Leading-Edge Asset Management

The impact of pipe segment length on break predictions in water distribution systems

Dr M. Poulton*, Y. Le Gat & Dr B. Brémond*****

* Cemagref, 50 avenue de Verdun 33612 CESTAS France, matthew.poulton@cemagref.fr

** Cemagref, 50 avenue de Verdun 33612 CESTAS France , yves.legat@cemagref.fr

*** Cemagref, 50 avenue de Verdun 33612 CESTAS France, bernard.bremond@cemagref.fr

Keywords Water distribution systems; break prediction; database management; pipe segment length

Abstract Break prediction models are important tools in asset management and rehabilitation planning of water distribution systems. To some extent though, they may depend on the configuration of the network as defined in the company database. This paper examines the effect of the length of pipe segments on break predictions and proposes methods for eliminating short sections of pipe sandwiched between longer pipes and for concatenating segments of the same (or similar) nature. A study was conducted on a large French network using a set of rules and length threshold values to prepare several different data files. These data files, containing the pipes' attribute information and associated break histories were used with the statistical model, LEYP, to make break predictions for each segment. The results were compared using a number of indicators. Ultimately, there would appear to be little benefit in concatenation as the model is likely to become less sensitive and less able to distinguish between risk factors.

Introduction

Water companies and their operators must manage their assets effectively and ensure the supply of drinking water to their customers. Leakage and breaks disrupt this service and the local environment; they remain one of the major factors for dictating rehabilitation of pipelines. Consequently, this has led to the development of break prediction models for defining rehabilitation strategies in drinking water networks: Clark *et al* (1982). The model, PHM: Eisenbeis (1994), Eisenbeis *et al* (2002) was included in the European Commission-funded project "Computer Aided Rehabilitation of Water Networks: CARE-W (2003).

The methods developed for predicting pipe breaks require the existence of a database that describes the network and an inventory of breaks recorded over several years: LeGat and Eisenbeis (2000). The network is subdivided into pipe segments. The segment is the elementary object of the study. The principal of the calculations is to predict for each segment, the number of breaks likely during a given time period. For each segment, the following information is necessary (or desirable):

- Asset data – diameter, material, length, year laid, (joint type…)
- Intervention data – identification of pipe concerned, date of intervention, type of incident, (reason for incident…)
- Environmental data – (soil type, surface type, traffic level, water pressure…)

A group of pipes of the same nature laid in the same year in an identical environment should display identical ageing behaviour and can from this point-of-view, be considered as a single segment. Often, segments registered in a geographic information system result from a splitting of a longer entity by the operator for functional purposes – e.g. to include hydraulic equipment such as valves and repair sleeves. Consequently, the number of segments is increased so as to be represented more precisely on the network plans. Thus it is common for these groups of homogenous pipes to be divided into many segments, often separated by a short segment of a different nature. These segments we call "sandwich pipes".

The object of this study is to determine the influence of this artificial splitting of pipe segments on the precision of break predictions. It has been carried out in collaboration with Véolia Water on French networks. The first part describes the method of analysis and the rules of concatenation pertinent to obtaining groups of pipe segments using different minimum lengths and the method for comparing the corresponding statistical analysis results. Then results are presented from a network originally comprising over 200000 pipe segments. Finally conclusions and proposed rules are presented.

Method

Obtaining data files for the study

Case of sandwich pipe segments

A sandwich pipe is a segment installed to support hydraulic equipment or to make a repair. It artificially separates two segments of the same nature and is itself, a different material. Only two adjacent pipes can be present – one at each node. In this study, the maximum length for a sandwich pipe is 10m.

Thus the first stage is to eliminate the sandwich pipes. **Figure 1** illustrates this procedure – three segments of the same material, diameter and date laid are separated by two sandwich segments. These are replaced by a single segment whose length is the total length of the original segments – including the sandwich segments.

Figure 1 "Sandwich" pipes

Concatenation procedure

Once the sandwich pipes have been eliminated, a concatenation procedure is implemented to obtain new segments at least as long as a determined threshold length. Two adjacent pipes can be concatenated if the following rules are respected:

- One of the segments has a length less than the predetermined threshold
- The segments are of the same material
- The segments have the same diameter
- The segments were laid within five years of each other
- The segments don't need to be in the same road

At the same time, the breaks attributed to each pipe are associated to the new concatenated entity. "Orphan pipes" are created when segment can not be concatenated to the threshold length.

