

Chapter 18

Water Pipe Failure Prediction: A Machine Learning Approach Enhanced By Domain Knowledge



Bang Zhang, Ting Guo, Lelin Zhang, Peng Lin, Yang Wang, Jianlong Zhou and Fang Chen

Abstract Drinking water pipe and waste water pipe networks are valuable urban infrastructure assets that are responsible for reliable water resource distributions and waste water collection. However, due to fast growing demand and aging assets, water utilities find it increasingly difficult to efficiently maintain their pipe networks. Pipe failures - drinking water pipe breaks and waste water pipe blockages - can cause significant economic and social costs, and hence have become the primary challenge to water utilities. Identifying key influential factors, e.g., pipes' physical attributes, environmental features, is critical for understanding pipe failure behaviours. The domain knowledge plays a significant role in this aspect. In this work, we propose a Bayesian nonparametric machine learning model with the support of domain knowledge for pipe failure prediction. It can forecast future high-risk pipes for physical condition assessment, thereby proactively preventing disastrous failures. Moreover, compared with traditional machine learning approaches, the proposed model considers domain expert knowledge and experience, which helps avoid the limit of traditional machine

B. Zhang (✉) · T. Guo · L. Zhang · P. Lin · Y. Wang · J. Zhou · F. Chen
DATA61, CSIRO, 13 Garden Street, Eveleigh, NSW 2015, Australia
e-mail: mattbang.zhang@data61.csiro.au

T. Guo
e-mail: Ting.Guo@data61.csiro.au

L. Zhang
e-mail: Lelin.Zhang@data61.csiro.au

P. Lin
e-mail: peng.lin.c@gmail.com

Y. Wang
e-mail: Yang.Wang@data61.csiro.au

J. Zhou
e-mail: Jianlong.Zhou@data61.csiro.au

F. Chen
e-mail: Fang.Chen@data61.csiro.au

learning approaches - learning only from what it sees - and improves prediction performance.

18.1 Introduction

Pipe networks are valuable urban infrastructure assets that are responsible for reliable water resource distributions and waste water collection. However, as urbanisation trends continue and urban populations rise, water utilities find it increasingly difficult to meet growing water demand with ageing and failing water pipe networks. Water pipe failures, which can cause tremendous economic and social costs have become the primary challenge to water utilities. In order to tackle the problem in a financially viable way, preventative risk management strategies are widely adopted by water utilities to prevent disastrous failures. The basic idea of the strategies is to proactively identify high-risk pipes and renew them in time to avoid potential failures. Meanwhile, replacement of pipes that are still in healthy condition is to be avoided. Accordingly, the strategies consist of two main steps: (1) high-risk pipe prioritisation, in which pipes are ranked based on their risk of failure, and (2) physical condition assessment, in which physical inspections are conducted on highly rated pipes to confirm their actual condition for replacements. The pipes, which are not identified as high-risk pipes at the prioritisation step, will only be renewed reactively. Hence, the success of the strategies relies heavily on the prioritisation step. To make accurate selections of high-risk pipes, the prioritisation step requires a failure prediction method that can give a precise estimation of pipe failure likelihood, based on which the estimated failure cost and renewal cost can be readily obtained.

The problem of estimating water pipe failure risk has been studied for many decades. There are two main methodologies for tackling the problem, namely data-driven modelling and domain knowledge-driven modelling.

For domain knowledge-driven physical modelling, a variety of models has been proposed for explaining and predicting the deterioration processes of water pipes. They usually consider an individual aspect of the problem based on the domain knowledge in the related area, such as pipe-soil interaction analysis, residual structural resistance, or hydraulic characteristics modelling. A comprehensive review can be found in [14]. For data-driven statistical machine learning-based modelling, it assumes that pipes with similar intrinsic attributes share similar failure patterns, and that failure patterns which have appeared before are likely to reappear in the future. The patterns can be learnt from the available factors and data sets.

Both methodologies have limitations. For domain knowledge-based physical models, they often just consider one aspect of the problem, e.g., corrosion, and lack the ability to learn knowledge from heterogeneous features. While, for data-driven statistical machine learning-based models, they usually learn from what they see, i.e., learning from the provided basic features, and lack the ability to identify and include the informative features that only domain experts are aware of, e.g., a significant proportion of the waste water pipe failures (blockages) are caused by tree

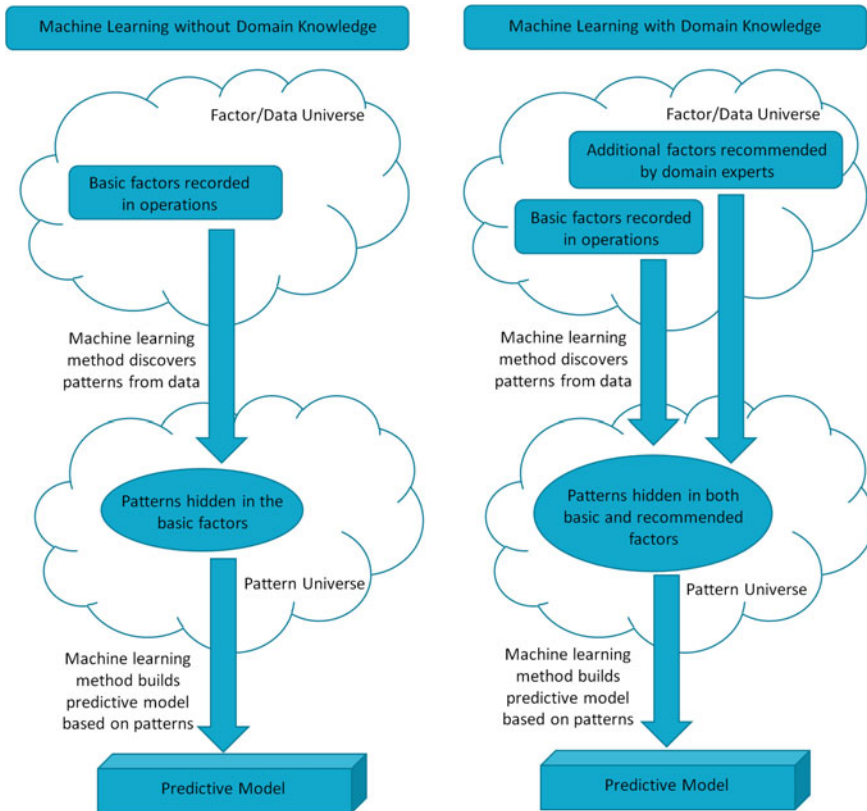


Fig. 18.1 Traditional machine learning versus machine learning with domain knowledge. Traditional machine learning methods suffer the limit of learning only from what they see. Domain knowledge can help avoid such limit via transferring the knowledge into machine learning models

root penetration. Therefore, in this work, we suggest and demonstrate that the incorporation of domain knowledge into machine learning methods could significantly improve the model performance. Figure 18.1 illustrates the difference between traditional machine learning methods and the machine learning methods considering domain knowledge.

