CHEMICAL ENGINEERING TRANSACTIONS
VOL. 33, 2013
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at: www.aidic.it/cet
Guest Editors: Enrico Zio, Piero Baraldi

ISBN 978-88-95608-24-2; **ISSN** 1974-9791
DOI: 10.3303/CET1333009

Tuning the Covariate Influence in a Gamma Process Clustering Algorithm

Edith Grall-Maës*, XuanZhou Wang, Pierre Beauseroy

Institut Charles Delaunay - CNRS UMR STMR- Université de Technologie de Troyes, Troyes, France
edith.grall@utt.fr

This paper addresses the problem of clustering stochastic deterioration processes in a Gamma process model-based framework. The process evolution is assumed to be described by realizations of Gamma processes. Complementary information is given by covariates which characterize the systems from which the realizations originate. The processes parameters depend on a partition of the covariate space. To identify the process parameters one need to cluster data taking into account the realizations and the covariates. The proposed method embeds both concerns by using an EM algorithm with side-information and a local *a posteriori* probability. The weight given to the Gamma process realizations with respect to the covariates depends on the importance of the neighborhood considered in the local *a posteriori* probability. Criteria related to the value of the parameter tuning this importance are proposed. Simulated data is used to illustrate the clustering method and to study the influence of the parameter tuning the covariate influence on the error result and on the proposed criteria.

1. Introduction

When studying deterioration processes of systems, we are eager to find a good stochastic model which fits the degradation measurements in order to do some prognostic for example, such as estimating the remaining useful life. Gamma process was successfully applied for the modelling of monotonic and gradual deterioration. It has been satisfactorily fitted to various data such as creep of concrete, fatigue crack growth, and thinning due to corrosion (Van Noortwijk 2009). In some cases the degradation process depends on some characteristics (covariates) of the monitored system. It follows that the stochastic law that governs the degradation process changes with those covariates. Most of the time, the only available information about the process is a database which contains realizations of degradation processes, and their corresponding covariates. Thus one has to estimate the Gamma process parameters while taking into account their dependency with the covariates.

This paper addresses the problem of clustering degradation measurements on systems with covariates, in a Gamma process model-based framework. The aim is to determine jointly the classes of the different process realizations of a given sample and the parameters of the Gamma process classes.

In comparison with classical data clustering, the process classification problem involves realizations that are composed of several observations. Thus the knowledge that all observations of a realization originate from the same process has to be considered as side information. On the other hand, the knowledge of covariate can be considered as spatial data. Thus the clustering method has to consider both the attribute data proximity and the spatial data proximity as in (Ambroise and Govaert, 1998; Hu and Sung, 2006). So, the considered problem is a problem of spatial clustering with side information.

The proposed method is based on the EM algorithm which is a classical approach in cluster analysis with mixture models (McLachlan and Basford 1988), and more specifically on the EM algorithm with side information proposed in (Shental et al., 2003). The spatial proximity is introduced by the consideration of the neighborhood whose influence can be tuned thanks to a parameter. The influence of this parameter is studied and criteria related to the value of this parameter are proposed.

The paper is organized as follows. A general formulation of the problem and notations are introduced in section 2. Section 3 describes firstly the proposed solution for clustering the process realizations with

Please cite this article as: Grall-Maes E., Wang X., Beauseroy P., 2013, Tuning the covariate influence in a gamma process clustering algorithm, Chemical Engineering Transactions, 33, 49-54 DOI: 10.3303/CET1333009

constraints of covariate and secondly criteria characterizing a clustering result. Numerical evaluations and discussions are reported in section 4 and we conclude the paper in section 5.

2. Problem formulation

The data we consider originates from N paths describing degradation process realizations. For each path n $(n=1\ldots N)$, the data is composed of the attribute data $X_n \in \Omega_{\mathbf{X}}$ and the covariate $Y_n \in \Omega_{\mathbf{Y}}$ characterizing the system from which the path originates.

