Journal de la Société Française de Statistique
Vol. 155 No. 3 (2014)

Integration of time-dependent covariates in recurrent events modelling : application to failures on drinking water networks

Title: Intégration des covariables temporelles dépendant du temps dans la modélisation d'évènements récurrents : application aux réseaux d'eau potable

Karim Claudio[1,2], Vincent Couallier[2] and Yves Le Gat[3]

Abstract: The standard Linear Extended Yule Process is a well-adapted stochastic model for water pipes failure modelling, as it takes into account the past number of pipe failures, the ageing and the effects of covariates in the failure rate. But fluctuations of failures along time in water network show the limits of the model, which does not consider time-dependent covariates, like frost effect. An improvement of the model actually used in a standard tool such as *PREVOIR*© *Canalisation* is therefore proposed that considers a time dependent covariate. This allows for dramatically improving the predictions of the number of pipe failures according to the climate.

Résumé : L'extension linéaire standard du processus de Yule est un modèle adapté pour la modélisation des défaillances sur les réseaux d'eau potable, dans la mesure où il inclut dans le calcul du taux de défaillance les effets des défaillances passées, de l'âge et de différentes variables. Cependant, les fluctuations du nombre de défaillances au cours du temps montrent les limites de ce modèle dans la mesure où il n'intègre pas l'effet de variables exogènes évoluant avec le temps comme l'effet du climat. Une amélioration du modèle actuellement utilisé dans des outils comme *PREVOIR*© *Canalisation* est ainsi proposée, incluant à présent des variables temporelles. Ceci permet d'améliorer la prédiction du nombre de défaillances selon l'évolution du climat.

Keywords: Dynamical intensity of counting process, Linear Extension of Yule Process (LEYP), Time-dependent covariate in water network failure models
Mots-clés : Extension linéaire du processus de Yule (LEYP), Intensité dynamique d'un processus de comptage, Variable temporelle dans les modèles de défaillances des réseaux d'eau potable
Classification AMS 2000 : 62N05, 90B25

1. Introduction

French drinking water networks stretch on 850 000 km and their value goes beyond $100 billions. This is why their maintenance and preservation have become a priority for water network managers. A study carried out in 2008 by the French Ministry of Environment shows that the global network efficiency reaches 80 % of global water production, which means that 20 % of the total supplied volumes are lost. Failures due to pipe ageing and environmental conditions explain a part of this loss, and may cause a deterioration of the service quality (disruptions in water supply, flooding risks, nuisance due to repair works).

[1] LyRE. E-mail: karim.claudio@lyonnaise-des-eaux.fr
[2] Institut de Mathématiques de Bordeaux. E-mail: vincent.couallier@u.bordeaux2.fr
[3] IRSTEA. E-mail: yves.legat@irstea.fr

 ISSN: 2102-6238

Many indicators have been established to quantify the losses on drinking network. Water network managers such as Lyonnaise des Eaux is committed to local authorities to maintain a linear indicator of failure (ILC: number of failures per kilometer per year) below a certain level.

In this perspective, Lyonnaise des Eaux has developed PREVOIR© Canalisation, a tool based on water pipe failures modelling, in order to predict failures on the water network and plan renewal of the network. It uses a stochastic model of water pipe failures: the Linear Extended Yule Process (Gat, 2009; Babykina, 2010). The LEYP model is defined in the framework of the counting process theory through a parametric formulation of its stochastic intensity that depends on the number of past failures (contrarily to the well known memoryless Poisson process). It thus considers that the risk of failure depends on the history of the pipe (Yule factor, Eisenbeis, 1994) and on its age (Weibull factor, Lawless, 1987). As it has been shown that the failures on a pipe in a water network are correlated with its length, diameter, ground corrosivity (Røstum, 2000) and water pressure (Zyl and Clayton, 2007), a third factor is considered in the model, the Cox regression factor, which includes these effects.

Thus, when applied to pipe failure data, the LEYP model is able to correctly represent the average trend of the number of failures. But it was observed that, in winter, the number of failures could greatly increase in case of cold weather (failure intensity is all the more important since winter is harsher) and the model is unable in its original setting to explain the variations linked to varying environmental conditions. Kleiner and Rajani (2001) have shown that a water pipe failure model (in their case a Poisson regression) could also include dynamic variables, and namely time-dependent covariates. Moreover, Babykina and Couallier (2012) showed the efficiency of a LEYP model comparing to a Non Homogeneous Poisson Process (NHPP), both using time-dependent covariates.

In this paper, an approach is proposed to model the influence of climate on failure intensity, by considering climate effect as a time-dependent factor. Based on the work of Kleiner and Rajani (2001) and Babykina *et al.* (Babykina, 2010; Babykina and Couallier, 2010, 2012), a time-dependant covariate is built starting from the temperature (low temperature in particular) and soil moisture. The new model, applied to data from the Urban Community of Bordeaux (CUB), is shown to improve the predictive ability of the tool and allows to estimate two quantities: first a retrospective estimation of the failures on a past period (explaining the variations of breaks during cold periods), and secondly a prediction of forthcoming failures, based on an hypothetic climatic scenario. The paper is organized as follows. Section 2 briefly presents the framework of the counting process theory to introduce the LEYP model. Section 3 gives the statistical analysis of the LEYP model from i.i.d. copies of the LEYP process observed on a time interval [a,b], where, for sake of readability, only one time-dependent covariate is considered. Section 4 contains an application on a real dataset from the Bordeaux city recorded between 2000 and 2010. Prospects and conclusions complete the paper.