For the modelling perspective, in order to improve high-risk pipe prioritisation for large-scale metropolitan pipe networks, we propose a Bayesian nonparametric statistical approach, namely the Dirichlet process mixture of hierarchical beta process model, for water pipe failure prediction. Unlike parametric approaches, the structure and complexity of the proposed model can grow as the amount of observed data increases. It makes the model invulnerable to faulty assumptions of model forms and adaptable to various failure patterns, thereby leading to more accurate predictions for different application scenarios.

It is worth noting that water pipe failure data is extremely sparse in reality. Very few pipes have failure records during the observation period. Such sparsity makes



Fig. 18.2 Water supply networks in the selected regions

traditional failure prediction methods incompetent for accurate pipe failure prediction since most pipes do not have failure data for training. The proposed approach deals with this issue by sharing failure data via a flexible hierarchical modelling of failure behaviours. The key component of the hierarchical modelling is a flexible grouping scheme. It clusters similar pipes together for modelling so that failure data can be shared by similar pipes for training.

Additionally, domain experts' experience, i.e., helping identify potential useful features for building the model and rejecting false correlated features, also helps tackle the data sparsity challenge.

The proposed method has been applied to the pipe network of an international metropolis that has a total population of near five million people. In this work, three representative regions are selected from the metropolis for comparison experiments. The regions and the networks are shown in Fig. 18.2. As we can see, the water supply network is constituted of two main categories of water pipes, critical water main (CWM) indicated by red lines and reticulation water main (RWM) indicated by blue lines. CWMs have larger diameters (300 mm and above), and RWMs have smaller diameters (smaller than 300 mm). Each water pipe is composed of a set of pipe segments connected in series. Failure records can be precisely matched with pipe segments, allowing the proposed method to model failure behaviours of pipe segments.

The rest of the chapter is organised as follows. Section 18.2 reviews the related work. Section 18.3 describes the details of the proposed method. Empirical studies and the importance of the domain knowledge are shown in Sect. 18.4. The conclusions are drawn in Sect. 18.5.

18.2 Related Work

In the past decades, a large number of statistical approaches have been proposed for water pipe failure prediction with significant success. However, most of them need

to pre-define the form of the model, hence lack the flexibility of modelling complex situations, where the recent Bayesian nonparametric machine learning strategy can readily solve the model selection problem. In this section, we briefly review the related work on statistical water pipe failure prediction methods and Bayesian nonparametric approaches.

18.2.1 Statistical Failure Prediction Methods

In recent decades, many statistical models have been proposed for water pipe failure prediction. In the early stages, various methods were developed for modelling the relationship between pipe age and pipe failure rate. For instance, the work in [15] proposed a time-exponential model, which formulates the number of failures per unit length per year as an exponential function of pipe age. Similarly, time-power model [12] and time-linear model [9] were developed with comparable performances.

Later, multivariate probabilistic models were suggested. They make predictions based on a variety of pipe attributes, such as age, material, length and diameter. One of the most popular multivariate approaches is the Cox proportional hazards model [3]. It is a semi-parametric method, in which the baseline hazard function has an arbitrary form and the pipe attributes alter the baseline hazard function via an exponential function multiplicatively. The Weibull model and its variants [2, 8] are also widely adopted in practice. They utilise either a Weibull distribution or a Weibull process for modelling pipe failure behaviours.

Recently, a ranking-based method [18] was proposed for predicting water pipe failures. It treats failure prediction as a ranking problem. Pipes are ranked based on their failure risk. The method performs failure prediction via a real-valued ranking function rather than an estimation of failure probability.

18.2.2 Bayesian Nonparametric Approaches

All the aforementioned methods are parametric or semi-parametric, which means the forms of the methods are predefined and fixed during the training process. If the assumptions made on the model form are not satisfied, accurate predictions cannot be achieved. In contrast, Bayesian nonparametric approaches do not make assumptions about the model structure. Instead, their model complexities grow as the amount of observed data increases, endowing Bayesian nonparametric approaches with flexibility for modelling complex real-world data.

The Beta process [5] and the Dirichlet process [4] are two Bayesian nonparametric approaches that were developed recently with tremendous success in a variety of domains. They have become the cornerstones for building more sophisticated Bayesian nonparametric models.

The Beta process was originally developed for survival analysis on life history data. It was utilised as a prior distribution over the space of cumulative hazard function. Later, the work in [17] extended the Beta process to more general spaces for different applications, such as factor analysis [13], image reconstruction [20, 21], image interpolation and document analysis [17]. One of its variants was also applied to water pipe failure prediction [10, 11].

The Dirichlet process [4] is a flexible Bayesian nonparametric prior for data clustering. It does not set any assumptions on the number of clusters. Instead, it allows the number of clusters to grow as the number of data points increases. It is the foundation of many nonparametric mixture models, and has been widely adopted in various applications, such as document analysis [16], musical similarity analysis [6] image annotation [19] and DNA sequence analysis [7].

18.3 The Proposed Method

The proposed Dirichlet process mixture of an hierarchical beta process model consists of two main components working with each other interactively: a hierarchical representation of water pipe failure behaviours and a flexible pipe grouping scheme. The grouping scheme generates a set of groups, on each of which the hierarchical representation can be constructed. The hierarchical representation provides a precise modelling of each group's failure behaviours, hence acts as the basis of grouping.

The two main components are described in Sects. 18.3.1 and 18.3.2 respectively. The details of the proposed model are given in Sect. 18.3.3.

18.3.1 Hierarchical Modelling of Water Pipe Failure Behaviours

The hierarchical beta process is adopted in this work as the hierarchical modelling of water pipe failure behaviours. We first briefly introduce the beta-Bernoulli process for modelling failure event and failure probability in Sects. 18.3.1.1 and 18.3.1.2. Then the details of the hierarchical modeling are given in Sect. 18.3.1.3.