The attribute set $\{X_n\}_{n=1\ldots N}$ is assumed to be a sample composed of K subpopulations which are all homogeneous Gamma process models. Each model is characterized by a parameter set $\theta_k = (a_k, b_k)$ where $a_k > 0$ and $b_k > 0$ are respectively defined as the shape and the scale parameters of the Gamma process. The latent cluster labels directly related to the parameters θ_k are described by $z = \{z_n\}_{n=1\ldots N}$ where $z_n = k$ means that the n^{th} realization originates from the k^{th} cluster.

The attribute data X_n is a Gamma process realization meaning that X_n is a time series: $X_n = \{(t_i^n, x(t_i^n))\}_{i=1\ldots|X_n|}$ where $|X_n|$ is the number of values in X_n and t_i^n is the instant number i of realization number n. The increments, given by $x(t_i^n) - x(t_{i-1}^n)$ with $t_0 = 0$ and $x(0) = 0$, are independent. The density distribution of the increments, which depends on the time and on the parameter θ_k, is given by

$$f_k\left(x\left(t_i^n\right) - x\left(t_{i-1}^n\right) \mid t_i^n, t_{i-1}^n, \theta_k\right) = \Gamma(a_k(t_i^n - t_{i-1}^n), b_k) \tag{1}$$

In the following, for simplicity we will use the notation $f_k(\Delta x_i^n \mid \theta_k)$.

Besides, θ_k is supposed to depend on covariates (also called spatial data in the following), and the formulation of that relation can be given by $\theta_k = \Theta(Y)$ with Θ an unknown function $\Omega_Y \to \{\theta_k\}_{k=1}^K$, which defines a partition on $\Omega_{\mathbf{Y}}$. In addition, we use the same regularity hypothesis as in most of the work of spatial clustering (Ambroise and Govaert, 1998): the data are considered to evolve slowly in geographic space $\Omega_{\mathbf{Y}}$, which means that two samples that are neighbors in $\Omega_{\mathbf{Y}}$ are likely to belong to the same sub-population k.

An example with 2 clusters and a two-dimension spatial data is given in figure 1 in section 4.

The objective is to find out the unknown cluster labels $\{z_n\}_{n=1}^N$ and consequently the Gamma process parameters a_k and b_k such that paths in the same cluster originate from a process model with the same parameters and that clusters are coherent from a geographic point of view. The performance of a clustering result described by z and a parameter set θ, without consideration of covariate influence, can be measured using the log-likelihood given by

$$l(\mathbf{z}, \theta) = \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{i=1}^{|X_n|} \ln f_{z_n}(\Delta x_i^n \mid \theta) \tag{2}$$

3. Clustering algorithm of Gamma process with covariate influence

3.1 Proposed algorithm

The proposed approach is based on the clustering method using Gaussian mixture models and the expectation-maximization (EM) algorithm (Celeux and Govaert, 1992). Besides, the side information is considered according to (Shental et al., 2003).

The EM algorithm is an iterative method that produces a set of parameters that locally maximizes the log-likelihood of a given sample, starting from a arbitrary set of parameters. It is often used to estimate the unknown parameters of a mixture. In that case the E step consists in calculating an estimation of the *a posteriori* probability for each observation and the expected log-likelihood. The M step consists in computing the parameters that maximize the expected log-likelihood found on the E step. When a hard partition is sought, it is suggested in (Celeux and Govaert, 1995) to add an intermediate classification step between the E and M steps.

In (Shental et al., 2003), an EM algorithm is introduced for computing Gaussian mixture models taking into consideration equivalent constraints between data points which determine whether points were generated by the same source. It is shown that the E step consists in computing the posterior probability by using the product of the conditional probabilities of all points into a chunklet. Using the notation introduced in section 2, the posterior probability $c_{nk}^{(m)}$ at iteration m that the path n belongs to class k, given X_n and the parameter $\theta^{(m-1)}$ writes according to

$$c_{nk}^{(m)} = p\left(z_n = k|X_n, \theta^{(m-1)}\right) = \frac{p_k^{(m-1)} \prod_{i=1}^{|X_n|} f_k\left(\Delta x_i^n | \theta^{(m-1)}\right)}{\sum_{r=1}^{K} p_r^{(m-1)} \prod_{i=1}^{|X_n|} f_r\left(\Delta x_i^n | \theta^{(m-1)}\right)} \quad (3)$$

For the problem of Gamma process clustering without covariate influence, the determination of the partition and the parameters is based on the mixture models with the constraint that observations in a same path belong to a same class. Since the aim is to determine a hard classification, an intermediate classification step is added between the E and M steps, as in (Celeux and Govaert, 1995). It follows that the algorithm is the following one.