2. The LEYP model

The construction of the Linear Extension of the Yule Process (LEYP) basically stems from the benchmarking of failure models carried out during the FP5 european research project CARE-W (Herz, 2002 and Saegrov, 2005). This collaborative work emphasized both strengths and weaknesses of the NHPP, initially proposed by Røstum (2000) as a water main failure modelling

 ISSN: 2102-6238

tool. Designed within the counting processes framework, the NHPP is able to handle left-truncated right-censored data in a natural way, to account for explanatory factors with the proportional hazard feature (Cox, 1972). A serious drawback is that the NHPP is memoryless, and cannot therefore account for the dependence of the failure process on past failures, and the consequential tendency of pipe failures to concentrate on the same network segments (as emphasized by Clark et al., 1982 and Eisenbeis, 1994). This fact induced an extension of the NHPP to define a process with an intensity keeping the memory of past events. Such a basic construct is known in the counting process literature as the so called Yule process (Greenwood and Yule, 1920 and Ross, 1983). Note that this model relies on a huge family of stochastic models with dynamically defined intensity process and could be opposed to models with frailty intensity (Peña, 2006, Aalen et al., 2008, p.332).

2.1. The Counting Processes Framework

It is convenient to model recurrent events using the general counting process framework covered in Aalen et al. (2008) and Andersen et al. (1993) and briefly presented in this section.

Let $(T_j)_{j\geq 0}$ be the times of successive failures of a system observed over a certain time-period with $T_0 = 0$ (time of installation). In majority of datasets devoted to successive failures of repairable systems, the repair times are either insignificant and/or not available, thus we will consider hereafter that moments of failures and repairs coincide. We also assume that failures are immediately detected.

Let $N(t) = \sum_{j=1}^{\infty} I(T_j \leq t)$ be the number of failures occurred up to t and $(N(t))_{t\geq 0}$ be the counting process. With respect to the filtration $\{\mathscr{N}_t\}$ generated by the counting process $(N(t))_{t\geq 0}$ itself, we can define via the Doob decomposition of $N(t)$ the compensator $C(t)$, a non decreasing predictible process such that

$$N(t) = C(t) + M(t)$$

where $M(t)$ is a zero mean martingale. We assume here that $C(t)$ is absolutely continuous, and thus satisfies

$$C(t) = \int_0^t \lambda(s)ds$$

where λ is called the (predictable) intensity process of $N(t)$. Following Aalen et al. (2008), $\lambda(t)$ is heuristically defined by

$$\lambda(t)dt = E\left[dN(t)\,|\,\mathscr{N}_{t-}\right], \tag{1}$$

and, in order to apply regression models to recurrent events data, we extend the information available by enlarging the filtration and thus define a new stochastic intensity process :

$$\lambda(t)dt = E\left[dN(t)\,|\,\mathscr{N}_{t-}, \mathbf{X}(t)\right], \tag{2}$$

where $\{\mathscr{N}_t\}$ is the filtration generated by the counting process $(N(t))_{t\geq 0}$ and $\mathbf{X}(t)$ is a (possibly vector-valued) covariate process recording external information. The function $\lambda(\cdot)$ in Eq. (2) can be interpreted as the instantaneous propensity of a system to fail during an infinitesimal interval

 ISSN: 2102-6238

dt, given the history of the failure process at time t and knowing the value of some covariates at time t or before time t.

In the application considered in this paper, a time segment is considered, with failure/maintenance records observed in the time interval $[a,b]$. Without loss of generality, the time axis stands for the segment age, and its origin $t=0$ is set when the pipe is put into service. In that case, no information is available in the time interval $[0,a]$ and we therefore face a left-filtered counting process: what is observed is in fact m observed failure times $0 \leqslant a = t_0 \leqslant t_1 < \ldots < t_j < \ldots < t_m \leqslant b = t_{m+1} < +\infty$.

This censoring mechanism does not change the definition of the stochastic intensity process (with respect to the natural filtration $\{\mathscr{N}_t\}$), but induces modifications in the estimation procedure (see section 3.1).

We start with the simplest and well known NHPP defined by:

$$\begin{cases} N(0) = 0 \\ \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{t-}\right) = \lambda(t)\mathrm{d}t \end{cases} \tag{3}$$

where $\lambda(\cdot)$ is a deterministic function (that can be parameterized and put into a regression model with covariates). It owns the basic property that its counting process is Poisson distributed: $N(t) \sim \mathcal{P}o(\Lambda(t))$ where $\Lambda(t) = \int_0^t \lambda(s)ds$. The memoryless property is revealed by the fact that the random number of failures in $[a,t]$, $N(t) - N(a)$ is independent of the past, for instance given by the variable $N(a)$, the number of failure in $[0,a]$. This implies that n i.i.d. copies of an NHPP process will have the same conditional distribution of $(N(t) - N(a)|N(a) = p)$. In maintenance analysis, this implies that the m previous maintenances do not modify the intensity of forthcoming failures. It could be preferable to handle a stochastic model where units with early failures will tend to have an higher intensity of failure at time t. This fact motivates the introduction of the Yule process.

Following Ross (1983), the Yule Process is a Markovian process $N(t)$ that records the size of a population initially composed of $k \in \mathbb{N}^*$ individuals at $t=0$, and characterized by a constant reproduction rate λ and a null death rate:

$$\begin{cases} N(0) = k \\ \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{t-}\right) = N(t-)\lambda\mathrm{d}t \end{cases} \tag{4}$$

It is well known that this counting process has a negative binomial distribution: $N(t) \sim \mathcal{NB}(k, e^{-\lambda t})$. Replacing the constant intensity λ in Eq. (4) by an age-dependent intensity function $\lambda(.)$ leads to the "Time-Dependent Yule Process" considered by Chang et al. (2002).

Definition 1. *The Linear Extension of the Yule Process (LEYP)*

The Linear Extension of the Yule Process (LEYP) is defined by considering an intensity that linearly depends on the number of past events (Gat, 2009, 2013):

$$\mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{t-}\right) = (1 + \alpha N(t-))\lambda(t)\mathrm{d}t \tag{5}$$

where $\alpha \in \mathbb{R}_+^*$ and $\lambda(\cdot)$ is a deterministic function.

Proposition 1. *Let us denote $\Lambda(t) = \int_0^t \lambda(s)ds$ and $\mu(t) = \exp(\alpha\Lambda(t))$. The number of events occured in the period $[0,t]$ follows a negative binomial distribution:*

$$N(t) \sim \mathcal{NB}\left(\alpha^{-1}, \frac{1}{\mu(t)}\right) \tag{6}$$

 ISSN: 2102-6238

A proof a Proposition 1 is developed in Gat (2009). For the practical use of the LEYP model in asset management, it is important to know the distribution of the possible number of failures in a prediction interval $[c,d]$ given m failures occurred in the observation interval $[a,b]$, as illustrated by Fig. 1.