18.3.1.1 Beta Process

On a measurable space Ω , a beta process H is defined as a positive Levy process, a positive random measure whose masses on disjoint subsets of Ω are independent. It is parameterised by a positive concentration function c and a base measure H_0 , which is also defined on space Ω . In simplified cases, where function $c(\omega_i)$ becomes a constant, we call c concentration parameter.

For disjoint infinitesimal partitions of Ω , the beta process can be generated as:

$$H(B_k) \sim \text{Beta}(cH_0(B_k), c(1 - H_0(B_k))), \quad (18.1)$$

where B_k indicates a partition, and $k \in \{1, \dots, K\}$ is the index. The process can be denoted as $H \sim BP(c, H_0)$.

When the base measure H_0 is discrete and has a set function form of $H_0 = \sum_i p_i \delta_{\omega_i}$, H turns to have atoms at the same locations as H_0 's and can be written in a set function form accordingly as:

$$\begin{aligned} H(\omega) &= \sum_i \pi_i \delta_{\omega_i}(\omega) \\ \pi_i &\sim \text{Beta}(cq_i, c(1 - q_i)) \end{aligned} \quad (18.2)$$

where $\delta_{\omega_i}(\omega) = 1$ when $\omega = \omega_i$ and 0 otherwise.

As defined in a general space Ω , the Beta process provides us a flexible Bayesian nonparametric prior for water pipe failure events which themselves can be modelled by the Bernoulli process.

18.3.1.2 Bernoulli Process

For a Bernoulli process $BeP(H)$, each of its draws X_j is again a measure on space Ω . j represents the draw index. H indicates a beta process on Ω , as defined before. It acts as the prior of the Bernoulli process. A draw of the Bernoulli process can also be represented via a set function form as:

$$\begin{aligned} X_j(\omega) &= \sum_i x_{ij} \delta_{\omega_i}(\omega) \\ x_{i,j} &\sim \text{Bernoulli}(\pi_i) \end{aligned} \quad (18.3)$$

where δ_{ω_i} corresponds to the same atom location of H . The random variable x_{ij} is generated from a Bernoulli distribution parameterised by π_i which is defined as Eq. 18.2. With x_{ij} as its elements, an infinite binary column vector, also denoted by X_j , can be used for representing a draw of the Bernoulli process. Then the draws of the Bernoulli process can form an infinite binary matrix X , with X_j representing a column and j representing the column index. Each row of the matrix corresponds to an atom location δ_{ω_i} . We can see that the beta process appears to be a proper Bayesian nonparametric prior for such infinite binary matrices.

It is worth noting that the Beta process is a conjugate prior of the Bernoulli process. Given a beta process prior $H \sim BP(c, H_0)$, and a set of m observations drawn from a Bernoulli process $X_j \sim BeP(H)$, the posterior is again a beta process, with parameters updated as follow:

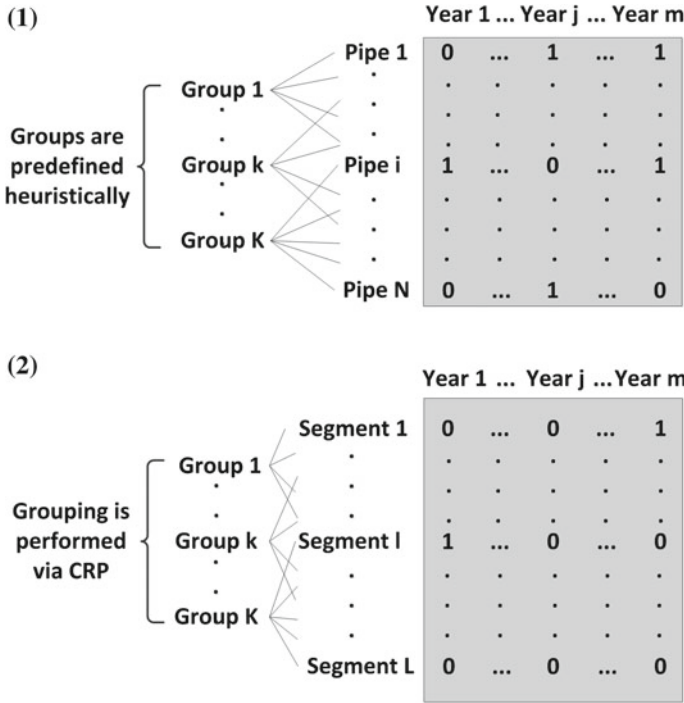


Fig. 18.3 Binary failure matrices for pipes and pipe segments

$$H|X_{1,\dots,m} \sim BP \left(c + m, \frac{c}{c + m} H_0 + \frac{1}{c + m} \sum_{j=1}^m X_j \right) \tag{18.4}$$

The conjugacy significantly simplifies the inference procedure for parameter estimation.

18.3.1.3 Hierarchical Modelling

With the aid of a Beta-Bernoulli process, a hierarchical representation can be developed for modelling water pipe failure behaviours. Firstly, failure events can be modelled by a Bernoulli process $BeP(H)$. Let an infinite binary matrix X , as illustrated in Fig. 18.3 (1), represent failure records of pipes. Each of its columns, X_j , can be treated as a draw from the Bernoulli process $BeP(H)$. It is an infinite binary column vector with the i -th element $x_{i,j}$ generated from $x_{i,j} \sim Bernoulli(\pi_i)$. $x_{i,j} = 1$ means pipe i failed in year j , and $x_{i,j} = 0$ otherwise. Then the beta process, $H \sim BP(c, H_0)$, defined as a positive Levy process on pipe space Ω , can be used as a prior of failure events, namely failure probability. Its set function form is defined as Eq. 18.3.

With beta process H as a prior, each row of the matrix X corresponds to an atom location δ_{ω_i} in the pipe space Ω , which can be infinitely large. We assume that two pipes share the same failure patterns if they have the same intrinsic attributes and environmental factors. Hence, we treat such two pipes as the same in the pipe space Ω . Considering all the possible combinations of pipe attributes and environmental factors, the number of “unique” pipes in the pipe space becomes infinite. Therefore, each column of the matrix X is an infinite binary vector that is drawn from a Bernoulli process. The beta process H is then a conjugate prior of the infinite binary matrix X . It models the failure probabilities of pipes via π_i .

