions. Using the AUC measure, region A has shown to have an increase of 12.03%, region B has shown to increase 14.77% and region C 10.81%.

References

[1] Z. Li, B. Zhang, Y. Wang, F. Chen, R. Taib, V. Whiffin, and Y. Wang, "Water pipe condition assessment: a hierarchical beta process approach for sparse incident data," *Machine learning*, vol. 95, no. 1, pp. 11–26, 2014.

[2] R. Thibaux and M. I. Jordan, "Hierarchical beta processes and the indian buffet process." in *AISTATS*, vol. 2, 2007, pp. 564–571.

[3] U. Shamir, C. Howard *et al.*, "An analytical approach to scheduling pipe replacement (pdf)," *Journal-American Water Works Association*, vol. 71, no. 5, pp. 248–258, 1979.

[4] K. Mavin, *Predicting the failure performance of individual water mains*. Urban Water Research Association of Australia, 1996.

[5] A. Kettler and I. Goulter, "An analysis of pipe breakage in urban water distribution networks," *Canadian Journal of Civil Engineering*, vol. 12, no. 2, pp. 286–293, 1985.

[6] J. G. Ibrahim, M.-H. Chen, and D. Sinha, *Bayesian survival analysis*. Wiley Online Library, 2005.

[7] S. Osaki, D. P. Murthy, and R. J. Wilson, *Stochastic models in engineering, technology and management*, 1995.

[8] R. Wang, W. Dong, Y. Wang, K. Tang, and X. Yao, "Pipe failure prediction: A data mining method," in *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*. IEEE, 2013, pp. 1208–1218.

[9] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[10] "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006. [Online]. Available: http://www.jstor.org/stable/27639773

[11] M. D. Hoffman, D. M. Blei, and P. R. Cook, "Content-based musical similarity computation using the hierarchical dirichlet process." in *ISMIR*, 2008, pp. 349–354.

[12] J. P. Huelsenbeck, S. Jain, S. W. Frost, and S. L. K. Pond, "A dirichlet process model for detecting positive selection in protein-coding dna sequences," *Proceedings of the National Academy of Sciences*, vol. 103, no. 16, pp. 6263–6268, 2006.

[13] H. C. White, S. A. Boorman, and R. L. Breiger, "Social structure from multiple networks. i. blockmodels of roles and positions," *American journal of sociology*, vol. 81, no. 4, pp. 730–780, 1976.

[14] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, "Learning systems of concepts with an infinite relational model," in *AAAI*, vol. 3, 2006, p. 5.

[15] D. I. Kim, M. Hughes, and E. Sudderth, "The nonparametric metadata dependent relational model," *arXiv preprint arXiv:1206.6414*, 2012.

[16] J. Lee, P. Müller, Y. Zhu, and Y. Ji, "A nonparametric bayesian model for local clustering with application to proteomics," *Journal of the American Statistical Association*, vol. 108, no. 503, pp. 775–788, 2013.

[17] N. L. Hjort, "Nonparametric bayes estimators based on beta processes in models for life history data," *The Annals of Statistics*, pp. 1259–1294, 1990.

[18] J. Pitman, *Combinatorial Stochastic Processes: Ecole D'Eté de Probabilités de Saint-Flour XXXII-2002*. Springer, 2006.

[19] R. M. Neal, "Markov chain sampling methods for dirichlet process mixture models," *Journal of computational and graphical statistics*, vol. 9, no. 2, pp. 249–265, 2000.