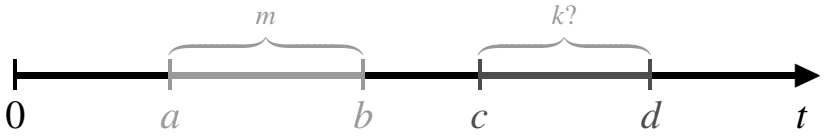

FIGURE 1. *Observation and prediction intervals*

Proposition 2. *The conditional number of events in* $[c,d]$ *given the number of events in* $[a,b]$ $(0 \leqslant a \leqslant b \leqslant c \leqslant d)$ *follows a negative binomial distribution:*

$$[N(d)-N(c) \mid N(b)-N(a)=m] \sim \mathcal{NB}\left(\alpha^{-1}+m, \frac{\mu(b)-\mu(a)+1}{\mu(d)-\mu(c)+\mu(b)-\mu(a)+1}\right) \tag{7}$$

The proof can be found in Gat (2009).

3. Statistical analysis for filtered data

3.1. The Likelihood function

In practice the model parameters are to be estimated with pipe maintenance data restricted within an observation window. In Western Europe, pipe failure time series are indeed rarely available in digital format before 1980, and actual datasets very often start after 1995. For parameter estimation purpose, it is thus crucial to establish the formula of the likelihood of the model parameter given failure times observed within a time interval (see Martinussen and Scheike, 2006, chap.3).

The LEYP parameters, namely α and the parameters of the function $\lambda(\cdot)$, compose the vector $\boldsymbol{\theta}$. The likelihood of the LEYP parameter $\boldsymbol{\theta}$ given the failure sequence $\left(t_j\right)_{j=1,\ldots,m}$ observed for one pipe in the time interval $[a,b]$ uses the mathematical concept of "product integral", which is a continuous generalisation of the discrete product operation and the product analogue of classical integration (*cf. e.g.* Gill and Johansen, 1990), and is used by Gat (2009) :

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{m} \mathrm{E}\left(\mathrm{d}N(t_i)|\mathscr{N}_{t_i-}\right) \prod_{i=1}^{m} \prod_{t\in[t_i,t_{i+1}]} \left(1-\mathrm{E}\left(\mathrm{d}N(t_i)|\mathscr{N}_{t_i-}\right)\right) \tag{8}$$

The product integral term $\prod_{t\in[t_i,t_{i+1}]}(1-\mathrm{E}\left(\mathrm{d}N(t_i)|\mathscr{N}_{t_i-}\right))$ stands for the joint probability of the non-occurrence of failures in each infinitesimal time interval between the i-th and $(i+1)$-th failures of the system. Finally, by using the conditional distribution of equation (7), equation (8) reduces to (see Gat, 2009 for a proof) :

$$L(\boldsymbol{\theta}) = \alpha^m \frac{\Gamma(\alpha^{-1}+m)}{\Gamma(\alpha^{-1})} \frac{\prod_{j=1}^{m}\mu(t_j)\lambda(t_j)}{(\mu(b)-\mu(a)+1)^{\alpha^{-1}+m}}. \tag{9}$$

 ISSN: 2102-6238

For an n-sample of segments assumed to fail independently, the global likelihood is simply the product of the individual likelihoods:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta})$$

3.2. *Statistical estimation and testing for the Yule-Weibull-Cox intensity*

In the application of the LEYP model for water asset management, the intensity of failure $\lambda(\cdot)$ has to be made explicit and takes the form of a Cox factor, with a vector $\boldsymbol{X(t)}$ of covariates of dimension $p+1$, composed by an intercept, $p-1$ fixed covariates and a time-dependent one.

For practical use of the LEYP model, the "Yule-Weibull-Cox" intensity related to a segment with age t, characterized by covariates vector $\boldsymbol{X(t)}$, and that has undergone j failures, is defined as the product of three factors:

- Yule factor $1+\alpha j$ that ensures the dependence on the number j of past events,
- Weibull factor $\delta t^{\delta-1}$ accounting for ageing, modelled as a power function of time,
- Cox factor $e^{\boldsymbol{X(t)}^T\boldsymbol{\beta}}$ accounting for covariate effects, possibly time-dependent.

The Yule-Weibull-Cox intensity (with p covariates) is defined by the formula:

$$\mathrm{E}_{\boldsymbol{\theta}}\left(\mathrm{d}N(t) \mid N(t-) = j, \boldsymbol{X}(t)\right) = (1+j\alpha)\delta t^{\delta-1} e^{\boldsymbol{X(t)}^T\boldsymbol{\beta}}\mathrm{d}t \tag{10}$$

with: $\boldsymbol{\theta} = \left(\alpha, \delta, \boldsymbol{\beta}^{\mathrm{T}}\right)^{\mathrm{T}}, \alpha > 0,\ \delta \geqslant 1,\ \boldsymbol{X}(t) = (1, X_1, X_2, \ldots, X_{p-1}, X_p(t))^{\mathrm{T}},\ \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^{p+1}$.

The "1" as first component of $\boldsymbol{X}$ and the corresponding β_0 in the vector of regression coefficients $\boldsymbol{\beta}$ specify the baseline intensity $(1+\alpha j)\delta t^{\delta-1} e^{\beta_0}$.