While the Beta-Bernoulli process is capable of modelling failure behaviours as described above, there are two issues of adopting it in practice. Firstly, the number of failures is extremely small compared with the number of pipes, especially for CWMs. Only a small portion of CWMs have failure records since most of the CWMs did not fail during the observation period. Thus, the majority of CWMs have no failure data for model training. Secondly, in addition to pipe failure histories, pipe attributes and environmental factors are also crucial for estimating failure probabilities. However, they are not properly considered in the Beta-Bernoulli process. The fact that the pipes with similar intrinsic attributes and environmental factors often share similar failure patterns is ignored by the Beta-Bernoulli process.

In order to address these issues, the hierarchical beta process (HBP) model [11, 17] can be adopted as a hierarchical modelling of water pipe failure behaviours. Given a water pipe grouping, e.g., grouping by intrinsic attributes, one more beta process can be added into the model hierarchy for modelling the failure behaviours of groups. The new beta process is on top of the existing beta process, serving as the prior of its mean parameter. The graphical model in Fig. 18.4 (1) illustrates the HBP model. It can also be described as the followings:

$$\begin{aligned} q_k &\sim \text{Beta}(c_0 q_0, c_0(1 - q_0)), \quad k \in [1, \dots, K], \\ \pi_i &\sim \text{Beta}(c_k q_k, c_k(1 - q_k)), \quad i \in [1, \dots, N], \\ x_{i,j} &\sim \text{Bernoulli}(\pi_i), \quad j \in [1, \dots, m_i], \end{aligned} \quad (18.5)$$

where π_i and x_{ij} are defined as before, modelling the failure probability of pipe i and failure history of pipe i in year j respectively. q_k and c_k are the mean and concentration parameters for group k . q_k can be regarded as modeling the failure rate of group k . q_0 and c_0 are the hyper parameters.

By adding one more hierarchy level, the HBP model estimates failure probabilities through the inferences on both group level and pipe level. Group level inference estimates the group failure rate q_k , and pipe level inference estimates the pipe failure probability π_i . Failure data can be shared by the same group of pipes for estimating group failure rate q_k . It helps to solve the failure data sparsity problem. The failure patterns that are shared by similar pipes are captured at the group level since the pipes within the same group share the same q_k . At the pipe level, the pipe failure probability π_i is estimated by considering not only the failure observations x_{ij} , but also the group similarity through the group failure rate q_k .

18.3.2 Flexible Water Pipe Grouping

Real world data is complicated and often demonstrates multi-modality property, which is the case for water pipe failures. Consequently, single-modality models become insufficient in such circumstances for modelling the whole data corpora. Mixture model is a widely adopted probabilistic approach for modelling the data arising from different modalities. It assumes that the final model consists of a set of mixture components, each of which can accurately model a portion of data.

For conventional parametric mixture models, the number of mixture components is required to be known in advance, which is unrealistic for many real world applications, such as water pipe grouping. Therefore, we adopt the Dirichlet process (DP), a nonparametric approach, for pipe grouping. It serves as a flexible prior for data partitioning and sets no assumptions on the number of partitions. Correspondingly, the Dirichlet process mixture model, which is built based on the Dirichlet process, can comprise a countably infinite number of components and adjust itself for fitting observed data.

In order to adopt DP as the prior of pipe grouping, we use the Chinese restaurant process (CRP) [1] as the constructive representation of DP. It exhibits the clustering property of DP via the following metaphor. Suppose there is a Chinese restaurant that has an infinite number of tables. A sequence of customers enters and select a table to sit. The first customer sits at the first table. The following customers sit at tables with a guide:

$$p(z_l = r | z_{-l}, \alpha) \propto \begin{cases} \frac{n_r}{n-1+\alpha} & \text{if } r \leq k \\ \frac{\alpha}{n-1+\alpha} & \text{if } r = k + 1. \end{cases} \quad (18.6)$$

z_l indicates a customer, z_{-l} denotes all the customers that appeared before z_l , r indicates a cluster index, and k represents the current number of clusters. n_r is the number of customers in cluster r and α is the concentration parameter for CRP, controlling the probability that a customer is assigned to an unoccupied table.

The CRP offers an exchangeable distribution over the table assignments z_l . The joint distribution is invariant to the order of customers. The procedure of assigning a table for a customer can be performed as he or she is the last customer entering the restaurant. As described by Eq. 18.6, the i -th customer sits at an occupied table with a probability proportional to the number of customers who are already sitting at that table. He or she sits at an unoccupied table with a probability proportional to the concentration parameter α . In this metaphor, customers correspond to data points and tables correspond to clusters. Fig. 18.4 (2) shows the Dirichlet process mixture model with the CRP as the constructive definition. Each data point x_i is drawn from a component of the mixture model. z_i is the component indicator for x_i . θ_k represents the parameter for component k .

With the aid of the CRP, we can group pipes adaptively for fitting data observations. As a result, pipes with similar failure behaviours are grouped together. Moreover,

the CRP helps to integrate the grouping process and the failure modeling process for achieving accurate performance.

18.3.3 *Dirichlet Process Mixture of Hierarchical Beta Process*

In this section, we give the detailed description of the proposed Dirichlet process mixture of the hierarchical Beta process (DPMHBP) model for water pipe failure prediction.

For the proposed DPMHBP model, a water pipe is treated as a set of pipe segments that are connected in series. The failure probability of a pipe segment is modelled by a beta process. It is different from the HBP model [11] where the Beta process is used for modelling failure probabilities of pipes.

Pipe length is an important attribute for estimating failure probability. The intuition is that longer pipes tend to have higher failure probabilities if other attributes and external factors are the same. However, the HBP model ignores the impact of the length attribute when estimating failure probabilities. It only focuses on pipe age attribute and failure histories. The significant variance of pipe lengths is neglected. In order to tackle the problem, the proposed approach suggests modelling the failure probabilities of pipe segments whose lengths are relatively constant with a very small variance.

Another difference between the HBP model and the proposed DPMHBP model is that the HBP model groups pipes based on heuristic domain information e.g., pipe age. Its grouping is predefined and fixed during the inference process. The number of the groups is also required to be set beforehand, which can be heuristic. In contrast, for the proposed DPMHBP method, the grouping process is integrated with the inference process via the DP mixture model. They interact with each other to achieve an optimal model. The number of groups is not fixed and can grow as the size of the training data increases.

