Discrete Nonparametric Kernel and Parametric Methods for the Modeling of Pavement Deterioration

Tristan Senga Kiessé, Tristan Lorino, Hussein Khraibani

Discrete Nonparametric Kernel and Parametric Methods for the Modeling of Pavement Deterioration

TRISTAN SENGA KIESSÉ, TRISTAN LORINO, AND HUSSEIN KHRAIBANI

Ifsttar, Bouguenais, France

This article is concerned with one discrete nonparametric kernel and two parametric regression approaches for providing the evolution law of pavement deterioration. The first parametric approach is a survival data analysis method; and the second is a nonlinear mixed-effects model. The nonparametric approach consists of a regression estimator using the discrete associated kernels. Some asymptotic properties of the discrete nonparametric kernel estimator are shown as, in particular, its almost sure consistency. Moreover, two data-driven bandwidth selection methods are also given, with a new theoretical explicit expression of optimal bandwidth provided for this nonparametric estimator. A comparative simulation study is realized with an application of bootstrap methods to a measure of statistical accuracy.

Keywords Bootstrap methods; Discrete associated kernel; Nonparametric regression; Nonlinear regression; Pavement design; Survival data.

1. Introduction

Pavement detoriation models are important inputs for the pavement management systems. These models are based on the analysis of observations of the pavement section condition and provide the evolution law of pavement deterioration. Different statistical models were used to model pavement conditions as classical linear, nonlinear regression models, and Markov Chain. But the limitation of these techniques in terms of discretization and their inefficiency in satisfying the goodness of fit led to use the survival data methodology to determine the evolution law of pavements (Lepert et al., 2003). Among these various investigations, the French Institute of Science and Technology for Transport, Development and Networks (Ifsttar) has been working on developping statistical modeling methods. The most used is a survival data analysis method (SDAM) based on a parametric Weibull model. However, some of the hypotheses in that approach may lead to biased

Address correspondence to Tristan Lorino, Ifsttar, Route de Bouaye CS4, 44344 Bouguenais, France; E-mail: tristan.lorino@ifsttar.fr

estimations of parameters. A recent alternative consists of a logistic mixed model (LMM) which reduces the bias of the estimations by taking into account the correlation between observations on the same pavement sections because of the random effects of this model (Khraibani et al., 2010).

In this article, we propose to complete the two previous parametric approaches with a purely discrete nonparametric one. More precisely, it consists of a nonparametric discrete kernel regression (NDKR) estimator using the associated methodology (Kokonendji et al., 2009; Kokonendji and Senga Kiessé, 2011). The specificity of this approach is the construction of discrete associated kernel estimators which both focus on ordinal variables (consequently, on their discrete realizations) and have the same "nature." They are different from the continuous kernel estimators applied until now on ordinal variables and which treat them as continuous variables without being restricted on their discrete realizations. This NDKR is of interest in terms of providing both discrete estimation and smoothing of fatigue cracking data which are on a discrete support $\mathcal{S}$ (e.g., the set of nonnegative integers $\mathbb{N}$) included in the real line number $\mathbb{R}$. In this approach, the estimation at a point $x \in \mathcal{S}$ is influenced both by the choices of discrete associated kernel and smoothing parameter which allow to take into account the observations in the neighbourhood of x.

The rest of this article is organized as follows. Section 2 briefly presents the survival data analysis method and the logistic mixed model. In Sec. 3, the NDKR method is recalled. Then, the discrete nonparametric regression kernel estimator is presented with some asymptotic properties; in particular, its almost sure convergence is shown. Two methods are also presented for optimal bandwidth choice. In Sec. 4, a comparison on the three approaches is realized based on discrete simulated fatigue cracking data. Section 5 contains the concluding remarks.

2. Parametric Methods

In this section, we recall the survival data analysis method and the nonlinear mixed-effects model without detailing all the mechanisms of these two approaches; for more details, one can refer to Khraibani et al. (2010).