Using the Yule-Weibull-Cox intensity for defining $\lambda(\cdot)$, the likelihood of the LEYP parameter $\boldsymbol{\theta}$ can be rewritten as follow:

$$L(\boldsymbol{\theta}) = L\left(\alpha, \delta, \boldsymbol{\beta}^{\mathrm{T}}\right) = \alpha^m \frac{\Gamma(\alpha^{-1}+m)}{\Gamma(\alpha^{-1})} \frac{\prod_{j=1}^{m} \exp\left(\alpha \int_0^{t_j} \delta s^{\delta-1} e^{\boldsymbol{X}(s)^T\boldsymbol{\beta}}\mathrm{d}s\right) \delta t_j^{\delta-1} e^{\boldsymbol{X}(t_j)^T\boldsymbol{\beta}}}{\left[\exp\left(\alpha \int_0^{b} \delta s^{\delta-1} e^{\boldsymbol{X}(s)^T\boldsymbol{\beta}}\mathrm{d}s\right) - \exp\left(\alpha \int_0^{a} \delta s^{\delta-1} e^{\boldsymbol{X}(s)^T\boldsymbol{\beta}}\mathrm{d}s\right) + 1\right]^{\alpha^{-1}+m}}.$$

As usual, the Maximum-Likelihood equations $\partial \ln L(\boldsymbol{\theta})/\partial\theta_k = 0$ do not have explicit solutions when $L(\boldsymbol{\theta})$ is given by Eq. (9). Extensive computer simulations have however been carried out to investigate the convexity of $\ln L(\boldsymbol{\theta})$, with always a positive outcome. The ML estimation of the LEYP model parameter can be therefore confidently performed. The classical Nelder-Mead numerical optimisation algorithm proposed by Nelder and Mead (1965) is used to that end; this method is in particular implemented in the free software Casses© (*cf.* Renaud et al., 2012).

Following the considerations above, we assume that the ML estimate $\hat{\boldsymbol{\theta}}$ has asymptotically a Gaussian distribution with expectation $\boldsymbol{\theta}$ and variance estimated from the second derivative of $\ln L(\boldsymbol{\theta})$:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = \left(-\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\theta^2}\right)^{-1}_{\hat{\boldsymbol{\theta}}}.$$

The validity of this assumption has been empirically obtained in Babykina and Couallier (2012) but remains to be mathematically proved.

 ISSN: 2102-6238

The significance of the effect of the covariates can then be tested using a classical Wald Chi-Squared test (test of the nullity of the parameters $\boldsymbol{\beta}$). As the parameters α and δ cannot be null (*cf.* Eq. 10), a Likelihood-Ratio Chi-Squared test (Pawitan, 2000) is used to compare δ to 1 (no ageing effect) and α against a value close to 0 (*e.g.* 0.1). Due to the left-truncation of the calibration window $[a,b]$, martingale residuals (Therneau et al., 1990) cannot be calculated since $N(t)$ is unknown when $a > 0$ (only $N(t) - N(a)$ is observed). It is possible yet to graphically compare the empirical and theoretical failure rates averaged over the sample of pipes, and plotted against age, as explained in (Andersen et al., 1993, pp. 230-234).

3.3. Graphical goodness of fit and validation of the LEYP predictions

Let us note by $[a,c]$ a period of observation (date format). The method proposed to validate the model and its predictions consists of:

– calibrating (on $[a,b]$) the model using e.g. the first 80 % of the duration of the available observation window,
– predicting (on $[b,c]$) for each pipe the number of failures in the last 20 % of the window.

For each pipe i, we denote by a_i, b_i, c_i the corresponding ages at dates a, b and c. This makes it possible to compare in the last 20 % of the observation window the actual number of failures versus the predicted ones. The relative comparison allows an assessment of the ability of the model to detect the pipes with highest failure risks, by measuring how actual failures are concentrated on the pipes with the highest expected failure rate. This is done by building a predictive performance curve (also known as a lift curve, see Hong and Weiss, 2001) in the following way :

Each pipe i of length l_i is assumed to have been observed both:

– within the calibration interval $[a_i, b_i]$, where it experienced m_i failures,
– within the validation interval $[b_i, c_i]$, where it experienced k_i failures.

The key point is the ranking of the pipes by descending expected failure rate (per length):

$$\frac{\hat{k}_i}{l_i(c_i - b_i)} \tag{11}$$

where $\hat{k}_i = \mathrm{E}_{\hat{\boldsymbol{\theta}}}(N_i(c_i) - N_i(b_i) \mid N_i(b_i) - N_i(a_i) = m_i, \boldsymbol{X}_i(t))$ is the predicted number of failures in $[b_i, c_i]$ knowing the past history. Considering the subsets of size $q = 1 \ldots n$ of pipes with the q highest expected failure rates, the relative risk rank weighted by the pipe length is calculated as:

$$r_{(q)} = \sum_{j=1}^{q} l_{(j)} / \sum_{j=1}^{n} l_{(j)} \tag{12}$$

and the relative number of failures observed within the validation interval:

$$\kappa_{(q)} = \sum_{j=1}^{q} k_{(j)} / \sum_{j=1}^{n} k_{(j)}$$

where parenthesized index (j) means that pipes are arranged in descending order of expected failure rate. The predictive performance curve is the graph $\left(r_{(q)}, \kappa_{(q)}\right)_{q \in \{1,\ldots,n\}}$, as illustrated by

 ISSN: 2102-6238

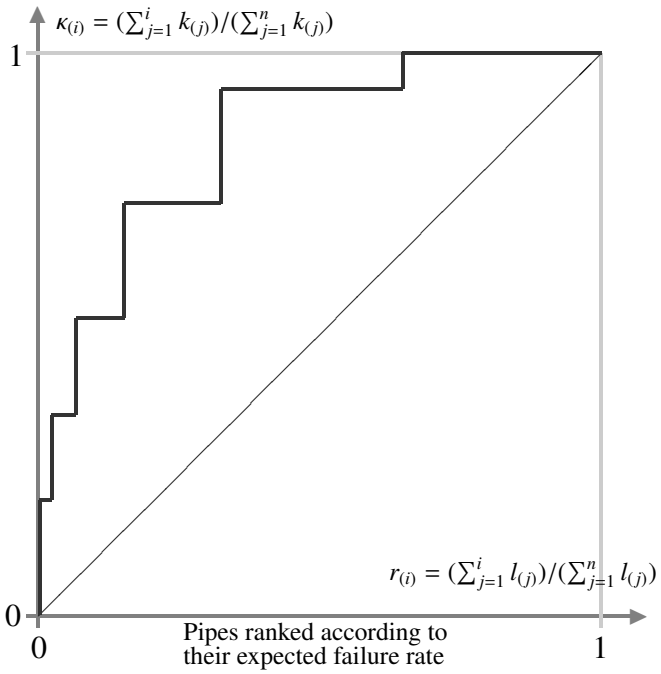

FIGURE 2. *Predictive performance curve*

Fig. 2. The area ξ under the step curve $\left(r_{(q)}, \kappa_{(q)}\right)_{q=1,\ldots,n}$ is calculated as:

$$\xi = \frac{\sum_{i=1}^{n} l_{(i)} \kappa_{(i)}}{\sum_{i=1}^{n} l_{(i)}}$$

Ratio ξ is all the greater since observed failures are more concentrated on pipes identified by the model as being the most likely to fail within $[b, c]$. Ratio $\kappa(q)$ can be interpreted as the proportion of the actual failures observed within $[b, c]$ that would have been avoided if a rate $r(q)$ of the total pipe length had been replaced at b, provided

– pipe replacement priorities had been set according to the predicted failure rate,
– the failure probability between ages 0 and $c - b$ is negligible.

Despite a formal resemblance with the Receiver Operating Curve used by Debón et al. (2010), the predictive performance curve cannot be similarly interpreted, since the response variable is not a Bernoulli random variable.

Remark : We are seeking to know whether the data let suppose the existence or not of a bias by checking that the sum of the actual numbers of failures $\sum_{j=1}^{q} k_j$ lies within the 95 % confidence interval of the sum of the expectations of the predicted numbers of failures $\sum_{j=1}^{q} \hat{k}_j$. This confidence interval can be estimated using the asymptotic normality of a sum of independent negative binomial variables with variance $\sum_{j=1}^{q} \mathrm{Var}(\hat{k}_j)$, provided the variances $\mathrm{Var}(\hat{k}_j)$ do not vary too much. This empirical checking does not formally prove the unbiasedness of model predictions, but ensures at least that this theoretical hypothesis is not contradicted by the studied data.

 ISSN: 2102-6238

Table 1. *Description of data used - CUB drinking water assets*

Number of pipes in 2010	33982
Network length (km)	3081
Mean age (2010)	47.9
Observation window	2000-2010 (11 years)
Mean number of failures per year	303
Failure rate on drinking main (nb failure/km/year)	0.1
Percentage of main with at least one failure	7

4. Application to data from the city of Bordeaux, 2000-2010.

The model is built and applied to the CUB data, composed by 3081 km of water network and a historic of failures from 2000 to 2010 described in Table 1.

First of all, according to physical knowledge of engineers, the model and the estimation of the parameters have to be stratified by material: cast iron/asbestos-concrete (52 % of total network), ductile iron/steel (34 %) and PVC/PE (14 %). Indeed, the effect of static and dynamic covariates considered depends on the material. In order to construct a synthetic time-dependent covariate, we have to take care of the phenomenon we are modelling and the time scale of recorded data.

To choose the time scale of $X_p(t)$, a balance between a high number of failures and a precise information on the time dependent covariate has to be found: the month seems to be the best candidate. The phenomenon to model is the frost (Kleiner and Rajani, 2002) so the time dependent covariate is built from cold temperature and soil moisture. It takes, at time t, the form:

$$X_p(t) = M_1(t) + M_2(t) + \sqrt{RS(t)}$$

where :

- $M_1(t)$ is the maximal number of consecutive days where minimal temperate is below -1 °C during month t,
- $M_2(t)$ the number of days where minimal temperature is below -3 °C and the difference between minimal and maximal temperature exceed +5 °C, during month t,
- $RS(t)$ the number of days with a relative soil moisture over 80 % over month t.

$M_1(t)$ and $M_2(t)$ have not the same order of magnitude as $RS(t)$. As the effect of temperature is greater than the effect of rainfall for pipe failure, it has been decided to take the square root of $RS(t)$ in order to minimize the impact on $X_p(t)$.

The model thus accounts for, as a time dependent phenomenon, long and intense cold events, abrupt changes of temperature and possible soil frost. Several covariates have been tested, but we however only present result for this one which gives the best result.

Figure 3 represents the evolution of the ILC (the linear indice of failures, expressed in number of failures per kilometre per month) through time and the value of X_p on the 2000-2010 period. As it can be seen, the climatic covariate chosen seems to be a good candidate to explain variation of failure over time (the linear correlation coefficient between the two quantities equals 0.71).

 ISSN: 2102-6238

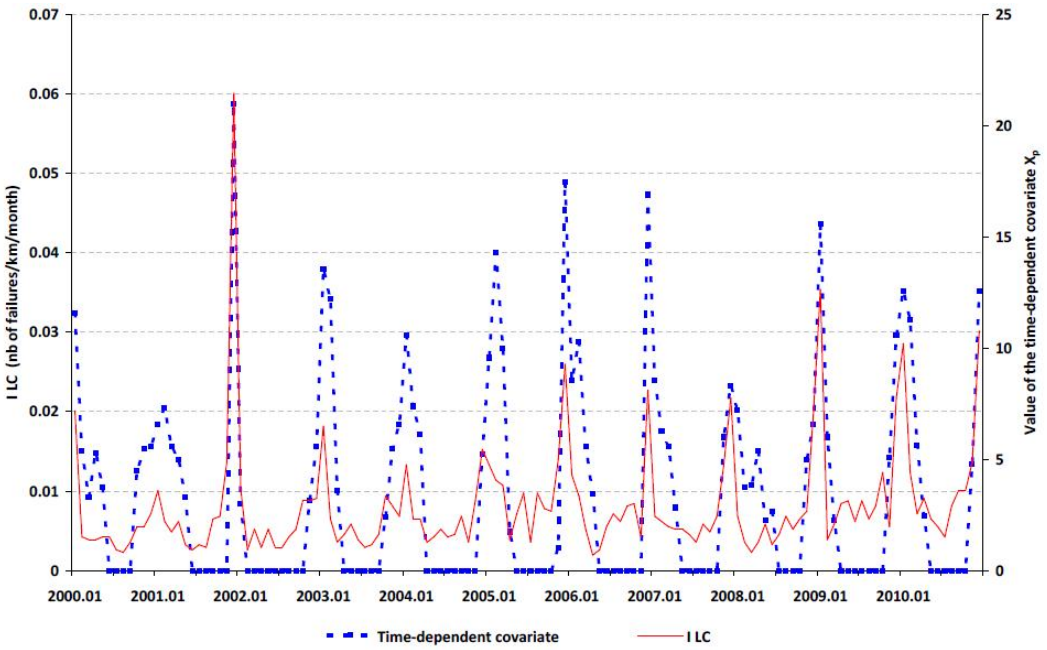

FIGURE 3. *Linear indicator of failure (ILC) and time-dependant covariate - CUB data between 2000 and 2010*