- Initialize the parameter set $\theta^{(0)}$
- Repeat until $l\left(\mathbf{z}^{(m)}, \theta^{(m)}\right) - l\left(\mathbf{z}^{(m-1)}, \theta^{(m-1)}\right) < \varepsilon$
  - Compute the $c_{nk}^{(m)}$ using relation (3)
  - Determine $\mathbf{z}^{(m)}$: choose $z_n^{(m)} = k$ corresponding to the largest value $c_{nk}^{(m)}$
  - Determine $\theta^{(m)}$ that maximizes $l(\mathbf{z}^{(m)}, \theta^{(m)})$
  - Compute the new value of $p_k^{(m)}$: $p_k^{(m)} = \frac{1}{N} \sum_{n=1}^{N} (z_n^{(m)} = k)$

For the problem of statistical process clustering with covariate influence, the idea of spatial proximity is added. To that end, we suggest using a local *a posteriori* probability. Instead of considering a given path, we consider this path and its neighbors (all paths with covariate which are close to the covariate of the given path). The neighborhood of a path n is defined by $V_\alpha(n)$: the set of path numbers contained in the neighborhood of n according to Y_n whose influence can be tuned thanks to a parameter α .
Then the algorithm for clustering with covariate influence is the same as above, except that $c_{nk}^{(m)}$ given by relation (3) is replaced by $\tilde{c}_{nk}^{(m)}$ which provides a local *a posteriori* probability. $\tilde{c}_{nk}^{(m)}$ is defined by:

$$\begin{aligned} \tilde{c}_{nk}^{(m)} &= p(z_n = k \mid X, V_\alpha(n), \theta^{(m-1)}) \\ &= \frac{p_k^{(m-1)} \prod_{\tilde{n} \in V_\alpha(n)} \prod_{i=1}^{|X_{\tilde{n}}|} f_k(\Delta x_i^{\tilde{n}} \mid \theta^{(m-1)})}{\sum_{r=1}^{K} p_r^{(m-1)} \prod_{\tilde{n} \in V_\alpha(n)} \prod_{i=1}^{|X_{\tilde{n}}|} f_r(\Delta x_i^{\tilde{n}} \mid \theta^{(m-1)})} \end{aligned} \quad (4)$$

3.2 Tuning the covariate influence

The influence of the covariate is tuned thanks to a parameter α which controls the number of paths contained in the neighborhood of each path for the local *a posteriori* probability computation. The neighborhood of a path n is a local region around its covariate Y_n , defined by $V_\alpha(n)$:

$$V_\alpha(n) = \left\{\tilde{n} \mid Y_{\tilde{n}} - Y_n \le R(\alpha) \ \ \forall \tilde{n} = 1 \ldots N\right\} \ \ \forall n = 1 \ldots N \quad (5)$$

where $R(\alpha)$ is a bound on the radius which depends on α .
Two approaches have been considered for tuning the covariate influence. The first one consists in choosing a fixed value for the radius then $R(\alpha) = \alpha$. The second one consists in choosing the number of increments values falling in the local region. Then $R(\alpha)$ is chosen such that at least a given number α of increments values belong to $V_\alpha(n)$, with the constraint that all the increments values of a same path belong to $V_\alpha(n)$.
It is expected than when the influence of the covariates increases, the homogeneity of the class labels in the covariate space increases and the likelihood decreases.
For a given clustering result described by z , a spatial homogeneity measure $G(z)$ can be defined by

$$G(z) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \delta_{z_i, z_j} \kappa_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} \kappa_{ij}} \quad \text{with} \quad \kappa_{ij} = \begin{cases} exp\left(-\left(\frac{Y_i - Y_j}{\rho}\right)^2\right) & \text{if } \left|Y_i - Y_j\right| < \rho \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where $\delta_{z_i,z_j}=1$ if $z_i=z_j$ and 0 otherwise. κ_{ij} is a measure of proximity of Y_i and Y_j. Then the value of G is between 0 and 1, and increases with the number of pairs sharing the same class label, and close, in a disk of radius ρ chosen arbitrarily equal to 2.