2.1. *Survival Data Analysis Method*

The SDAM presumes that the age T_τ at which a section reaches a deterioration threshold of τ is a random variable that follows a Weibull law characterized by a parametric vector (α, λ) and a corresponding hazard function $h(t, x) = \alpha t^{\alpha-1}\lambda$; the function h is defined as the event rate at time t conditional on survival until time t. The objective of this method is to estimate the parameters α and λ, then to identify the probability law of T_τ. In that purpose, the probability of the realization of all the observations made on each section is calculated: this is the likelihood function denoted by $L(\alpha, \lambda)$.

One particularity of survival data analysis is the presence of a "censoring random variable" reflecting the possible non observation for a given section to reach the threshold exactly at the time t.

Hence, the likelihood function $L(\alpha, \lambda)$ is obtained as

$$L(\alpha, \lambda) = F(T_i)^{\delta_{1i}} S(T_i)^{\delta_{2i}} \{F(T_{1i}) - F(T_{2i})\}^{\delta_{3i}} f(T_i)^{\delta_{4i}}$$

by combining the contributions of the failure time through the probability density function $f(T)$; the left-censored observation reflected by the cumulative distribution function $F(T) = \Pr(T < t)$; the right-censored observation through the survival function $S(T) = 1 - F(t)$; and, the interval-censored observation through the factor $F(T_2) - F(T_1)$ with δ_{pi}, $p = 1, 2, 3, 4$, are indicative functions of the types of events observed. Then this analysis is repeated for all the evolution thresholds τ standing between 0% and 100%, by 5% increments. Finally, to determine the evolution law of a specific section, we use the notion of robustness; this notion states that if a given section $sec = 1, 2, \ldots, k, \ldots, M$, with $M \in \mathbb{N}\backslash\{0\}$, has evolved more quickly that k other at a given age, then this section will always evolve more quickly than the other k sections.

2.2. *Logistic Mixed Model*

The nonlinear mixed-effects (NLME) framework is widely used for describing nonlinear relationship between a response variable and parameters and covariates in the repeated measurements data that are grouped according to a cluster factor. The NLME models were initially proposed in biostatistics literature; here, we follow the generalized NLME models proposed by Lindstrom and Bates (1990).

For pavement section sec with $n^{(sec)}$ repeated measurements, $sec = 1, 2, \ldots, M$, $M \in \mathbb{N}\backslash\{0\}$, the generalized NLME model for pavement cracking data can be expressed as

$$y_{sec,j} = m(\beta_{sec}; a_{sec,j}, x_{sec,j}) + e_{sec,j}, \quad j = 1, 2, \ldots, n^{(sec)},$$

where $y_{sec,j}$ is the measured value of the deterioration at time j; $a_{sec,j}$ denotes the age (in years) on time j, m is the nonlinear function relating the response variable to age and to the other possible covariates $x_{sec,j}$ varying with each section and time, β_{sec} is a vector with the parameters of nonlinear function, and $e_{sec,j}$ is a normally distributed within section error term.

The form adopted for predicting the cracking measurements is assumed to be sigmoïdal. Therefore, the pavement sections can be described by the logistic model:

$$y_{sec,j} = \frac{\theta_1}{1 + \exp\{-(\frac{t_{sec,j} - \theta_2}{\theta_3})\}} + e_{sec,j}, \tag{1}$$

where $y_{sec,j}$ is the percentage of cracking for the sec-th section at the jth measurement time $t_{sec,j}$, $sec = 1, 2, \ldots, M$ with $M \in \mathbb{N}\backslash\{0\}$, $j = 1, 2, \ldots, n^{(sec)}$; and $e_{sec,j}$ represents an independent and identically normally distributed within section error term with zero mean and variance σ^2. The parameter θ_1 corresponds to the value of the limit of cracking growth at which roads will be completely degraded and is set to 100% of cracking. The parameter θ_2 is the midpoint, the time at which $y_{sec,j} = \theta_1/2 = 50\%$. The parameter θ_3 is the scale parameter and represents the distance on the time axis between the midpoint and the point where the response is $\theta_1/(1 + e^{-1}) = 73\%$ of cracking (see Fig. 1).