There is an undeniable correlation between the frost (represented by the covariate X_p) and the failure process. However this doesn't prove that X_p has a significant effect on the failure process.

4.1. Estimation of the model parameters

As previously said, the model, namely the estimation of the parameters, is stratified by material. The different covariates have not the same effect indeed on a pipe according to its nature. For instance, the ageing is more marked on ductile iron pipes than on grey iron pipes (as ageing of grey iron pipes is hardly visible on such a short time period, we consider that ageing is negligible for these pipes and put $\delta = 1$).

The tables 2, 3 and 4 contain the values of the parameters, for each type of material, estimated according to the maximum likelihood method, and the p-value of the null hypothesis test (comparing $\boldsymbol{\beta}$ to 0^{p+1}, θ to 1 and α to 0.1 unilaterally). Moreover, a comparison with the parameter estimates of the model without a time dependent covariate can be done. Only significant covariates (*i.e.* p-value $\leqslant 0.05$) have been retained. Hence, Table 4 only contains the model without time-dependent covariate.

Except for the plastic pipe model (there is no effect of frost on plastic pipe failure), the time dependent covariate has always a significative effect in the model, according to the result of the null hypothesis test. But the addition of the covariate has an effect on the other parameters too.

 ISSN: 2102-6238

TABLE 2. *Parameter estimation (and p-value) - case of ductile iron/steel pipes*

Variable (parameter)	Parameters estimated with time dependent covariate	Parameters estimated without time dependent covariate
Historic failures (α)	5.76 (< 1E-16)	6.91 (< 1E-16)
Age (δ)	1.25 (6.2E-05)	1.12 (1.9E-02)
Intercept (β_0)	-4.24 (< 1E-16)	-3.93 (< 1E-16)
Length	0.59 (< 1E-16)	0.53 (< 1E-16)
Pressure	0.26 (9.5E-02)	-0.06 (7.4E-01)
Diameter	-0.002 (1.8E-02)	-0.001 (2.4E-02)
Ground corrosivity	0.54 (8.5E-06)	0.54 (6.2E-06)
Connecting pipe density	2.77 (1.0E-03)	2.28 (8.0E-03)
Climate	0.07 (3.2E-11)	

TABLE 3. *Parameter estimation (and p-value) - case of grey iron/asbestos pipes*

Variable (parameter)	Parameters estimated with time dependent covariate	Parameters estimated without time dependent covariate
Historic failures (α)	1.49 (< 1E-16)	1.54 (< 1E-16)
Age (δ)	1	1
Intercept (β_0)	-2.84 (< 1E-16)	-2.25 (< 1E-16)
Length	0.49 (< 1E-16)	0.47 (< 1E-16)
Pressure	0.05 (1.7E-05)	0.05 (1.7E-05)
Diameter	-0.003 (< 1E-16)	-0.002 (< 1E-16)
Laid before 1931	-0.36 (< 1E-16)	-0.38 (< 1E-16)
Laid between 1931 and 1945	-0.10 (1.3E-03)	-0.11 (4.4 E-03)
Ground corrosivity	0.27 (1.1E-15)	0.26 (1.8E-15)
Connecting pipe density	0.81 (3.4E-04)	0.76 (5.6E-04)
Climate	0.11 (< 1E-16)	

 ISSN: 2102-6238

TABLE 4. *Parameter estimation (and p-value) - case of PVC/PE pipes*

Variable (parameter)	Parameters estimated without time dependent covariate
Historic failures (α)	9.90 (< 1E-16)
Age (δ)	1.31 (1.1E-16)
Intercept (β_0)	-2.27 (9.2E-09)
Length	0.51 (1.1E-11)
Material	-0.96 (1.9E-08)

First of all, in the new model, there is an increase in the effect of ageing and a decrease in the effect of past failures on the failure process. Then, in particular in the ductile iron model, the effect of pressure becomes significant and the connecting pipe density has a greater effect in the risk of failure in the new model. Finally, the effects of length, diameter and ground corrosivity are the same in both models. Considering a time dependent covariate enhances thus the explanatory power of the model. However, the estimation of the parameters only proves that the climate influences the failure process (p-value is lower than 5%) by stimulating it (the sign of the parameter is positive). But does this covariate explain the fluctuations observed in the number of failures?

4.2. Accuracy of the model

A first way to check the improvement of the model due to the climate-related covariate is to analyse its accuracy: the estimation of the number of failures is carried out on the same population and on the same period as the parameter calibration. Then a comparison with the model without time dependent covariate is performed, based on observed and estimated failure numbers.

On the 2000-2010 period, 3,334 failures were recorded. The original LEYP model predicts on this period 3,314 failures, while the improved one predicts 3,313. There is an equivalent performance of both models on a long period (11 years) but the difference appears on a short-time period.

As shown by Figure 4, the fluctuations of the number of failures, which are smoothed by the model without climate-related covariate, are better accounted for by the improved model. This means that the yearly variation of the number of failures can be explained in a satisfactory way as a weather effect. To sum up, the climate-related covariate allows the LEYP model to fit better the observed fluctuations of the failure process used to calibrate it. But to be practically useful, the model should not only be able to explain past events but should predict forthcoming events too.