Calculating the break predictions

The calculations are performed using the statistical model LEYP (Linearly Extended Yule Process): Le Gat (2007). Repeated breaks suffered by a pipe in a water distribution system under pressure are considered as events of the same nature occurring at random instants:

$$T_j, \quad j \in \{0,1,...,\infty\}$$

By convention, the instant T_0 is non random and fixed at the date the pipe was laid ($T_0 = t_0 = 0$). The LEYP model is governed by an intensity function which is supposed to depend on:

- The age, t of the process (i.e. the age of the segment considered)
- The number of previous events (realisation of *N(t)*)
- The vector of covariates, **Z**

The analytical form of the intensity function combines the product of

- The influence of previous events in a form derived from the Yule process (linear extension of this model)
- The influence of age in the form of the Weibull model (power of time)
- The influence of the covariates represented as in the Cox proportional hazards model

Thus: $\lambda(t;\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left(dN(t) \middle| N(t-) = j, \mathbf{Z}\right)/dt = (1+\alpha j)(\delta-1)t^{\delta}\exp\left(\mathbf{Z}^T\boldsymbol{\beta}\right)$

The expectation operator, $\mathrm{E}_{\boldsymbol{\theta}}$ relates to the LEYP parametric probability measure with parameter:

$$\boldsymbol{\theta}' = (\alpha \quad \delta \quad \boldsymbol{\beta})$$

The calculation of the break predictions requires two phases:

- A calibration during a period of observation during which breaks were recorded on the network. The model is calibrated (estimation of $\boldsymbol{\theta}$) by a method to maximise the likelihood function of breaks observed during the period of observation.
- The calculation of the predictions is carried out for a period subsequent to that used in the calibration phase. The result is, for each segment, a number of breaks and break rate along with intervals of confidence for the total number of breaks.

In the present study, the period of observation includes the period of calibration and the period of predictions. This allows results to be compared with actual breaks observed in what is called the period of validation.

Method for comparing results

The criteria for comparing results are:

- The capacity of identifying the segments the most at risk (ranking predicted number of breaks or predicted break rate)
- The accuracy of the predictions (total number of breaks predicted during a validation period)

Capacity of identifying the segments the most at risk

The pipe segments are sorted by descending break rate. To judge the quality of the criteria, the study proposed some indicators:

- **Al** – Area under the curve of predicted performance (**figure 2**). The abscissa plots the relative rank of the segments sorted by predicted break rate. The ordinate plots the percentage of cumulative breaks actually observed.

  The equation of this algebraic curve is:

$$F_L(r) = \frac{\sum_{i=1}^{r} l_i}{\sum_{i=1}^{n} l_i}, \qquad F_B(r) = \frac{\sum_{i=1}^{r} b_i}{\sum_{i=1}^{n} b_i},$$

  Where, $r \in \{1,..,n\}$ is the rank of the pipes sorted by decreasing predicted break rate, l_i is the length of pipe I, b_i is the observed break number on pipe I within the validation period

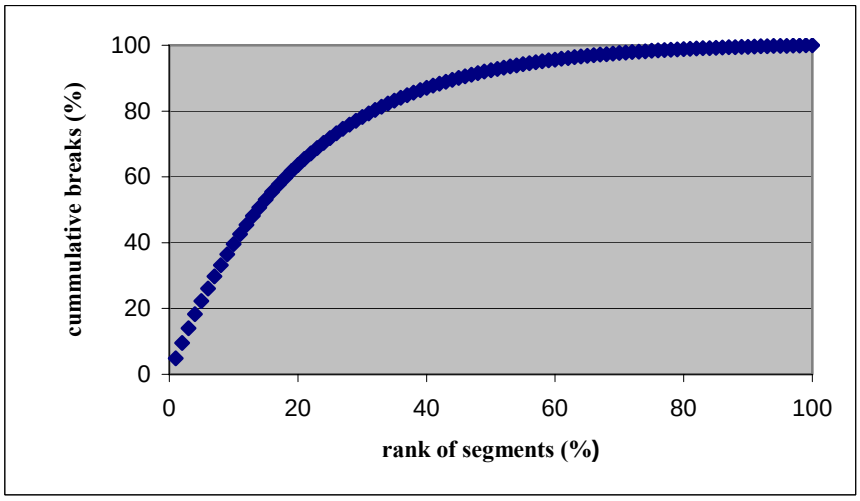

Figure 2 Example of prediction performance curve

- **CLx,** – percentage of breaks really observed on the top x% of length of pipes sorted by decreasing predicted break rate, in other words, rehabilitating top x% of the pipes should avoid Clx % of the breaks.

$$CL_x = F_B(r) \quad \text{with } r \text{ such that } F_l(r) = x$$

The accuracy of the predictions

This indicator calculated over the period of validation allows a direct comparison between actual and real break numbers.

