# 1

# Introduction



Examples of recurrent failures abound in the literature devoted to the reliability of technical objects, and in many cases, the occurrence rates tend to increase not only with the ageing of the object, but also with the number of past failures. The effect of ageing can be relevantly modeled using the now classical non-homogeneous Poisson process (NHPP), a comprehensive presentation of which can be found in [LAW 87], and a good example of application to drinking water pipe failures in [RØS 00]. In this same context of pipe failures, the PhD work of [EIS 94] emphasizes the critical importance of past failures. The consideration of the dependency of the failure process on its past is not a trivial question, and motivates a theoretical effort which the present book attempts to contribute to.

The basic concept of a *stochastic process* underlies all developments of the present work. A stochastic process must be understood as a function $X()$ of time $t$, each $X(t)$ being considered as a random variable (r.v.).

The stochastic process theory is the *natural* mathematical framework for studying the repetition of random events of the same kind. As presented by [COO 02], this question can be addressed from two alternative perspectives, which are equivalent and respectively consist of modeling:

– either the distribution of successive inter-arrival times;

– or the distribution of the number of events that occur in a given time interval.

The method chosen by [EIS 94] arises from the first approach. The "classical" presentation of [ROS 83] arises from the second approach. The linear extension of the Yule process (called LEYP throughout the rest of the book) aims at building a failure occurrence model that cumulates the advantages of both NHPP and [EIS 94]'s approaches. This involves a theoretical setup, focused on the *counting process* concept, which is to be developed throughout the next two chapters.

A counting process is a particular stochastic process, simply designed to count repeated events, as presented in section 1.2.1.

As this presentation is to have a general scope, the entity subjected to repeated failures will be called a *technical object* or more simply an *object*; this term will be replaced by "water main" or "water pipe" when the context refers more specifically to failures that affect a water network.

## 1.1. Notation

The following mathematical notations will be used throughout this book:

– $\mathbb{N}$ and $\mathbb{N}^*$ respectively denote the sets of natural integers $\{0, 1, 2, \ldots, \infty\}$ and the set of strictly positive natural integers $\{1, 2, \ldots, \infty\}$;

– $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{R}_+^*$ are the real sets $]-\infty, +\infty[$, $[0, +\infty[$ and $]0, +\infty[$ ;

– $\mathrm{P}(A)$ and $\mathrm{P}(A \mid B)$ respectively denote the probability of the event $A$, and the conditional probability of $A$ given that the other event $B$ occurs;

– $\mathrm{P}(A \cap B)$ and $\mathrm{P}(A, B)$ equivalently denote the joint probability of events $A$ and $B$ ; $\mathrm{P}\left(\bigcap_j A_j\right)$ more generally stands for the joint probability of events $A_j$;

– $t \in \mathbb{R}_+$ is a positive time variable that stands for the age of a technical object;

– $N(t) \in \mathbb{N}$ is an integer-valued step function that counts the failures;

– $\mathrm{d}N(t) = N(t+\mathrm{d}t) - N(t)$ is the differential of $N(t)$, i.e. $\mathrm{d}N(t) = 1$ whenever a failure occurs within $[t, t + \mathrm{d}t[$, $\mathrm{d}N(t) = 0$ otherwise;

– $\Delta N(t) = N(t) - N(t-)$ stands for the increment of $N(t)$ at $t$;

– $\mathcal{N}_{[a,t[}$ stands for the auto-exciting $\sigma$-algebra generated by the process $N(t)$ within $[a, t[$ ;

– $\mathcal{N}_{t-}$ stands for the auto-exciting $\sigma$-algebra $\mathcal{N}_{[0,t[}$;

– $\boldsymbol{Z}$ is a vector of failure factor values specific to a given technical object, also called "covariates";

– $\mathcal{F}_{[a,t[} = \mathcal{N}_{[a,t[} \vee \sigma(\boldsymbol{Z})$ denotes the information on the process $\mathcal{N}_{[a,t[}$ increased by the knowledge of the covariates $\boldsymbol{Z}$, or more technically the smallest $\sigma$-algebra that contains all events composed with events of $\sigma$-algebras $\mathcal{N}_{[a,t[}$ and $\sigma(\boldsymbol{Z})$;