4.3. Predictive performance of the model

As explained in section 3.3, the principle of the validation method is simple: the calibration of the parameters is done using the first 80 % of the observation window (2000-2008) and prediction

 ISSN: 2102-6238

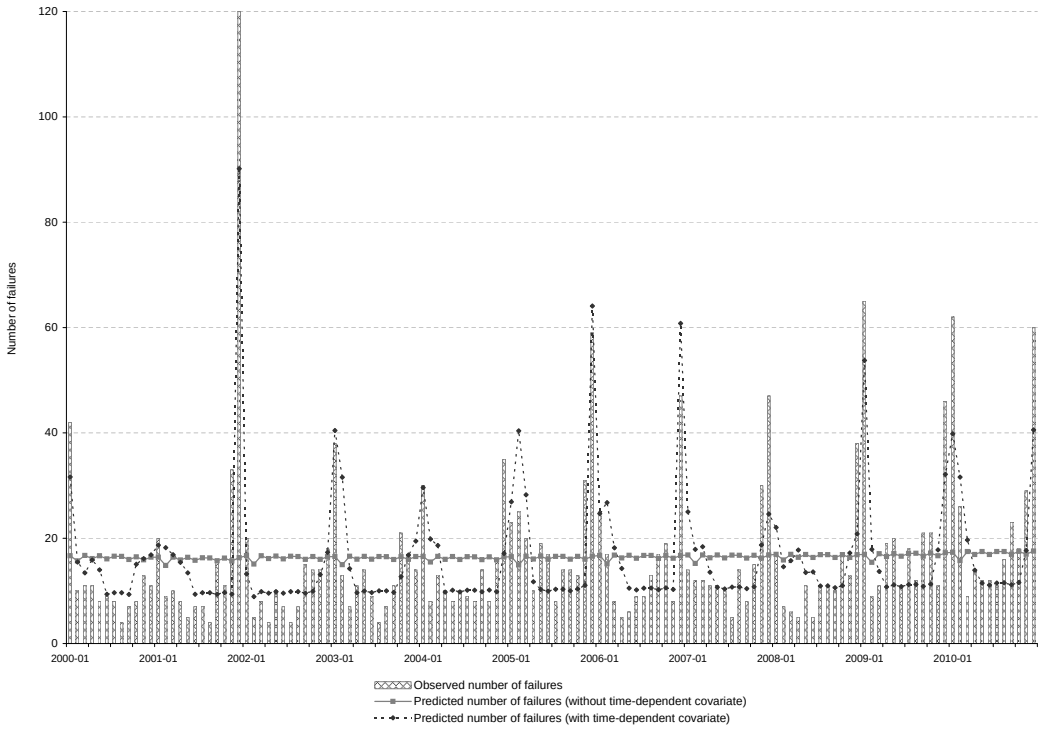

FIGURE 4. *Comparison of the monthly numbers of failures observed and estimated using LEYP models with and without time dependent covariate, on the whole CUB network*

is carried out with the last 20 % (2009-2010).

After calibrating the parameters on the 2000-2008 period, a probability distribution of the failure number is calculated for each pipe. The pipes are then ordered by descending expected failure rate (see Eq. (11)). Figure 5 plots the cumulative relative numbers of observed failures ($\kappa_{(q)}$) and predicted failures ($\hat{\kappa}_{(q)}$) against the relative risk ranks of the sorted pipes (see Eq. (12)), as explained in section 3.3.

First of all, both curves (observation and prediction) are very close to each other for the pipes with the highest theoretical failure rate, which means that the model is correctly able to identify the pipes with the highest failure proneness.

Figure 5 can also be interpreted as follows: the replacement of the first 7% of the pipes with a highest expected failure rate would have avoided 40% of the failures that occured in 2009 and 2010. So not only observed and predicted curves have to be close to each other, but they must also have a high gradient on the first pipes (those which have a high failure probability) in order to have a model with a good predictive power. This high proportion of failures avoided by the replacement of a given low percentage of the pipes with the highest expected failure rates characterizes the good predictive power of the model.

So finally, the climate-related covariate efficiently improves the LEYP model, as it is able to

 ISSN: 2102-6238

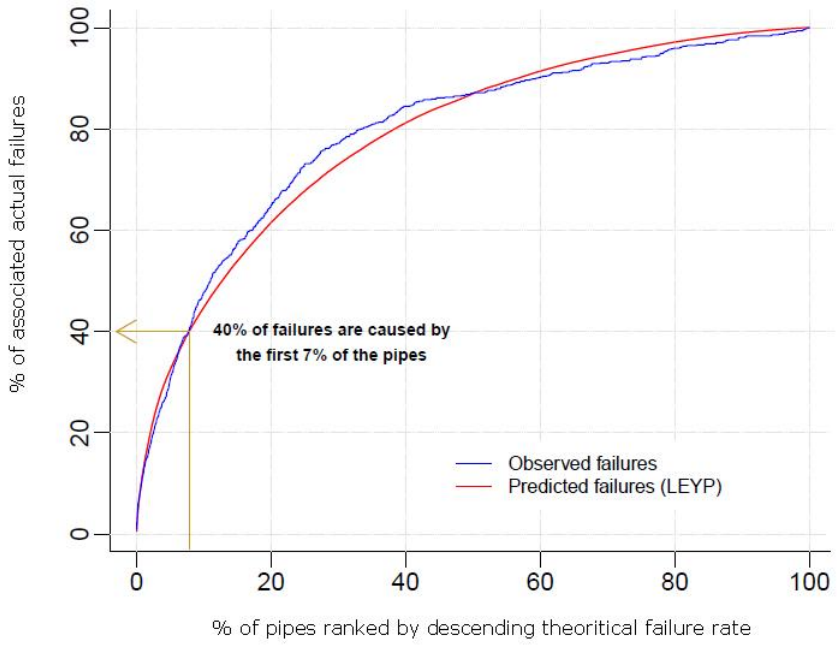

FIGURE 5. *Percentage of breaks associated to the pipes according to the model and the observations*

explain the temporal fluctuations of the failure process.