Considering all the issues mentioned above, the DPMHBP model can finally be given as follows:

$$\begin{aligned}
 q_k &\sim \text{Beta}(c_0 q_0, c_0(1 - q_0)), & k \in [1, \dots, K], \\
 z_l &\sim \text{CRP}(\alpha), & z_l \in [1, \dots, K], \\
 \rho_l &\sim \text{Beta}(c_{z_l} q_{z_l}, c_{z_l}(1 - q_{z_l})), & l \in [1, \dots, L], \\
 y_{l,j} &\sim \text{Bernoulli}(\rho_l), & j \in [1, \dots, m_l], \\
 \pi_l &= 1 - \prod_{l=1}^{s_i} (1 - \rho_l), & l \in [1, \dots, s_i].
 \end{aligned} \tag{18.7}$$

The failure probability estimation is conducted on three levels: segment group level, segment level and pipe level. The failure events are recorded for segments rather

than pipes. The grouping is performed on segments via the CRP, as illustrated by Fig. 18.3 (2). At segment group level, q_k denotes the failure rate of segment group k . z_l represents the group index for segment l . At segment level, ρ_l indicates the failure probability of segment l . Once the segment level estimation is obtained, pipe failure probability π_i can be readily computed via the failure probability of a series of connected segments. Figure 18.4 (3) shows the graphical model of the DPMHBP model.

It is worth noting that the Bernoulli process is more suitable for modelling segment failures than modelling pipe failures because it is very rare for a segment to fail twice in a year.

Regarding the inference of the model parameters from the training data, since no analytical solution is available for the proposed model, we use a Markov chain Monte Carlo (MCMC) sampling algorithm for inference. Gibbs sampling is the MCMC-based method that has been widely used for DP mixture models when conjugacy exists between prior and likelihood. However, for the DPMHBP model, such conjugacy is broken by the extra hierarchy of the HBP model. Therefore, we choose to utilise a Metropolis-within-Gibbs sampling method for inference.

18.4 Experiments

In this section, we conduct comparison experiments on the metropolitan water supply network data to demonstrate the superiority of the proposed DPMHBP model. We first introduce the pipe network data and the failure data in Sect. 18.4.1. The features that are suggested by domain experts and used in the experiments are explained in Sect. 18.4.2. Then the compared methods are listed in Sect. 18.4.3. Finally, we give the comparison results and discuss the impact of the proposed method in Sect. 18.4.4.

18.4.1 Data Collection

Three representative regions from the metropolis are selected to perform the experiments. Region A is a local government area with a population around 210,000, which is one of the most populous local government areas in its state. Its population density is 629 people per km². Region B is a local government area with a high population density of 2,374 people per km². Its population is about 182,000. Region C is a low density suburban local government area, which has a population of 205,000 and a population density of 300 people per km².

For each region, both network data and failure data are collected. Network data consists of pipe IDs, pipe attributes, pipe locations and environmental factors. Pipe location is represented as a set of connected line segments, each of which corresponds to a pipe segment. Failure data contains pipe IDs, failure dates and failure locations.

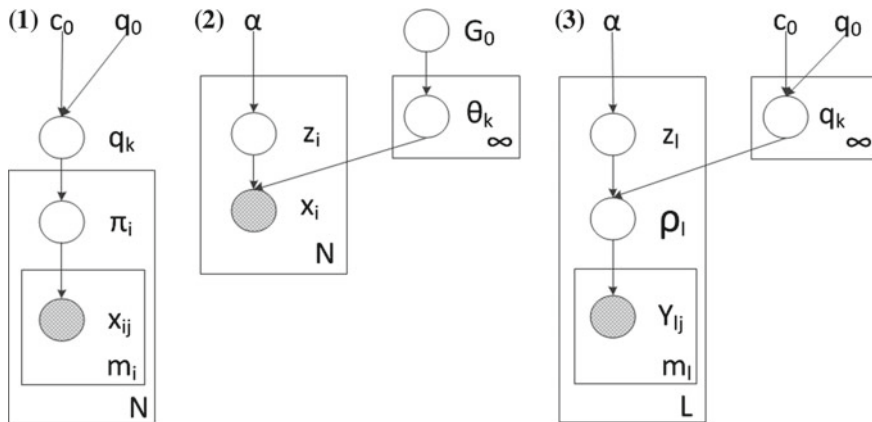


Fig. 18.4 Graphical models for **1** Hierarchical Beta process, **2** Dirichlet process mixture model (with Chinese restaurant process as the constructive definition), **3** Dirichlet process mixture of hierarchical beta process

Table 18.1 Summary of pipe network data and pipe failure data

		# Pipes	# Failures	Laid years	Observation period
Region A	All	15189	4093	1930–1997	1998–2009
	CWM	3793	520	1930–1997	1998–2009
Region B	All	11836	3694	1888–1997	1998–2009
	CWM	2457	432	1888–1997	1998–2009
Region C	All	18001	4421	1913–1997	1998–2009
	CWM	5041	563	1913–1997	1998–2009

Pipe amount, failure amount, laid year range and observation period are summarised for different pipe types in Table 18.1. As we can see, CWMs only take a small portion of the network, 24.97% for region A, 20.76% for region B, and 28.00% for region C. The ratio between CWM failures and all the failures is even smaller, 12.71% for region A, 11.70% for region B, and 12.74% for region C.

The observation period covers 12 years, spanning from 1998 to 2009. It is short compared with pipe life span which can be more than 100 years as shown in Table 18.1. The majority of the pipes did not fail or just failed once during the observation period. If considering pipe segment failures, the failure events are even more sparse. Hence, the sparsity assumption holds for the proposed approximated sampling algorithm.

Failure locations are used for matching failures with pipe segments. It enables the proposed DPHBP model to work on pipe segment level for estimating failure probabilities.

Table 18.2 Pipe attributes and environmental factors

	Property and factors	Description
Pipe attributes	Protective coating	Categorical value indicating the type of coating
	Diameter	Continuous value indicating pipe diameter
	Length	Continuous value indicating pipe length
	Laid date	Laid date for pipe
	Material	Categorical value indicating the type of pipe material
Environmental factors	Soil corrosiveness Soil expansiveness Soil geology Soil map	Categorical value indicating soil property for the corresponding soil factor
	Distance to traffic intersection	Continuous value indicating the distance between pipe segment and the closest traffic intersection

As mentioned before, we focus on CWMs for comparison experiments since both physical condition assessment and proactive replacement are conducted for CWMs. For comparing the performances of different approaches, we use the first 11 years' failure records as training data and the last year's failure records as testing data. All the compared methods have the same setting for fair comparison.