Study data

Raw data

The calculations were performed on data supplied by Véolia Water. The study was limited essentially to grey cast iron pipes that account for most of the recorded breaks:

- 79536 pipes
- 3736 km

The distribution of segment length in the original database is shown in **figure 3**. 13% of the pipe segments are less than 5m and 38% less than 10m.

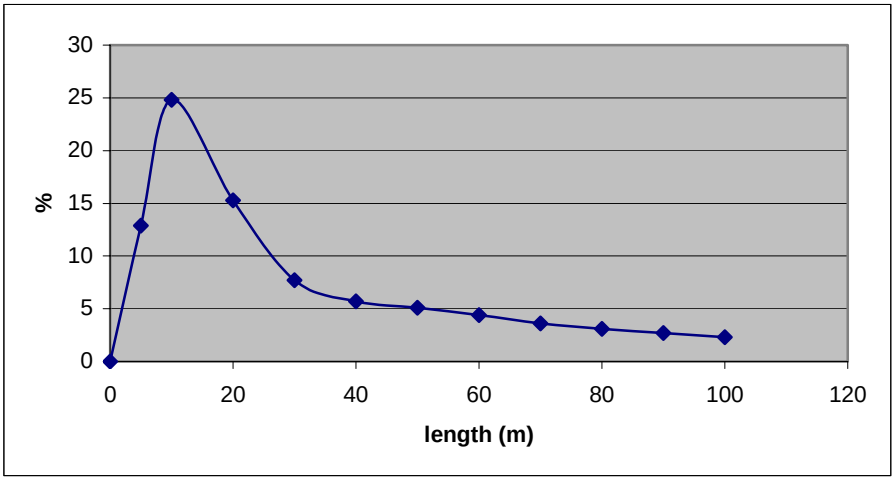

Figure 3 Distribution of pipe segment lengths

Elimination of sandwich pipes

The maximum length for a sandwich pipe was fixed at 10m. This value resulted from analysis of the number of sandwich pipes as a function of an imposed limit. **Figure 4** shows that 94% of sandwich pipes have a length less than 10m.
As a result of the "sandwiching" process, the number of cast iron pipes decreased from 79536 to 78941.

Concatenation procedure

Following the concatenation procedure explained in this paper, separate data files were created for different threshold lengths between 5m and 100m. **Table 1** summarises the characteristics of these data files. A certain number of segments with a length less than the threshold that could not be concatenated under the rules applied (the orphans) were kept. The number of orphans increases with the increase in threshold length.

The distributions of pipe number and pipe lengths are shown in **figure 5** and **figure 6** for the different length thresholds (raw is for original data and 0 for sandwich pipes removed)