– $\lambda(t)$ is a real positive function bounded on any compact interval, and its integral is $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$;

– $\mathrm{E}X$ and $\mathrm{E}(X \mid A)$ respectively denotes the expectation of the random variable (r. v.) $X$ and its conditional expectation given $A$;

– $\mathrm{Var}(X)$ denotes the variance of the r. v. $X$;

– $\mathcal{U}_E$ stands for the uniform distribution on the set $E$;

– $\mathcal{U}_{[0,1]}$ denotes in particular the uniform distribution on interval $[0, 1]$ ;

– $\mathcal{N}(\mu, \sigma^2)$ stands for the Gaussian distribution with expectation $\mu$ and variance $\sigma^2$;

– $\mathcal{P}o(\mu)$ is the Poisson distribution with expectation $\mu \in \mathbb{R}_+$;

– $\mathcal{NB}(\theta, p)$ is the negative binomial distribution with two parameters $\theta \in \mathbb{R}_+^*$ and $p \in [0, 1]$;

– $\mathcal{NM}(\theta, (p_j)_{j=1,\ldots,n})$ is the negative multinomial distribution with $n + 1$ parameters $\theta \in \mathbb{R}_+^*$ and $p_j \in [0, 1]$;

– $\mathcal{M}(k, (p_j)_{j=1,\ldots,n})$ is the multinomial distribution with $n + 1$ parameters $k \in \mathbb{N}^*$ and $p_j \in [0, 1]$, where $\sum_{j=1}^n p_j = 1$;

– $\chi^2(k)$ is the Chi-squared distribution with $k \in \mathbb{N}^*$ degrees of freedom;

– $L(\boldsymbol{\theta})$ stands for the likelihood of a theoretical process with parameter $\boldsymbol{\theta}$ given a sequence of observed events;

– $\prod\!\!\!\!\int$ stands for the product integral operator, which plays the same role for products as the integral operator $\int$ plays for sums;

– the indicator function $\mathrm{I}(p)$ of proposition $p$ takes value 1 if $p$ is true, 0 otherwise;

– $s \wedge t$ gives the minimum of scalars $s$ and $t$;

– the operator min() gives the minimum of a collection of values;

– the operator max() gives the maximum of a collection of values.

The calculation lines that build up the proof of a proposition will be closed by a right-justified □ symbol. The text lines that express a remark will be typed in italic and closed by a right-justified △ symbol.

## 1.2. General theoretical framework

The theoretical approach adopted throughout this book builds on two essential reference textbooks. The pioneering *Statistical Models Based on Counting Processes* [AND 93], by P.K. Andersen, Ø. Borgan, R.D. Gill and N. Keiding, emphasizes the power of the concepts of the counting process and intensity function to rigorously process survival data. More recently, *Survival and Event History Analysis* [AAL 08], by O.O. Aalen, Ø. Borgan and H.K. Gjessing, explicitly extends the theoretical framework to properly handle recurrent event data.

### 1.2.1. *The concept of a counting process*

We consider a technical object which is observable in continuous time and is likely to undergo events of interest, also called failures, at random times $T_j$, with $j \in \mathbb{N}$ denoting the rank of the failure. The time variable $t$ is measured since the object considered was put into service, i.e. at $t = 0$, and we will often use the terms "time" and "age" indifferently. By convention, the failure time $T_0$ is not random and fixed at the time the object began to be observed, at age 0 or later. The random variable $T_j$ might then be either the age at the first failure, or at the first observed failure. The time interval within which the object is observed will be denoted by $[a, b]$, with $a \in \mathbb{R}_+, b \in \mathbb{R}_+^*$.