18.4.2 Considered Features - The Importance of Domain Knowledge

In this section, we describe the pipe attributes and the environmental factors that we used in the experiments. As mentioned before, by considering the domain experts' knowledge, informative features can be readily identified and considered in the model. Without the support of domain knowledge, important features could be ignored by the model and false correlated features could be incorporated into the model, in which case, the model performance would be significantly reduced.

For drinking water pipe, there are five pipe attributes utilised in the experiments including protective coating, diameter, length, laid date, and material. Two types of environmental factors are considered in the experiments. One is the surrounding soil condition, and the other is the distance between pipe segment and its closest traffic intersection. These features are summarised in Table 18.2.

For pipe attributes, protective coating and material are categorical features indicating the type of coating and material. Typical protective coatings are a polyethy-

lene sleeve and tar coating. Typical materials are cast iron cement lined (CICL) and polyvinyl chloride (PVC). Diameter, length, and laid date are continuous features.

Surrounding soil condition is one of the most complex and important environmental factors for water pipe failure prediction. It directly impacts on the pipe degradation process. In the experiments, four different soil features are considered including soil corrosiveness, soil expansiveness, soil geology and soil map. They depict different perspectives of soil characteristics.

Soil corrosiveness describes the risk of pipe pitting (metal corrosion) which is essentially an electrical phenomenon and can be measured by a linear polarisation resistance test. Soil expansiveness describes the a shrinking and swelling of expansive clays in response to moisture content change. It is a phenomenon that affects clay soil and can be measured by shrink swell test. Soil geology depicts the information of rocks, e.g., sandstone and shale. A soil map represents the landscape information, e.g., fluvial, colluvial and erosional. It also includes information on the soil types that are associated with different landscapes.

Each soil factor is a categorical feature containing several distinct values. The selected local government areas are partitioned into small regions according to the distinct values of soil factors. Pipe segments falling into the same region share the same soil factor value.

A large portion of CWMs are buried underneath roads. It makes the change of road surface pressure another important environmental factor for estimating water pipe failures. It has been shown that frequent pressure changes can lead to high failure rate. One of the main sources causing road surface pressure change comes from traffic intersections due to the frequent vehicle starting and stopping. In order to measure the impact of road surface pressure change, we calculate the distance between each pipe segment and its closest traffic intersection. The obtained continuous value is regarded as a feature of the pipe segment for predicting its failure probability.

For the waste water pipes, tree root coverage percentage, soil evaporation and soil moisture are also considered based on domain experts' knowledge. A key cause of waste water pipe failures is the intrusion of tree roots. Roots have three basic functions; they anchor the plant and hold it upright, store food, and absorb water and nutrients. The extent of the tree root system is dependent on the species, the age of the tree, the nutrient availability from surrounding decaying organic matter and the physical limitations of the surrounding soil (soil depth, soil density/pore size, oxygen and moisture content). A constant soil temperature and adequate moisture availability lead to horizontal growing roots, in day soil condition tends to lead to vertical growing roots. In temperate conditions, tree root growth is most active during spring and autumn. In this work, we use tree canopy area (obtained by satellite image recognition) as the estimation of the tree root area. Figure 18.5 illustrates the relationship between tree root canopy coverage and the waste water pipe failures. Figure 18.6 demonstrates the relationship between soil moisture and waste water pipe failures.

As we can see in Figs. 18.5 and 18.6, both tree canopy coverage and soil moisture have a strong positive correlation with waste water pipe blockage. It demonstrates

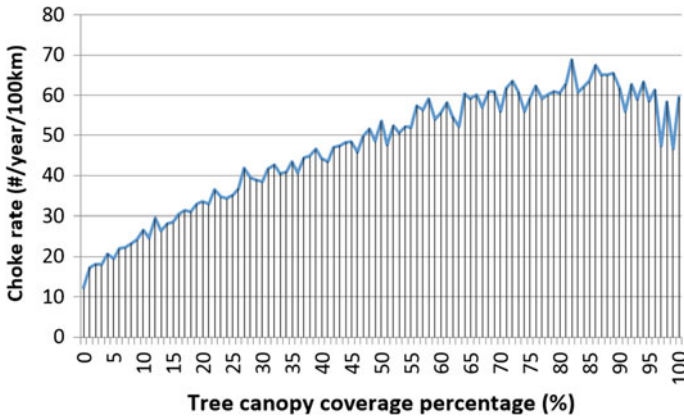


Fig. 18.5 The relationship between tree canopy coverage and waste water pipe failure (choke)

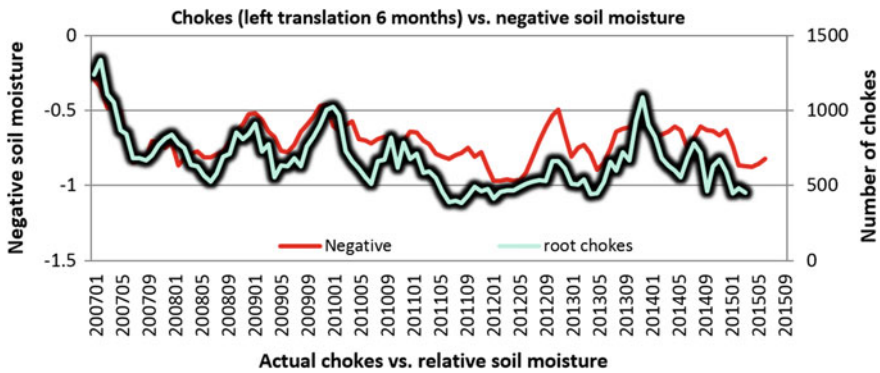


Fig. 18.6 The relationship between soil moisture and waste water pipe failure (choke)

that domain experts' knowledge can help identify important factors and later improve model performance.

18.4.3 Compared Approaches

In order to evaluate the proposed approach, four state-of-the-art methods are compared in the experiments including the Cox proportional hazard model, the Weibull model, the HBP model and a support vector machine (SVM) based ranking method. Additionally, different grouping methods are used with the HBP model as comparisons for demonstrating the advantage of the grouping scheme of the proposed approach.

The Cox proportional hazard model [3] is one of the most popular approaches for survival analysis. It is a semi-parametric approach, in which the form of the baseline hazard function can be arbitrary, and the explanatory features put impacts on the baseline hazard function via an exponential function multiplicatively. Formally, the Cox proportional hazard model can be described as:

$$h(t, z) = h_0(t)e^{b^T z}, \quad (18.8)$$

where h_0 indicates the baseline hazard function, z indicates the explanatory features of water pipe, and b is the parameter vector that can be learned from training data via a partial likelihood maximisation procedure.