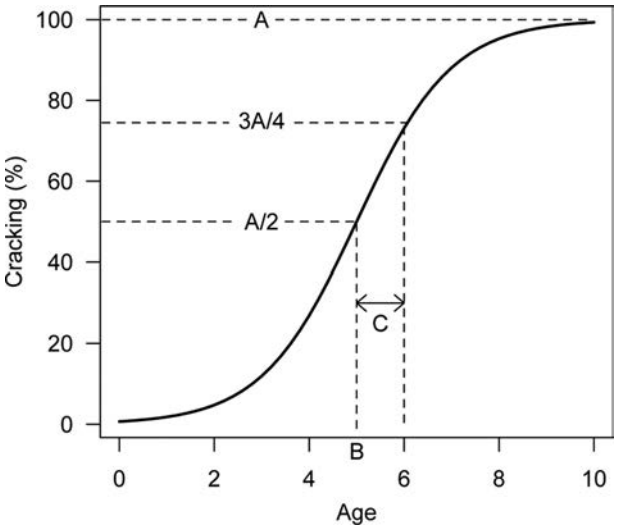

Figure 1. Growth curve of logistic model.

In order to account for the variation on the same pavement section, random components were intoduced into model (1) yielding the following nonlinear mixed-effects model:

$$y_{sec,j} = \frac{100}{1 + \exp[-\{\frac{t_{sec,j} - (\theta_2 + b_{sec})}{(\theta_3 + c_{sec})}\}]} + e_{sec,j}, \tag{2}$$

where we have the assumptions $(b_{sec},\ c_{sec}) \sim\ \mathcal{N}(0, \sigma^2_{sec})$, $e_{sec,j} \sim \mathcal{N}(0, \sigma^2)$ and $b_1, b_2, \ldots, b_M,\ c_1, c_2, \ldots, c_M,\ e_1, e_2, \ldots, e_M$, respectively, are independent (with M the number of section). The parameter θ_2 is replaced by $\theta_2 + b_{sec}$ to account for the correlation between observations on the intra-individual variability in the midpoint time. The parameter θ_2 is called the fixed effect, b_{sec} is called the random effect and represents the individual section departure from the average time of the midpoint. Similarly, the fixed effect θ_3 represents the mean level of the growth time for the population and c_{sec} is the individual section departure from the mean level of the growth time. In the application, we will fit data with the model (2), assuming that the random effects are added to the formula.

3. Nonparametric Discrete Kernel Regression

In this section we present the discrete associated kernel methodology introduced by Kokonendji and Senga Kiessé (2011); then we also provide the NDKR estimator (Kokonendji et al., 2009). Some asymptotic properties and the bandwidth optimal choice are studied for this estimator.

3.1. *Discrete Associated Kernel*

Let us recall some notions about the discrete associated kernel approach. First, the kernel $K_{x,h}(\cdot)$ is a pmf with support $\mathcal{S}_x$ which contains x and does not depend on h such as

H1. $\lim_{h\to 0} \mathrm{E}(\mathcal{K}_{x,h}) = x$,
H2. $\lim_{h\to 0} \mathrm{Var}(\mathcal{K}_{x,h}) = 0$,

with $\mathcal{K}_{x,h}$ the discrete r.v. of pmf $K_{x,h}(\cdot)$. Here, we propose the following new assumptions less general than H1–H2:

H3. $\Pr(\mathcal{K}_{x,h} = x) = 1 - h\mathrm{U}(\mathcal{K}_{x,h}) + O(h^2)$,
H4. $\mathrm{Var}(\mathcal{K}_{x,h}) = h\mathrm{V}(\mathcal{K}_{x,h}) + O(h^2)$,

with $\sum_{y\in\mathcal{S}_x\setminus\{x\}} \Pr(\mathcal{K}_{x,h} = y) = h\mathrm{U}(\mathcal{K}_{x,h}) + O(h^2) \to 0$ as $h \to 0$; one can verify that H3–H4 lead to H1–H2. These new assumptions can be useful both for specifying the rate of convergence of $\mathcal{K}_{x,h}$ to x and explaining an explicit expression of optimal bandwidth. Note that the notations "$\mathrm{U}(\mathcal{K}_{x,h})$" and "$\mathrm{V}(\mathcal{K}_{x,h})$" reflects that U and V are connected to $\mathcal{K}_{x,h}$ but do not obligatory depend on x and h as we will see in the following example from Kokonendji and Senga Kiessé (2011). Now we give a class of symmetric discrete associated kernels $K_{x,h}(\cdot)$ that fullfill assumptions H3–H4 (and, consequently, H1–H2).