As illustrated in Figure 1.1, the counting process $N(t)$ is a right continuous and left-limited integer-valued function that starts at $N(0) = 0$ and increases

by one unit at each $T_j$:

$$\forall t \in \{T_j : j = 1, \ldots, \infty\}, \quad \mathrm{d}N(t) = 1$$

$$\forall t \in ]T_j, T_{j+1}[ : j = 0, \ldots, \infty, \quad \mathrm{d}N(t) = 0$$

It is moreover assumed that at most one failure can occur at a given time, and that the process cannot "explode", i.e. the counting function keeps a finite value at any finite time:

$$\forall t \in \mathbb{R}_+, \begin{cases} \mathrm{P}(\mathrm{d}N(t) > 1) = 0 \\ \mathrm{P}(N(t) < \infty) = 1 \end{cases}$$

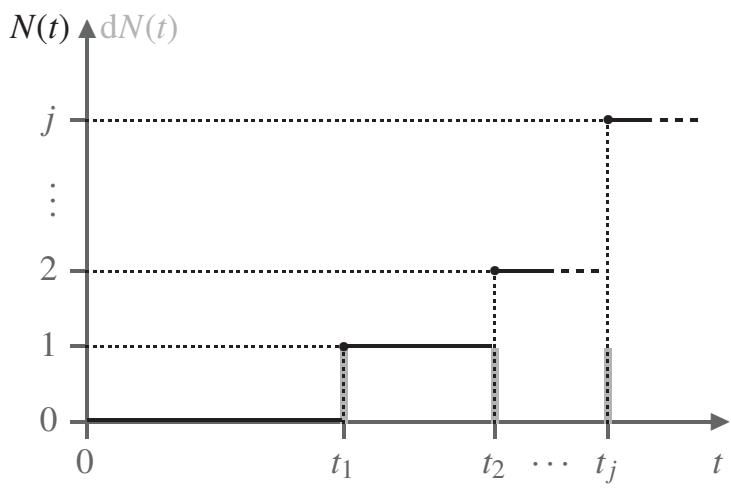


**Figure 1.1.** *Counting process* $N(t)$ *and differential* $\mathrm{d}N(t)$

## 1.2.2. *The intensity function of a counting process*

Let $\mathcal{N}_{[a,t[}$ denote the $\sigma$-algebra $\sigma\left(N(s) - N(a)\right)_{s \in [a,t[}$. Informally called the *past* in [AAL 08], $\mathcal{N}_{[a,t[}$ can be seen as the knowledge available about the process since the beginning of its observation until just before $t$. This information is qualified as *left-truncated* if the failure process is not observed

since the object was put into service ($a > 0$), so nothing is known about the process within $[0, a[$.

The intensity function of $N(t)$, which we will denote by $\eta(t)$, can be heuristically defined as the probability density of a one unit jump at $t$, conditional on the past:

$$\mathrm{P}\left(\mathrm{d}N(t) = 1 \mid \mathscr{N}_{[a,t[}\right) = \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right)$$

Remark 1.1.– It is here to be stressed that the main modeling effort presented in this book has consisted of searching for a parametric form as suitable as possible for $\mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right)$. This conditional expectation assumes an underlying probability distribution for the r.v. $N(t) - N(a) \mid \mathscr{N}_{[a,t[}$, which will generally depend on the parameter denoted by $\boldsymbol{\theta}$; to emphasize the role of $\boldsymbol{\theta}$, the intensity will sometimes be written as $\mathrm{E}_{\boldsymbol{\theta}}\left(\mathrm{d}N(t) \mid \mathscr{N}_{[a,t[}\right)$. △

## 1.3. The non-homogeneous Poisson process

The NHPP model, as presented by [ROS 83], can be defined as:

Definition 1.1.– *The NHPP is defined by the system of equations:*

$$\forall t \in \mathbb{R}_+,\ \begin{cases} N(0) = 0 \\ \mathrm{E}\left(\mathrm{d}N(t) \mid \mathscr{N}_{t-}\right) = \mathrm{E}\left(\mathrm{d}N(t)\right) = \lambda(t)\mathrm{d}t \end{cases}$$

Pivotal properties of NHPP are:

– the intensity depends on age $t$, hence the term *non-homogeneous*;

– $N(t)$ is Poisson distributed with parameter $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$;

– $N(t)$ is *Markovian*, i.e. its distribution does not depend on the trajectory it took between 0 and $t-$.

The particular intensity function $\lambda(t) = \delta t^{\delta-1} e^{\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}}$ is presented by [LAW 87] as tractable for practical use. It is the product of two factors:

– an ageing factor $\delta t^{\delta-1}$, sometimes called *Weibull factor* (see [AAL 08]),

– a scale factor $e^{\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}}$, often called *Cox factor*, for it has initially been proposed by [COX 72].