For the Weibull model [2, 8], water pipe failures are modelled as a set of stochastic events governed by a time dependent stochastic process, namely the Weibull process. It can be regarded as a nonhomogeneous Poisson process whose intensity varies as time changes. The intensity function can be formally given as:

$$\lambda(t) = \alpha\beta t^{\beta-1}, \quad (18.9)$$

where t represents pipe age, α and β are parameters that need to be learned from training data. Similar to the Cox proportional hazard model, the explanatory features can also be utilised via an exponential function multiplicatively.

Analogous to the method proposed in [18], an SVM-based ranking approach is compared. This approach formulates pipe failure prediction as a ranking problem. It ranks pipes according to their failure risks without estimating their actual failure probability. It learns a real-valued ranking function H that maximises the objective function:

$$\sum_{z \in P, z' \in N} \frac{I(H(z) > H(z'))}{|P| \cdot |N|}, \quad (18.10)$$

where P and N represent the positive class dataset (failure dataset) and negative class dataset respectively. $I(\cdot)$ is the indicator function. $|P|$ and $|N|$ indicate the numbers of data points in the positive and negative class datasets respectively.

The HBP model proposed by [11] is also compared. In order to evaluate the grouping scheme of the proposed approach, three different grouping methods are integrated with the HBP model for comparison. They group pipes based on pipe attributes according to domain expert suggestions. Specifically, pipes are grouped based on material, diameter and laid year.

For fair comparison, the features described in the previous section are used for all the compared methods. For HBP and DPMHBP, the features are applied multiplicatively similar to the Cox proportional hazard model and the Weibull model. A linear kernel is used for the SVM-based ranking approach.

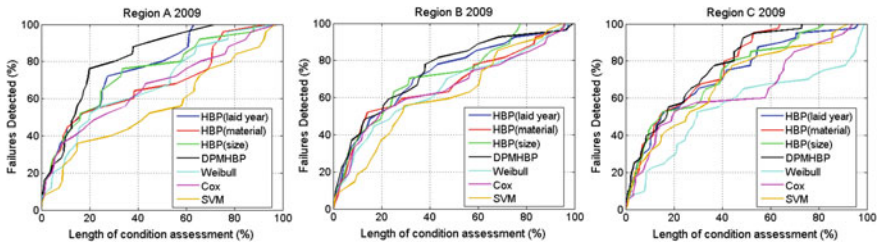


Fig. 18.7 Failure prediction results for the selected regions by different models

Table 18.3 AUC of different approaches. The second row shows the AUC when 100% of CWMs are inspected. The third row shows the AUC when 1% of CWMs are inspected

	Region A			Region B			Region C		
	DPM HBP Cox	HBP SVM	Weibull	DPM HBP Cox	HBP SVM	Weibull	DPM HBP Cox	HBP SVM	Weibull
AUC (100%)	82.67% 66.91%	77.05% 56.45%	68.44%	74.51% 65.53%	72.56% 61.90%	65.20%	78.37% 64.50%	73.54% 69.48%	55.84%
AUC (1%)	8.09‰ 4.67‰	5.64‰ 4.32‰	5.84‰	4.21‰ 2.46‰	3.60‰ 3.41‰	2.70‰	5.11‰ 2.50‰	2.48‰ 1.73‰	2.98‰

18.4.4 Prediction Results and Real Life Impact

In this section, we compare the prediction results to demonstrate the superiority of the proposed approach. As mentioned before, the historical failure data from 1998 to 2008 is used for training and the failures which occurred in 2009 are used for testing. Water pipes are ranked by different methods based on their estimated failure risks. The failure prediction results are shown in Fig. 18.7. The x -axis represents the cumulative percentage of the inspected water pipes, and the y -axis indicates the percentage of the detected pipe failures.

Additionally, we calculate AUC for measuring the performances of different approaches. The results are shown in Table 18.3. Statistical significance tests, particularly the one-sided paired t -test at 5% level of significance, are performed on AUC to evaluate the significance of the performance differences. The results are shown in Table 18.4. For Tables 18.3 and 18.4, only the results from the best groupings are shown for the HBP model.

As we can see from Fig. 18.7 and Table 18.3, the proposed DPMHBP model consistently gives the most accurate prediction for all the three regions, whereas the other methods only perform accurately for some of the regions. It demonstrates the adaptability of the proposed approach to the diversity of failure patterns. The significance test results, listed in Table 18.4, show that the proposed model significantly outperforms the other methods.

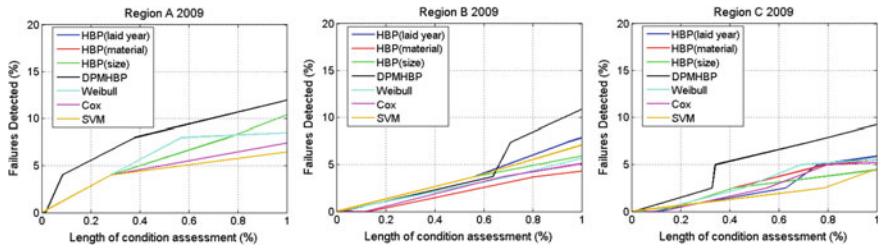


Fig. 18.8 The detection results with 1% of pipe network length inspected

Table 18.4 Statistical significance test (t-test) results for the proposed method and the others. The second row shows the results when 100% of CWMs are inspected. The third row shows the results when 1% of CWMs are inspected

	Region A		Region B		Region C	
	versus HBP	versus Weibull	versus HBP	versus Weibull	versus HBP	versus Weibull
	versus Cox	versus SVM	versus Cox	versus SVM	versus Cox	versus SVM
AUC	2.56(= 0.08)	9.37(<0.05)	3.12(= 0.05)	22.01(<0.05)	7.83(<0.05)	43.55(<0.05)
(100%)	10.58(<0.05)	18.88(<0.05)	21.17(<0.05)	30.11(<0.05)	26.08(<0.05)	15.63(<0.05)
AUC	44.29(<0.05)	40.46(<0.05)	1.26(<0.05)	4.64(<0.05)	65.90(<0.05)	53.43(<0.05)
(1%)	62.44(<0.05)	69.01(<0.05)	5.53(<0.05)	1.99(<0.05)	65.43(<0.05)	61.72(<0.05)

In addition to the comparison studies shown above, we also demonstrate the real-life impact of the proposed method by showing its improvements in its real-world application. Different from the standard performance measurement, domain experts often adopt evaluation criteria that can reflect the constraints encountered in reality. In the context of water pipe failure prediction, as mentioned before, only a small portion of the pipes can be physically inspected each year. Specifically, due to budget constraint, only 1% of the total CWMs can be inspected every year. Therefore, we show the performance curves with 1% of CWMs inspected in Fig. 18.8. AUC and significance test results are also given in Tables 18.3 and 18.4 for the situation of inspecting 1% of CWMs. As we can see, the proposed approach significantly outperforms the other methods for all the three regions. In region C, the proposed approach nearly doubles the number of detected failures compared with the second best method.