Example 3.1. Let a be a fixed integer and $h > 0$ be a smoothing parameter. For any fixed $x \in \mathcal{S} = \mathbb{N}$, consider the discrete r.v. $\mathcal{DT}_{a;x,h}$ defined on support $\mathcal{S}_{a,x} = \{x, x \pm 1, \ldots, x \pm a\}$ and whose pmf is given by

$$K_{a;x,h}(y) = \frac{(a+1)^h - |y-x|^h}{A(a,h)}, \quad \forall y \in \mathcal{S}_{a;x},$$

where $A(a,h) = (2a+1)(a+1)^h - 2\sum_{k=1}^{a} k^h$ is the normalizing constant. We have the mean $\mathbb{E}(\mathcal{DT}_{a;x,h}) = x$ and the variance

$$\mathrm{Var}(\mathcal{DT}_{a;x,h}) = \frac{1}{A(a,h)} \left\{ \frac{a(2a+1)(a+1)^{h+1}}{3} - 2\sum_{k=1}^{a} k^{h+2} \right\} = V(a,h)$$

which does not depend on x and tends to 0 when $h \to 0$. The R package for symmetric discrete triangular distributions is available Senga Kiessé et al. (2009).

First, for h sufficiently small, the modal probability of $\mathcal{DT}_{a;x,h}$ can be approximate by

$$\begin{aligned}
\Pr(\mathcal{DT}_{a;x,h} = x) &= \frac{(a+1)^h}{A(a,h)} \\
&\simeq \frac{1 + h\log(a+1)}{1 + h\{(2a+1)\log(a+1) - 2\sum_{k=1}^{a}\log(k)\}} \\
&= 1 - h\left\{2a\log(a+1) - 2\sum_{k=1}^{a}\log(k)\right\} + O(h^2) \\
&= 1 - 2hU(a) + O(h^2).
\end{aligned}$$

Furthermore, we have the following expansion for the variance of $\mathscr{DT}_{a;x,h}$:

$$\begin{aligned}\mathrm{Var}(\mathscr{DT}_{a;x,h}) &\simeq \left[\frac{a(a+1)(2a+1)}{3}\{1+h\log(a+1)\} - 2\sum_{k=1}^{a} k^2\{1+h\log(k)\}\right] \\ &\quad \times \left[1+h\left\{(2a+1)\log(a+1) - 2\sum_{k=1}^{a}\log(k)\right\}\right]^{-1} \\ &= \left\{\frac{a(2a^2+3a+1)}{3}\log(a+1) - 2\sum_{k=1}^{a} k^2\log(k)\right\} h + O(h^2) \\ &= 2hV(a) + O(h^2),\end{aligned}$$

which also holds for h sufficiently small. Note that U and V depend on parameter a but not on x and h as said previously.

Remark 3.1. The class of standard discrete kernels built from usual discrete probability distributions (Poisson, binomial, and negative binomial) does not fulfill hypotheses H3–H4 but only H1. For all $x \in \mathbb{N}$ and $h > 0$, the modal probability at x of these kernels does not tend to 1 as h goes to 0; moreover, the r.v. associated to these discrete kernels satisfies $\lim_{h\to 0}\mathrm{Var}(\mathscr{K}_{x,h}) \in \mathscr{V}(0)$ instead of H2 where $\mathscr{V}(0)$ is a neighbourhood of 0. See Kokonendji and Senga Kiessé (2011) for more details on these discrete kernels.