$\boldsymbol{Z}$ is a vector of explanatory variable values, or *covariates*, which can be either categorical or quantitative, and characterize the technical object or its environment. $\boldsymbol{\beta}$ is a vector of regression coefficients that account for the effects of the covariates on the process intensity. The first components of $\boldsymbol{Z}$ and $\boldsymbol{\beta}$ are respectively 1 and $\beta_0$, and define the *baseline* intensity, when all other covariate values are 0. The exponential form in the Cox factor make covariates act multiplicatively on the intensity, which makes us qualify this form of NHPP as *proportional hazard model* (PHM), sometimes also called *Cox model*.

## 1.4. The Eisenbeis model

In the model of [EIS 94], which from now will be refered to as the *Eisenbeis model*, the successive inter-event times are random variables $X_j = T_j - T_{j-1}$ defined by $\mathbb{R}_+$, which are indexed by the event occurrence rank $j \in \mathbb{N}$, and follow Weibull distributions with parameters $\mu_j$ and $\delta_j$ that depend on $j$. The cumulative distribution function (CDF) of $X_j$ is written as:

$$\forall x \in \mathbb{R}_+, \forall j \in \mathbb{N}, \mathrm{P}\left(X_j \leq x \mid \mu_j, \delta_j\right) = 1 - \exp\left(-x^{\delta_j} e^{\mu_j}\right)$$

The parameter $\mu_j$ is moreover defined as a linear combination $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_j$ of explanatory variables (covariates), which can be either categorical or quantitative, and characterize the technical object or its environment. In the technical context of the Eisenbeis model, water mains are characterized by their diameter, length, location under roadway or sidewalk, type of embedding soil, etc. $\boldsymbol{\beta}_j$ is a parameter vector, specific to event rank $j$. As NHPP, this model is thus also a PHM. The components of vectors $\boldsymbol{Z}$ and $\boldsymbol{\beta}_j$ are indexed by convention from 0 to $q$, where $q$ is the actual number of covariates; a numerical covariate counts indeed for one, whereas a categorical covariate with $m$ possible values counts for $m - 1$ actual covariates (i.e. $m - 1$ indicator variables).

The Eisenbeis model can also be reformulated as the counting process $N(t)$ of the number of events undergone by the object within interval $[0, t]$:

DEFINITION 1.2.– *The Eisenbeis model is defined by the system of equations:*

$$\forall t \in ]T_{j-1}, T_j], \forall j \in \mathbb{N}^*,$$
$$\begin{cases} N(0) = 0 \\ \mathrm{E}\left(\mathrm{d}N(t) \mid N(t-) = j-1\right) = \delta_j (t - T_{N(t-)})^{\delta_j - 1} e^{\mathbf{Z}^{\mathrm{T}} \boldsymbol{\beta}_j} \mathrm{d}t \end{cases}$$

*where by convention* $T_0 = 0$ *at installation of the water main.*

To not have to estimate too many parameters, [EIS 94] proposes to simplify the dependency of $\delta_j$ and $\boldsymbol{\beta}_j$ on $j$ by grouping the values of $j$ into three strata:

– Stratum I for $j \in \{1\}$ ,

– Stratum II for $j \in \{2, 3, 4\}$

– and Stratum III for $j \in \{5, 6, \ldots\}$,

and by fixing also $\delta_{III} = 1$ in the third stratum.

The respective definitions 1.2 and 1.1 of Eisenbeis and NHPP models highlight an essential difference: the intensity of Eisenbeis model strongly depends on the failure rank, whereas the NHPP is mainly driven by the process age. The counting process based on the Eisenbeis model is additionally *not Markovian*, as its distribution depends on the ages at the previous failures.

## 1.5. Other approaches for water pipe failure modeling

There is an extensive amount of international literature devoted to the modeling of repeated water pipe failures. A relevant overview covering publications since 1979 is given by [KLE 01], more recently completed by [BER 08, BUR 10] and [STC 12]. It is to be noticed that, except for the works focused on inter-failure times, the theoretical framework of stochastic processes is never mentioned. This tendency seems to want to last, since most recent publications, such as [DEB 10] and [YAM 09], promote generalized linear models; [DEB 10] considers the occurrence of at least one failure within time intervals of some years as Poisson distributed, whereas [YAM 09] considers shorter time intervals of some months and the binomial distribution.