A risk map, as shown in Fig. 18.9, is another widely used method for visualising real-life impact. As illustrated in the figure, the prioritisation of pipes is coded by different colours. For instance, red lines indicate the top 10% high-risk pipes predicted by our method. Black stars in the figure denote the failures which occurred in the testing year. As we can see, many failures could be prevented and significant economic and social savings could be brought to the water utility if the proposed method were applied.

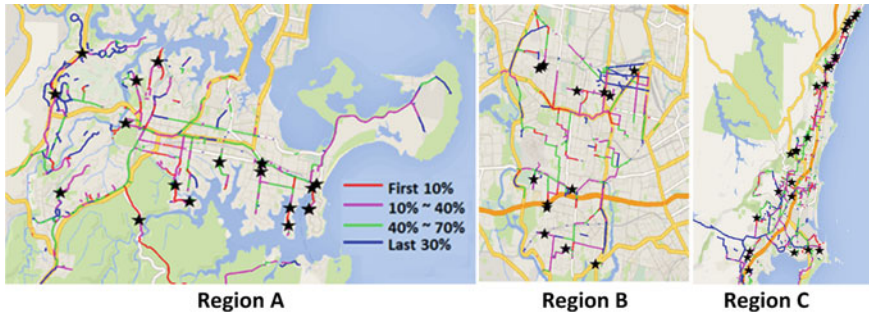


Fig. 18.9 Risk maps for the selected three regions

18.5 Conclusion

In this work, we present the Dirichlet process mixture of the hierarchical beta process model for water pipe failure prediction. The model demonstrates high adaptability to the diversity of failure patterns. Its structure and complexity can grow as the number of data points increases. It tackles the sparse failure data problem by sharing failure data through pipe grouping. An efficient Metropolis-within-Gibbs sampling algorithm is also proposed for handling large-scale datasets. The empirical studies conducted on the real water pipe data verify the superiority of the proposed approach. The domain expert knowledge also gave significant impact on the model development and the informative factor identification. It would be extremely difficult, if not impossible, to discover the key informative factors without the support of domain knowledge. Besides, the incorporation of domain experts' knowledge and experience can help enhance domain users' trust in the model as it improves their understanding of the model and makes them trust the basis of the model development.

References

1. Aldous, D.J.: Exchangeability and Related Topics. Springer, Berlin (1985)
2. Constantine, A.G.: Pipeline reliability: Stochastic Models in Engineering Technology and Management. World Scientific, Singapore (1996)
3. Cox, D.R.: Regression models and life-tables. In: Journal of the Royal Statistical Society. Series B Methodological, pp. 187–220. (1972)
4. Ferguson, T.S.: A bayesian analysis of some nonparametric problems. *Ann. Stat.* **1**, 209–230 (1973)
5. Hjort, N.L.: Nonparametric bayes estimators based on beta processes in models for life history data. *Ann. Stat.* **18**, 1259–1294 (1990)
6. Hoffman, M.D., Blei, D.M., Cook, P.R.: Content-based musical similarity computation using the hierarchical dirichlet process. In: ISMIR, pp. 349–354 (2008)
7. Huelsenbeck, J.P., Jain, S., Frost, S.W., Pond, S.L.K.: A dirichlet process model for detecting positive selection in protein-coding dna sequences. *Proc. Natl. Acad. Sci.* **103**(16), 6263–6268 (2006)

8. Ibrahim, J.G., Chen, M.H., Sinha, D.: Bayesian Survival Analysis. Wiley Online Library (2005)
9. Kettler, A., Goulter, I.: An analysis of pipe breakage in urban water distribution networks. *Can. J. Civil Eng.* **12**(2), 286–293 (1985)
10. Li, B., Zhang, B., Li, Z., Wang, Y., Chen, F., Vitanage, D.: Prioritising water pipes for condition assessment with data analytics. *OzWater* (2015)
11. Li, Z., Zhang, B., Wang, Y., Chen, F., Taib, R., Whiffin, V., Wang, Y.: Water pipe condition assessment: a hierarchical beta process approach for sparse incident data. *Mach. Learn.* **95**(1), 11–26 (2014)
12. Mavin, K.: Predicting the failure performance of individual water mains. Urban Water Research Association of Australia (114) (1996)
13. Paisley, J., Carin, L.: Nonparametric factor analysis with beta process priors. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 777–784. ACM (2009)
14. Rajani, B., Kleiner, Y.: Comprehensive review of structural deterioration of water mains: physically based models. *Urban Water* **3**(3), 151–164 (2001)
15. Shamir, U., Howard, C.: An analytic approach to scheduling pipe replacement. *Am. Water Works Assoc.* **71**(5), 248–258 (1979)
16. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. *J. Am. Stat. Assoc.* **101**(476), a (2006)
17. Thibaux, R., Jordan, M.I.: Hierarchical beta processes and the Indian buffet process. In: International Conference on Artificial Intelligence and Statistics, pp. 564–571 (2007)
18. Wang, R., Dong, W., Wang, Y., Tang, K., Yao, X.: Pipe failure prediction: a data mining method. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE), pp. 1208–1218. IEEE (2013)
19. Yakhnenko, O., Honavar, V.: Annotating images and image objects using a hierarchical dirichlet process model. In: Proceedings of the 9th International Workshop on Multimedia Data Mining: Held in Conjunction with the ACM SIGKDD 2008, pp. 1–7. ACM (2008)
20. Zhou, M., Chen, H., Ren, L., Sapiro, G., Carin, L., Paisley, J.W.: Non-parametric bayesian dictionary learning for sparse image representations. In: Advances in Neural Information Processing Systems, pp. 2295–2303 (2009)
21. Zhou, M., Yang, H., Sapiro, G., Dunson, D.B., Carin, L.: Dependent hierarchical beta process for image interpolation and denoising. In: International Conference on Artificial Intelligence and Statistics, pp. 883–891 (2011)