5. Conclusion and prospects

Despite its good efficiency, the LEYP model suffers from some limitations in its original setting. The evolution through time of the predicted number of failures is only accounted for by the Weibull and the Yule factors. But according to their form, the only possibility of temporal evolution is monotonous. Adding a time dependent covariate in the Cox factor allows to make the prediction temporarily fluctuate. If this new covariate summarizes seasonal effects linked to the weather then the model accuracy in predicting the annual number of failures is improved.

An obvious drawback of the new model is that its use to forecast future failures requires climatic forecast input data, which are not available with a reasonable accuracy beyond a very short term. Using the LEYP model with a climate-related covariate to perform medium or long term failure forecasts requires therefore to hypothesise climatic scenarios.

There are anyway three main advantages in considering a climatic covariate within the LEYP setup:

1. an improvement of the estimation of the Yule α and Weibull δ parameters: potential estimation biases in these parameter estimates are likely to be corrected by accounting for climatic fluctuations of the failure rate within the time period considered for calibrating the model.
2. a more objective view of the network reliability: as the model enables to explain the effect of the weather on the water network reliability, a temporary increase in the failure rate may

 ISSN: 2102-6238

be clearly imputable to the frost and not to a poor infrastructure asset management policy.

3. an improved tool for prospective studies: in the climate change context, various network renewal policies can be tested and adjusted according to various possible climatic scenarios.

The LEYP model with time dependent covariates is moreover open to various applications (*e.g.* gas network) and developments, beyond the assessment of the effect of the climate on the failure intensity. A first example could be the assessment of the effect of active leak detection (in addition to climate effect), that decreases the water losses on the one hand, but artificially increases the failure rate on the other hand. It is to be noted that leak detection is an example of a covariate that varies in intensity both in time and space, as during a given year only some zones of the network may undergo leak detection campaigns. An other example of such a space-time dependent covariate is provided by the pressure modulation, which is ensured by an hydraulic device that decreases the service pressure on a part of the network during periods of the day when the water demand is lower. Pressure modulation aims at mitigating the water losses due to latent leaks. It is also likely to decrease the failure rate, which beneficial effect could be assessed using the LEYP model with space-time dependent covariates.

Acknowledgment : The authors wish to thank the two anonymous referees for their valuable comments to the manuscript and their constructive suggestions.

References

Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008). *Survival and Event History Analysis*. Springer, New York, 1st edition.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York, 1st edition.

Babykina, G. (2010). *Modélisation statistique d'événements récurrents. Exploration empirique des estimateurs, prise en compte d'une covariable temporelle et application aux défaillances des réseaux d'eau.* PhD thesis, Ecole doctorale de Mathématiques et Informatique, Université de Bordeaux.

Babykina, G. and Couallier, V. (2010). Events for repairable systems under worse than old assumption. In *"Advances in Degradation Modeling Applications to Reliability, Survival Analysis, and Finance, Birkhäuser Boston, 2010, ch. Modeling Recurrent"*, pages 339–354.

Babykina, G. and Couallier, V. (2012). Empirical assessment of the maximum likelihood estimator quality in a parametric counting process model for recurrent events. *Computational Statistics and Data Analysis*, 56(2):297–315.

Chang, C. C. H., Chan, W., and Kapadia, A. S. (2002). The analysis of recurrent failure times : The time-dependent Yule process approach. *Communication in Statistics - Theory and Methods*, 31(7):1203–1213.

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

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220.

Debón, A., Carrión, A., Cabrera, E., and Solano, H. (2010). Comparing risk of failure models in water supply networks using ROC curves. *Reliability Engineering and System Safety*, 95:43–48.

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

Gat, Y. L. (2009). *"Une extension du Processus de Yule pour la modélisation stochastique des événements récurrents - Application aux défaillances de canalisations d'eau sous pression (Extending the Yule Process to model recurrent events: Application to the failures of pressure water mains)*. PhD thesis, Engref-AgroParisTech, Paris, France.

 ISSN: 2102-6238

Gat, Y. L. (2013). Extending the yule process to model recurrent pipe failures in water supply networks. In Press.

Gill, R. and Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, 18(4):1501–1555.

Greenwood, M. and Yule, G. U. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society, Series A*, 83:255–279.

Herz, R. K., editor (2002). *International Conference - Computer Aided Rehabilitation of WaterNetworks CARE-W*. TU Dresden, Chair of Urban Engineering.

Hong, S. J. and Weiss, S. M. (2001). Advances in predictive models for data mining. *Pattern Recognition Letters*, 22(1):55 – 61.

Kleiner, Y. and Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: Statistical models. *Urban Water*, 3(3):131–150.

Kleiner, Y. and Rajani, B. (2002). Forecasting variations and trends in water main breaks. *Journal of Infrastructure Systems*, 8(4):122–131.

Lawless, J. F. (1987). Regression methods for poisson process data. *Journal of the American Statistical Association, Theory and Methods*, 82(399):808–815.

Martinussen, T. and Scheike, T. H. (2006). *Dynamic regression models for survival data*. Statistics for Biology and Health. Springer, Dordrecht.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7:308–313.

Pawitan, Y. (2000). A reminder of the reliability of the wald statistic: likelihood explanation. *The American Statistician*, 54(1):54–56.

Peña, E. A. (2006). Dynamic modelling and statistical analysis of event times. *Stat Sci.*, 21(4):1–26.

Renaud, E., Le Gat, Y., and Poulton, M. (2012). Using a break prediction model for drinking water networks asset management: From research to practice. *Water Science & Technology: Water Supply*, 12(5):674–682.

Ross, S. (1983). *Stochastic Processes*. John Wiley and Sons, New York.

Røstum, J. (2000). Statistical modelling of pipe failures in water networks. Doctor engineer dissertation, Norwegian University of Science and Technology, Department of Hydraulic and Environmental Engineering, Trondheim, Norway.

Saegrov, S. (2005). *CARE-W Computer Aided Rehabilitation for Water Networks*. IWA Publishing.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77(1):147–160.

Zyl, J. V. and Clayton, C. (2007). The effects of pressure on leakage in water distribution system. In *Proc. I.C.E Water Management*, pages 104–109.

 ISSN: 2102-6238