## 1.6. Why mobilize the Yule process?

Definition 1.2 of the Eisenbeis model involves an important limitation: estimating parameters by means of observed data is only possible provided that the technical objects are observed since their installation; if observation is oppositely restricted to an age interval $[a, b]$ where $a > 0$, event ranks are unknown and the model cannot therefore be applied.

Practical applications, reported by [LEG 00], have however been carried out to get around the left-truncation issue:

– by consenting to consider that $t = 0$ at the beginning of the observation window;

– and by introducing, in log transforms, the age at the previous failure as well as observed failure ranks as covariates.

Results are interesting on the whole, and show an advantage over NHPP in detecting the water pipes that are the most likely to fail. The Eisenbeis model turns out to be an interesting tool for prioritizing water main renovations or replacements. Predictions of future failure numbers have however always included an embarrassing overestimation tendency. The NHPP, on the other hand, poorly detects water pipes likely to fail, but provides unbiased average predictions. Implementing the Eisenbeis model requires moreover time consuming Monte Carlo computations to get around the impossibility of literally calculating the convolution of Weibull distributions. By contrast, NHPP allows very simple and quick prediction computations.

Investigating the use of the Yule process is then fully justified by the search of a model that would combine the advantages of both Eisenbeis and NHPP models, namely a good ability to detect the objects most likely to undergo future failures, and to provide unbiased and easy to compute predictions. The intensity of the searched process should increase both with age and past failures. The idea to exponentially combine distributed inter-arrival times, the parameter of which depends on the event rank, is mentioned by [LEG 01], who refers to Furry distribution (sometimes also known as the Yule–Furry distribution). The work of [PEL 99] is also to be mentioned, which presents a rigorous solution to handle the Eisenbeis model with observations restricted to age intervals $[a, b]$ that do not start at $a = 0$, by explicitly calculating probabilities $\mathrm{P}\left(N(b) - N(a) = m \mid N(a) = j\right)$, and then

their expectation over $j$. The idea to mobilize the Yule process is thus greatly indebted to [LEG 01] and [PEL 99]. The theoretical basis of the Yule process is moreover well presented by [ROS 83].

## 1.7. Structure of the book

After this introductory chapter, Chapter 2 will be devoted to preliminary concepts and tools of probability theory; these preliminaries will particularly concern the binomial and multinomial distributions, the negative binomial and multinomial distributions and power series, and their link with the Yule process. Chapter 3 presents the most general form of non-homogeneous birth process (NHBP), insisting particularly on a general formula for the conditional probability of the number of events within a given time interval $[a, b]$ given the number of events that occurred within interval $[0, a[$. This result is then applied in Chapter 4 to the case where the intensity of NHBP linearly depends on the number of past events, which defined the so-called *linear extension of the Yule process* (LEYP); analytical formulas of the negative binomial probability of the number of events within a time interval given increasingly general past observation interval configurations will be established, as well as a negative multinomial generalization for the joint probability of several time intervals (adjacent and non-overlapping). Chapter 5 will establish the likelihood function of a LEYP process given randomly observed sequences of failures, undergone by technical objects characterized by known covariate values; this result is essential to implement an estimation procedure of the LEYP parameters, for which task the interest of the box-constrained Nelder–Mead optimization algorithm will be emphasized. An important extension of LEYP model will then be presented in Chapter 6, aiming at accounting for the selective survival phenomenon; this arises when the LEYP process is not observed since the object installation, and the objects which can be observed are likely to be the most robust among their cohort. This setup involves considering technical objects with a limited service life, and a decommissioning process that depends on the failure process, and at the same time is susceptible to truncate and censor it. This development gives rise to the so-called *LEYP2s*, the likelihood of which is then studied in Chapter 7; this chapter also presents a numerical validation of LEYP2s model parameter estimation procedure. A case study LEYP2s model application, that uses water network data kindly provided by Lausanne (CH) water utility, is

presented in Chapter 8, and allows us to check the practical interest of such a statistical tool. Chapter 9 concludes the book.

Chapters 7 and 8 involve some computations, either based on random or actual data, which were all carried out by the author for specific illustration purpose of the book. This whole computational work has been implemented in R scripts [RD 11].