For the log likelihood, it is suggested to use a normalised log-likelihood introduced in (Wang et al., 2012). The maximum log-likelihood value is obtained when the clustering is done without covariate influence, and noted l_K. The minimum value is obtained when the covariate influence tends to infinity, which corresponds to assigning all paths to a single class. Thus we note it l_1. Consequently the normalised log-likelihood $l'(z)$ for a given clustering result described by z, is defined by

$$l'(z)=\frac{l(z)-l_1}{l_K-l_1}$$

4. Application

The performance of the proposed approach has been evaluated by using simulated data, supposed to belong to $K=2$ Gamma processes. For the experiment we have chosen $(a_1,b_1)=(18,1.5)$ or $(a_1,b_1)=(24.5,1.75)$ and $(a_2,b_2)=(12.5,1.25)$. The data for the covariate have been chosen in dimension 2, with an uniform distribution in $[0,10]^2$. The chosen boundary between the two classes is given by: $y_1=-(y_2/4)^3+8.25$, where y_1 and y_2 are the covariates. The number of realizations is equal to 100 and each realization contains 4 observations. An example of different paths and covariates for both classes is given on figure 1.

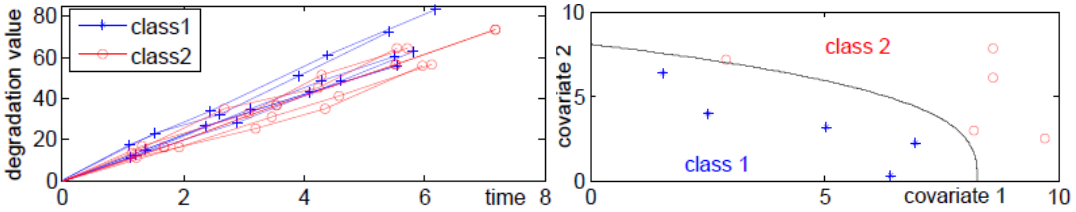

Figure 1: An example of different paths (left) and covariates (right)

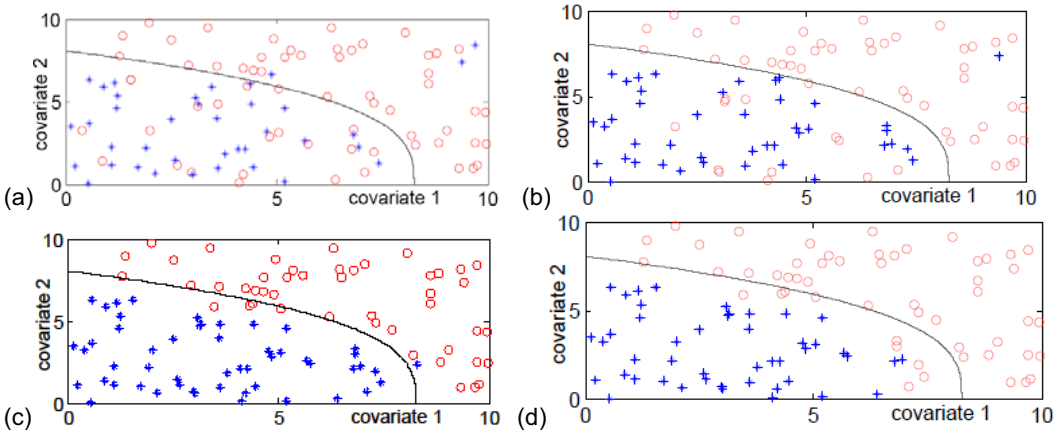

Figure 2: An example of classification results (+: class 1, o: class 2, line: theoretical boundary) without covariate influence (a) and with different values of radius α : 0.5(b), 1.5(c) and 2.5 (d)