3.2. *Nonparametric Regression Estimator*

Consider the nonparametric regression model of y_i on x_i,

$$y_i = m(x_i) + e_i,$$

where y_i denotes the observation of the response variable Y_i in $\mathbb{R}$, x_i is the realization of the explanatory variable X_i in $\mathscr{S} = \mathbb{N}$, e_i is assumed to be the residual from the real random variable ϵ_i satisfying commonly $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$, and $m : \mathbb{N} \mapsto \mathbb{R}$ is an unknown discrete bounded regression function such as $m(x_i) = \mathbb{E}(Y_i|X_i = x_i)$ and $\mathrm{Var}(Y_i|X_i = x_i) < \infty$. The NDKR estimator of Nadaraya-Watson type for m is proposed by Kokonendji and Senga Kiessé (2011) as

$$\tilde{m}_n(x) = \sum_{i=1}^{n} \frac{Y_i K_{x,h}(X_i)}{\sum_{j=1}^{n} K_{x,h}(X_j)}, \quad x \in \mathbb{N}. \tag{3}$$

In the following we study some asymptotic properties of NDKR estimator $\tilde{m}_n$.

3.2.1. *Asymptotic Properties.* First, for any fixed $x \in \mathbb{N}$, assume that the discrete r.v. X has a pmf such that $f(x) = \Pr(X = x) > 0$. Under assumptions H1–H2, the pointwise bias and variance of the estimator $\tilde{m}_n(x)$ are given in Kokonendji et al. (2009) and Kokonendji and Senga Kiessé (2011), depending on the modal probability $\Pr(\mathscr{K}_{x,h} = x)$ and variance $\mathrm{Var}(\mathscr{K}_{x,h})$. Therefore, by taking into account

the hypotheses H3–H4 we have

$$\text{Bias}\{\tilde{m}_n(x)\} = \frac{h}{2}\left\{m^{(2)}(x) + 2m^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right\} V(\mathcal{K}_{x,h}) + O\left(\frac{1}{n} + h^2\right) + o(h), \quad (4)$$

$$\text{Var}\{\tilde{m}_n(x)\} = \frac{\text{Var}(Y|X=x)}{nf(x)}\{1 - hU(\mathcal{K}_{x,h})\}^2 + o\left(\frac{1}{n}\right) + O(h^2), \quad (5)$$

where $f^{(1)}$, $m^{(1)}$, and $m^{(2)}$ are finite differences used instead of the usual differentiation on the real line numbers $\mathbb{R}$. Then, the pointwise consistency of the discrete semiparametric estimator $\tilde{m}_n$ is obtained through the asymptotic behaviour of its mean square error (MSE) as

$$\text{MSE}(x, n, h, K, f) = \text{Bias}^2\{\tilde{m}_n(x)\} + \text{Var}\{\tilde{m}_n(x)\} \to 0, \quad x \in \mathbb{N}.$$

Indeed, under assumptions H1–H4 and $\text{Var}(Y|X=x) < \infty$, the asymptotic expansions of the bias and variance of $\tilde{m}_n$ are such as $\text{Bias}\{\tilde{m}_n(x)\} \to 0$ and $\text{Var}\{\tilde{m}_n(x)\} \to 0$, as $h = h(n) \to 0$ when $n \to \infty$. In addition, the global consistency of $\tilde{m}_n$ such as

$$\text{MISE}(n, h, K, f) = \sum_{x \in \mathbb{N}} \text{MSE}(x) \to 0 \quad \text{when } h = h(n) \to 0 \quad \text{and} \quad n \to \infty \quad (6)$$

comes from adding the condition

$$\sum_{x \in \mathbb{N}} \left| \left\{m^{(2)}(x) + 2m^{(1)}(x)\left(f^{(1)}/f\right)(x)\right\} \right| < \infty$$

to the previous assumptions.

In what follows we present the theorem on almost sure consistency of the NDKR estimator $\tilde{m}_n$ in (3), to this we need first to recall the following lemma (Hoeffding, 1963).