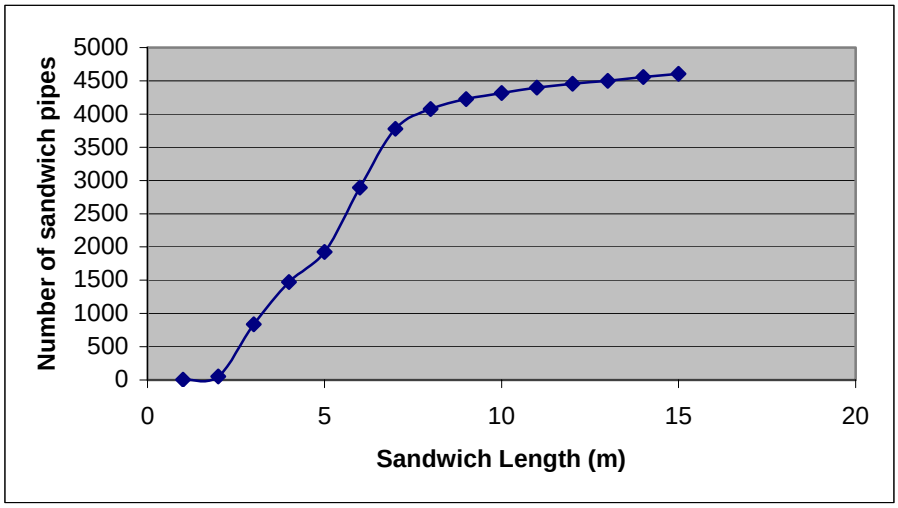

Figure 4 Distribution function of sandwich length

Table 1 Characteristics of data files

Length threshold (m)	Cast iron segments	Sandwich pipes	Orphan pipes	Maximum number of pipes joined
0 (original data)	79536	455	0	
0 (with sandwich pipes removed)	78941	0	0	19
5	68946	0	5785	7
10	57978	0	14595	7
20	51118	0	20637	10
50	45754	0	28040	15
100	41289	0	35274	19

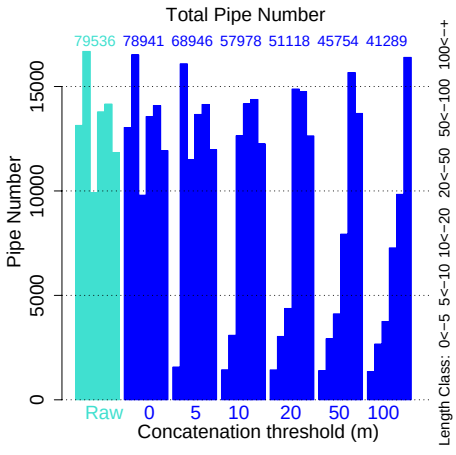

Figure 5 Pipe number distribution

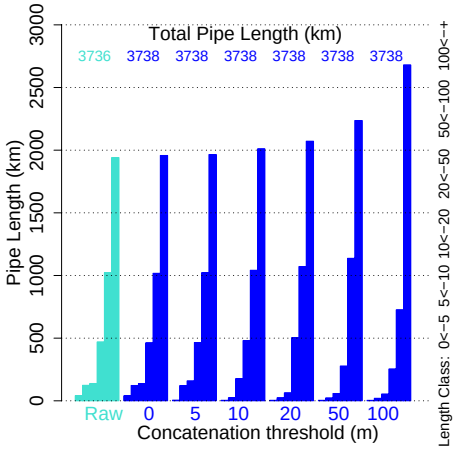

Figure 6 Pipe length distribution

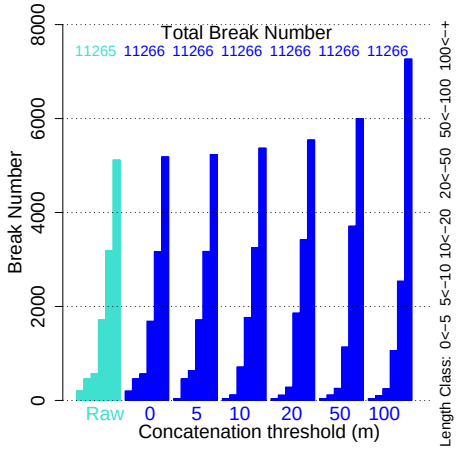

Figure 7 Pipe break distribution

During the observation period the total number of breaks was 11266. This number is independent of the concatenation procedure. However, the concatenation process re-allocates the breaks on new segments. **Figure 7** shows the distribution of break number for the seven different length thresholds.

Calculations and results

The calibration period used was 9 years (01/01/1995 – 31/12/2003). The total number of breaks during this period was 8228. The validation period used was 2 years (01/01/2004 – 31/12/2005). The total number of breaks during this period was 1997.

The impact of pipe concatenation is first assessed with respect to the estimates of the parameter values. For this purpose the parameters of interest are:

- α , which controls the tendency of breaks to occur repeatedly on the same pipes;
- β_0 , the model intercept which characterizes the baseline failure rate, related to pipe of length 1 m, diameter ≥ 300 mm and laid out between 1946 and 1964;
- β_1, the regression coefficient of the natural logarithm of the length, *i.e.* the power of the length the failure rate is proportional to;
- β_2 , the regression coefficient of the diameter category 40-60mm;
- β_3 , the regression coefficient of the diameter category 70-80mm;
- β_4 , the regression coefficient of the diameter category 90-100mm;
- β_5 , the regression coefficient of the diameter category 110-150mm;
- β_6 , the regression coefficient of the diameter category 160-250mm;
- β_7 , the regression coefficient of the construction period 1850-1929 (pit cast iron);
- β_8 , the regression coefficient of the construction period 1930-1945 (old spun cast iron).