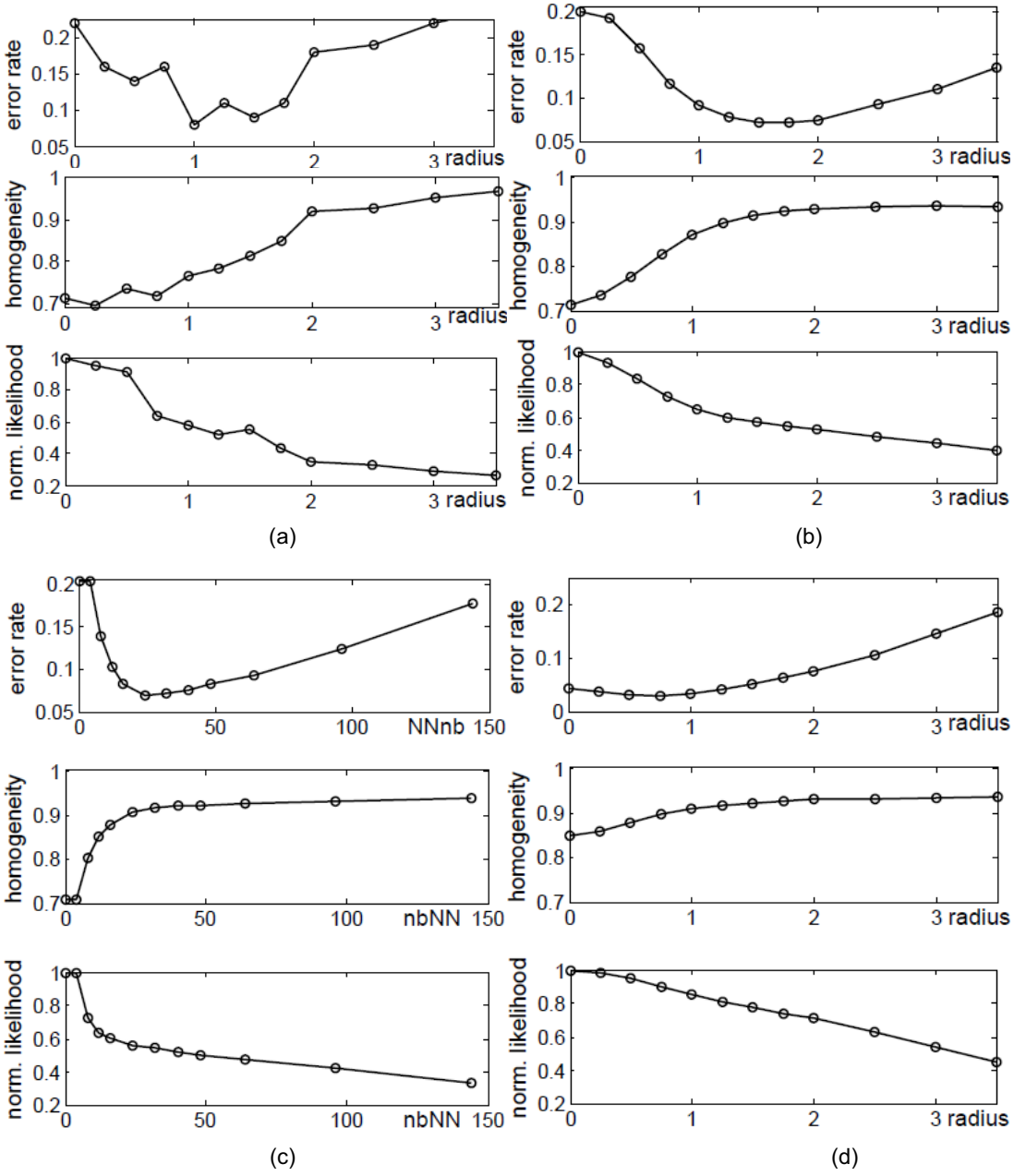

Figure 3: Error rate, homogeneity and normalized likelihood, in relation with the parameter α in four different cases : (a) $(a_1, b_1) = (18, 1.5)$, parameter: radius, 1 simulation (b) $(a_1, b_1) = (18, 1.5)$, parameter: radius, mean of 200 simulations (c) $(a_1, b_1) = (18, 1.5)$, parameter: nearest neighbors, mean of 200 simulations (d) $(a_1, b_1) = (24.5, 1.75)$, parameter: radius, mean of 200 simulations.