Lemma 3.1. *Let $Z_1, Z_2, \ldots, Z_n$ be i.i.d. random variables with finite second moments. If there exist constants a and b such that $Pr(Z_i \in [a, b]) = 1$, then given $\epsilon > 0$ we have*

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \geq \epsilon\right) \leq 2\exp\left\{-\frac{n\epsilon^2}{\epsilon(b-a) + 2Var(Z_1)}\right\}.$$

Now we present the theorem on almost sure consistency of $\tilde{m}_n$.

Theorem 3.1. *For any fixed $x \in \mathbb{N}$, under assumptions H1–H2, the nonparametric estimator $\tilde{m}_n(x)$ converges almost surely to $m(x)$ as follows*:

$$\tilde{m}_n(x) \xrightarrow{a.s.} m(x).$$

The notation "$\xrightarrow{a.s.}$" stands for "almost sure convergence."

Proof. Let us write the nonparametric regression estimator as $\tilde{m}_n = \tilde{g}_n(x)/\tilde{f}_n(x)$ where

$$\tilde{g}_n(x) = \frac{1}{n}\sum_{i=1}^{n} W_{x,h}(X_i)Y_i \quad \text{and} \quad \tilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} W_{x,h}(X_i)$$

with $W_{x,h}(X_i) = K_{x,h}(X_i)/\mathbb{E}\{K_{x,h}(X_1)\}$ such as $\mathbb{E}\{W_{x,h}(X_i)\} = 1$. The estimator $\tilde{m}_n$ can be expressed as:

$$\tilde{m}_n(x) - m(x) = \frac{\tilde{g}_n(x) - \mathbb{E}\{\tilde{g}_n(x)\}}{\tilde{f}_n(x)} - [m(x) - \mathbb{E}\{\tilde{g}_n(x)\}] - \frac{m(x)}{\tilde{f}_n(x)}\{\tilde{f}_n(x) - 1\}.$$

Firstly, we have

$$\begin{aligned} m(x) - \mathbb{E}\{\tilde{g}_n(x)\} &= m(x) - \mathbb{E}[\mathbb{E}\{W_{x,h}(X_i)Y_i\}|X_i] \\ &= m(x)\mathbb{E}\{W_{x,h}(X_i)\} - \mathbb{E}\{W_{x,h}(X_i)m(X_i)\} \\ &= \mathbb{E}[\{m(x) - m(X)\}W_{x,h}(X_i)]. \end{aligned}$$

The continuity of $m : \mathbb{N} \to \mathbb{R}$ at $x \in \mathbb{N}$ can be defined as follows:

$$\forall\, \epsilon > 0, \quad \exists\, \eta > 0 : \forall\, y \in]x - \eta, x + \eta[\cap\mathbb{N} \Rightarrow |m(y) - m(x)| < \epsilon; \tag{7}$$

we easily deduce that any crf is continuous in the sense of this definition. Note that, for $\eta > 0$ in (7), the notion of discrete neighbourhood $]x - \eta, x + \eta[\cap\mathbb{N}$ of x can be reduced to the single point $\{x\}$. The continuity of the crf m results in $m(x) - \mathbb{E}\{\tilde{g}_n(x)\} \to 0$.

Secondly, let us write $\tilde{g}_n(x) - \mathbb{E}\{\tilde{g}_n(x)\} = (1/n)\sum_{i=1}^{n} Z_i$ with $Z_i = W_{x,h}(X_i)Y_i - \mathbb{E}\{W_{x,h}(X_i)Y_i\}$. For any $x \in \mathbb{N}$, there exists $0 < M < \infty$ such that we have $|Z_i| \leq |Y_i| \leq |M|$ then successively

$$\begin{aligned} \mathrm{Var}(Z_i) &= \mathrm{Var}\{W_{x,h}(X_i)Y_i\} \\ &\leq \mathbb{E}\{W_{x,h}^2(X_i)Y_i^2\} \\ &\leq \mathbb{E}\{Y_i^2\} < \infty, \end{aligned}$$

since $0 \leq W_{x,h}(\cdot) \leq 1$ and $m(x)$ is bounded with $\mathrm{Var}(Y|X = x) < \infty$. Therefore, according to Lemma, one has

$$\Pr\left[|\tilde{g}_n(x) - \mathbb{E}\{\tilde{g}_n(x)\}| \geq \epsilon\right] = \Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-n\epsilon^2}{2\epsilon M + 2}\right) \leq 2\exp\left(\frac{-n\epsilon}{2}\right),$$

for any $\epsilon > 0$. Consequently, Borel-Cantelli lemma leads to get $\tilde{g}_n(x) - \mathbb{E}\{\tilde{g}_n(x)\} \xrightarrow{a.s.} 0$ since $\sum_{n\geq 1} \Pr\left[|\tilde{g}_n(x) - \mathbb{E}\{\tilde{g}_n(x)\}| \geq \epsilon\right] < \infty$.