The ageing parameter, δ , does not show in this study any evidence of being $\neq 1$; while there is no manifest ageing, the main deterioration factor of the pipe reliability seems to ground in the repetition of past breaks.

Table 2 LEYP model parameters from different simulations

Parameter	Raw Data	Concatenation Threshold (m)					
		0	5	10	20	50	100
α	2.112	2.106	2.092	2.029	1.966	1.901	1.813
β_0	-7.329	-7.327	-7.268	-7.131	-7.038	-6.889	-6.773
β_1	0.521	0.521	0.511	0.487	0.474	0.450	0.435
β_2	0.619	0.619	0.610	0.594	0.580	0.555	0.531
β_3	0.494	0.494	0.487	0.478	0.468	0.448	0.431
β_4	0.554	0.554	0.548	0.538	0.529	0.513	0.501
β_5	0.407	0.407	0.404	0.399	0.396	0.388	0.384
β_6	0.298	0.298	0.295	0.291	0.291	0.284	0.280
β_7	-0.381	-0.382	-0.383	-0.386	-0.389	-0.390	-0.386
β_8	-0.608	-0.609	-0.606	-0.600	-0.593	-0.591	-0.580

The consequences of concatenation on the model parameters are shown in table 2. The four following tendencies can be noticed:

- Except a slight effect on the couple (α, β_0), the sole removal of "sandwich" pipe segments does not significantly affect the model.
- The main impact of concatenation seems to decrease α, and increase β_0 accordingly, thus transferring break risk explanation from the past breaks to the baseline failure rate.
- The model parameterised with raw data has a β_1 close to 0.5; the failure risk varies as the square root of the pipe length. The concatenation makes β_1 decrease significantly, and makes the length factor consequently lose some explanation power.
- Another interesting effect is the decrease in the covariate modulation, i.e. the decrease in the deviation between regression coefficients related to the modalities of a given qualitative factor; this holds for both diameter category (β_2, β_3, β_4, β_5, β_6) and construction period (β_7, β_8).

In a nutshell, one can conclude that the general consequence of concatenation lies in a decrease in the models ability to discriminate the failure risk and there is hence a fear that the model tends towards "averaging" of the predictions.

Concerning the predictive efficiency of the model, the results are evaluated using the five parameters:

- Area under the performance curve Al
- CL0.5
- CL1
- CL5
- Total number of observed/predicted breaks

In figures 8, 9 and 10, CLx are presented multiplied by the actual number of breaks in the validation period (1997); the number of breaks one can expect to avoid by randomly replacing 100x% of the total length is 1997x.

Table 3 Performance indicators from different simulations

Threshold Length (m)	Area under performance curve Al	CL0.5	CL1	CL5	Breaks observed	Breaks predicted
Original data	0.646	0.0225	0.0396	0.150	1997	1873
0	0.645	0.0225	0.0391	0.152	1997	1874
5	0.645	0.0235	0.0381	0.153	1997	1875
10	0.647	0.0210	0.0386	0.154	1997	1877
20	0.650	0.0215	0.0421	0.152	1997	1880
50	0.653	0.0200	0.0416	0.149	1997	1884
100	0.655	0.0185	0.0381	0.149	1997	1890

Table 3 shows the consequences of concatenation on the predictive performance indicators of the model. The four following tendencies can be noticed:

- The predicted breaks number tends to increase with the concatenation process, underestimating slightly less at 100m threshold (5%) than for the raw data (6 %)
- The area under the performance curve also tends to slightly increase with concatenation
- This performance gain is however not uniform since it mainly concerns pipes that are not in the top of the list of most at-risk; except for 5 m threshold, CL0.5 (see **figure 8**) tends to decrease whereas an increase is observed in CL1 (**figure 9**) at 20 and 50 m thresholds and in CL5 (**figure 10**) until 10m.
- The sandwich removal has very little effect.

One can conclude that there is little interest to concatenate since usual practical yearly replacement rate is less than 1% of the total network length.