The figure 2 shows an example of different classification results obtained without covariate influence $(\alpha = 0)$ and with covariate influence tuned according to the value of the radius α. It shows that when α is too small, some isolate data are misclassified. Increasing α allows to obtain a classification result more homogeneous. However, when it is too large, some errors arise near the boundary. When α grows, the classification result tends to the classification of all observations into a unique class.

The figure 3a gives the value of three criteria: the error rate, the homogeneity and the normalized log-likelihood, in relation with the value of the radius α in the case of $(a_1, b_1) = (18, 1.5)$ for an example. The figure 3b gives the mean values of these three criteria estimated with 200 simulations. The minimum error rate is 7.1% and is obtained for $\alpha = 1.75$ whereas it is equal to 20% without influence of covariates. The

figure 3c gives the mean values of these three criteria estimated with 200 simulations, in relation with the number of neighbors. The minimum error rate is 7.1% and is obtained for $\alpha = 32$. It is the same value than using the radius as the parameter for selecting the neighbors.
The figure 3d gives the mean values of the three criteria, in relation with the radius, in the case of $(a_1, b_1) = (24.5, 1.75)$ which means that the two processes are more different that in the case of the figure 3b. The minimum error rate is 2.9% and is obtained for $\alpha = 0.75$ whereas it is equal to 4.4% without influence of covariates.
In the case of real data, the error rate is not known then the value of α has to be selected according to an efficient criterion. The curves on figure 3 show that the homogeneity and the normalized log-likelihood provide good criteria for that. In particular, a change in the slope of the curves of the homogeneity and of the normalized log-likelihood appears around the optimal value of the parameter α . A threshold on the homogeneity value could also be used. However to find the optimal value, the criterion has to take into account that the curves are not smooth.

5. Conclusion

In this paper the problem of clustering stochastic deterioration processes in a Gamma process model-based framework has been tackled. The data set is composed of realizations describing the process evolution in time and covariates which describe the systems from which the realizations originate. Thus this is a problem of spatial clustering with side information.
The proposed procedure is an iterative algorithm including a step of estimation of *a posteriori* probabilities and a step of determination of Gamma process parameters. The influence of the covariates is achieved by a local *a posteriori* probability and the importance of the covariates proximity is tuned thanks to a parameter. Empirical results lead to believe that this algorithm converges. A formal study of the convergence has to be developed in the future.
Two parameters for controlling the influence of covariates have been considered: the size of the local neighborhood, and the number of neighbors. Simulation results have shown that they lead to equivalent results.
Results on simulated data show that the probability of error can be significantly reduced when the covariates are taken into account. An homogeneity criterion and a normalized log-likelihood criterion provide values that are directly related to the parameter that controls the influence of covariates. Thus they can be used for selecting an optimal value of the parameter. This will be addressed in future work.

References

Ambroise, C, & Govaert, G, 1998, Convergence of an EM-type algorithm for spatial clustering, *Pattern Recognition Letters*, 19, pp. 919-927

Celeux, G & Govaert, G, 1992, A classification EM algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3), pp. 315-332

Celeux, G & Govaert G, 1995, Gaussian parcimonious clustering models. *Pattern Recognition*, 28, pp. 781-793

Hu, T & Sung, S, 2006, A hybrid EM approach to spatial clustering. *Computational Statistics & Data Analysis*, 50, pp. 1188-1205

McLachlan G.J., Basford K.E., 1988, Mixture models. inference and applications to clustering. Statistics: Textbooks and Monographs, New York: Dekker

Shental, N, Bar-Hillel, A, Hertz, T, & Weinshall, D, 2003, Computing gaussian mixture models with EM using side-information. *Proc. of the 20th International Conference on Machine Learning*.

Van Noortwijk, JM, 2009, A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety*, 94(1), pp. 2-21

Wang, XZ, Grall-Maes, E & Beauseroy, P, 2012, A normalized criterion of spatial clustering in model-based framework, *Proc. of the 11th International Conference on Machine Learning and Applications (ICMLA),* pp. 542 – 547, 12-15 Dec.