Thirdly, we have $\tilde{f}_n(x) - 1 \xrightarrow{a.s.} 0$. For the demonstration, we just express $\tilde{f}_n(x) - 1 = (1/n)\sum_{i=1}^{n} Z_i'$ with $Z_i' = K_{x,h}(X_i) - 1$. It comes that

$$-1 \leq Z_i' \leq 0 \text{ and } \mathrm{Var}(Z_i') \leq \mathbb{E}\{K_{x,h}^2(X_1)\} \leq 1,$$

then the rest of the demonstration for this third part is similar to the second part.

Hence, the theorem is shown.

3.2.2. *Bandwidth Optimal Selection.* Here, we investigate two approaches for selecting the optimal parameter.

Minimization of Mean Integrated Squared Error. The first possibility is the minimization of MISE of $\tilde{m}_n$ in (6) that we obtain from the bias and the variance in (4) and (5), respectively. Then, the differentiation of the approximate MISE (AMISE) with respect to h is such as

$$\frac{d}{dh}\text{AMISE} = -\frac{1}{n}\sum_{x\in\mathbb{N}}\frac{\text{Var}(Y|X=x)}{f(x)}U(\mathcal{K}_{x,h}) - hU^2(\mathcal{K}_{x,h})\}$$
$$+\frac{h}{4}\sum_{x\in\mathbb{N}}\left[\left\{m^{(2)}(x)+2m^{(1)}(x)\left(\frac{f^{(1)}}{f}\right)(x)\right\}V(\mathcal{K}_{x,h})\right]^2.$$

From the equation $d/dh(AMISE) = 0$, it ensues the explicit expression

$$h_{opt}(n) = \frac{\sum_{x\in\mathbb{N}}\text{Var}(Y|X=x)U(\mathcal{K}_{x,h})/f(x)}{\sum_{x\in\mathbb{N}}\text{Var}(Y|X=x)U^2(\mathcal{K}_{x,h})/f(x) + (n/4)\sum_{x\in\mathbb{N}}\{W(x)\times V(\mathcal{K}_{x,h})\}^2}$$

with $\sum_{x\in\mathbb{N}}W^2(x) = \sum_{x\in\mathbb{N}}\left\{m^{(2)}(x)+2m^{(1)}(x)\left(f^{(1)}/f\right)(x)\right\}^2 < \infty$ and $h_{opt}(n)\to 0$ as $n\to\infty$.

However, this explicit expression h_{opt} cannot be always used because it requires to knwon the true functions m and f. In the following we present an alternative approach to find an optimal h-value.

Cross-validation Procedure. The selection of the optimal bandwidth parameter also can be realized by the cross-validation method. For a given discrete associated kernel, let us write the nonparametric count regression estimator $\tilde{m}_n$ of the crf m as

$$\tilde{m}_n(x) = \sum_{i=1}^{n}\omega_{x,h}(X_i)Y_i$$

with $\omega_{x,h}(X_i) = K_{x,h}(X_i)/\sum_{j=1}^{n}K_{x,h}(X_j)$. The optimal smoothing parameter is $h_{cv} = \arg\min_{h>0}CV(h)$ with

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\tilde{m}^2_{n,-i}(X_i;h) - \frac{2}{n}\sum_{i=1}^{n}\tilde{m}_{n,-i}(X_i;h)Y_i,$$

where

$$\tilde{m}_{n,-i}(X_i;h) = \sum_{j\neq i}^{n}\frac{Y_jK_{X_i,h}(X_j)}{\sum_{j\neq i}^{n}K_{X_i,h}(X_j)} = \sum_{j\neq i}^{n}\omega_{X_i,h}(X_j)Y_j$$

is the leave-one-out kernel estimator of $\tilde{m}_n(X_i;h)$ being computed by excluding X_i (Kokonendji et al., 2009).