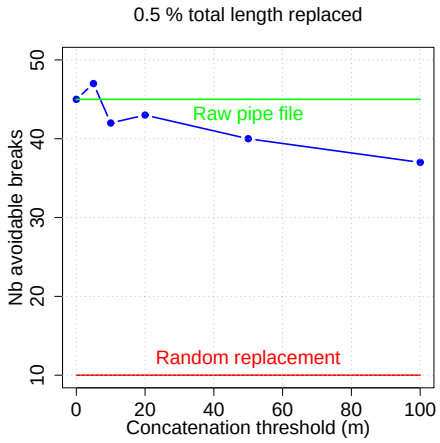

Figure 8 Impact of concatenation on breaks avoidable by 0.5% replacement

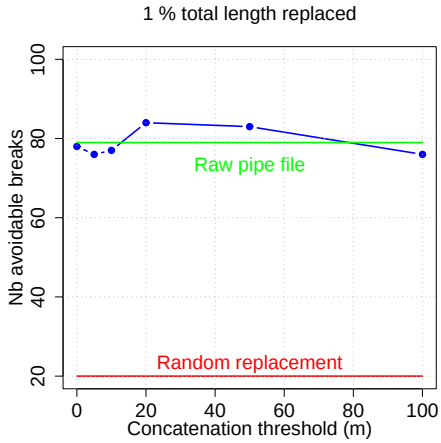

Figure 9 Impact of concatenation on breaks avoidable by 1% replacement

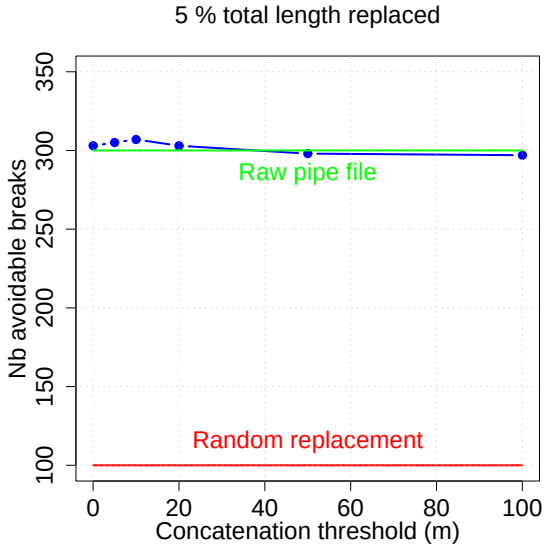

Figure 9 Impact of concatenation on breaks avoidable by 5% replacement

Conclusions

- This paper presented the theoretical interest of concatenation of small segments of water distribution networks, derived from a GIS database, in order to improve the quality of break prediction models.
- A method for concatenation has been proposed together with the definition of several performance indicator parameters to enable different configurations to be evaluated.
- Concatenation seems to dampen the model parameters and consequently weakens the model's ability to discriminate the break risk.
- Contrary to what was at first feared, the model is not sensitive to datasets with numerous very short segments. If for file size or computation time reasons, one wishes to simplify the input data set, a 5 m threshold concatenation is however a reasonable option. These results have to be confirmed with other case studies.
- Future research should look at the opposite problem with the focus on the splitting of the longest pipes (provided breaks are geo-referenced and can therefore be re-allocated without error risk).
- The concatenation after the model has been calibrated deserves also to be studied in order to design practical replacement projects. This operation inevitably involves gathering in the same project, pipes with various failure risks and therefore needs optimisation procedures to maximise the aggregated failure risk while matching operational constraints.

References

CARE-W (2003), CARE W-consortium, editor. *CareW Website: http://care-unifr.it/, 2003.*

Clark R.M., Stafford C.L., Goodrich J.A (1982).. Water distribution systems: A spatial and cost evaluation. *Journal of Water Ressources Planning and Management. 108(3): 243-256, 1982.*

Eisenbeis P (1994), *Modélisation statistique de la prevision des défaillances sur les conduits d'eau potable.* PhD thesis, Université Louis Pasteur Strasbourg, 1994.

Eisenbeis P, Le Gat Y, Poulton M (2002), Failureforecast and hydraulic reliability models for rehabilitation decision aid. In *Proceedings of the International Conference CARE-W*, 97-106, Dresden, Germany, November 1st 2002. TU Dresden.

Le Gat Y (2007), Etude du Processus de Yule Non Homogène – Application à la modélisation du risque de casses en réseau d'AEP, PhD thesis, ENGREF Paris, 2007 (in preparation).

Le Gat Y and Eisenbeis P(2000), Using maintenance records to forecast failures in water networks, *Urban Warter*, 2:173-181, 2000.