4. Illustrations

In this section, the SDAM, LMM, and NDKR associated kernel regression are illustrated on simulated performance data. The data set was generated using a fatigue-crack propagation model whose random parameters follow a Weibull probability distribution. These simulated data contained $M = 100$ maintained

sections. For each section, 70–134 measurements varying from 0–100% were considered. In order to evaluate the performance of the different models we compare the root mean squared error *RMSE* defined as

$$RMSE^{(sec)} = \sqrt{\frac{\sum_{j=1}^{n^{(sec)}} (y_{sec,j} - \hat{y}_{sec,j})^2}{n^{(sec)}}}, \tag{8}$$

where $\hat{y}_{sec,j}$ is the ajustement of the j-th observation $y_{sec,j}$ of percentage of cracking for the sec-th section with a number $n^{(sec)}$ of repeated measurements, $sec = 1, 2, \ldots, 100$. For nonparametric estimator, the bandwidth optimal choice is realized using cross-validation procedure; we omit here to present the optimal h-value obtained. All computations are done with the R statistical language (Senga Kiessé et al., 2009).

4.1. *Global Study*

The illustrations are realized on the sections $sec \in \{1, 13, 30, 41\}$ that some characteristics are given in Table 1.

Table 2 presents the values of the $RMSE^{(sec)}$ resulting from the application of each model on the sections $sec \in \{1, 13, 30, 41\}$. In addition, Fig. 2 provides graphic illustrations corresponding to sections 1 and 30 which are taken as examples, because the sections 1, 13, and 41 have closed characteristics (see Table 1). Looking at these examples, the discrete nonparametric kernel regression estimator $\tilde{m}_n$ provides the discrete estimations whom are closest to observations. Thus, in terms of goodness of fit, the estimator $\tilde{m}_n$ provides the best estimations followed by the LMM and, at last, the SDAM.

Note that the performance of the parametric models is slightly improved by introducing covariables but not sufficiently to make them better than the NDKR. Indeed, the order of performance between the three approaches is not changed in this case that we do not present here.

The calculation of individual $RMSE^{(sec)}$ for a section is completed with a more robust evaluation using the Monte-Carlo simulations (see Table 3 and Fig. 3). It consists of a bootstrap approximate $\widehat{RMSE}^{(sec)}$ of the $RMSE^{(sec)}$ obtained by resampling the observations of the section sec chosen. Indeed, from the actual sample $y_1, y_2, \ldots, y_{n^{(sec)}}$, of a section, we draw with replacement N random samples $y_1^*, y_2^*, \ldots, y_{n^{(sec)}}^*$, of the same size $n^{(sec)}$. Thus, we obtain N bootstrap samples on whom we apply SDAM, LMM, and the regression estimator $\tilde{m}_n$ in (3). From the

Table 1
Some characteristics of sections

Section sec	$n^{(sec)}$	Mean	Variance	Coefficient of variation
1	71	0.524	0.135	0.702
13	72	0.528	0.134	0.692
30	134	0.565	0.140	0.662
41	70	0.521	0.133	0.700

Table 2
RMSE (in %) coming from fitting some pavement sections

	$\text{RMSE}^{(sec)}$		
Section	SDAM	LMM	NDKR*
1	2.25	1.14	0.231
13	2.42	1.22	0.084
30	4.02	0.93	0.224
41	2.46	1.17	0.271

*With $a = 2$

adjustement of each simulated sample $i = 1, 2, \ldots, N$, of the fixed section, it ensues the corresponding $RMSE_i^{(*)}$ and the calculation of the average:

$$\overline{RMSE^{(*)}} = \frac{1}{N}\sum_{i=1}^{N} RMSE_i^{(*)} =: \widehat{RMSE}^{(sec)}$$

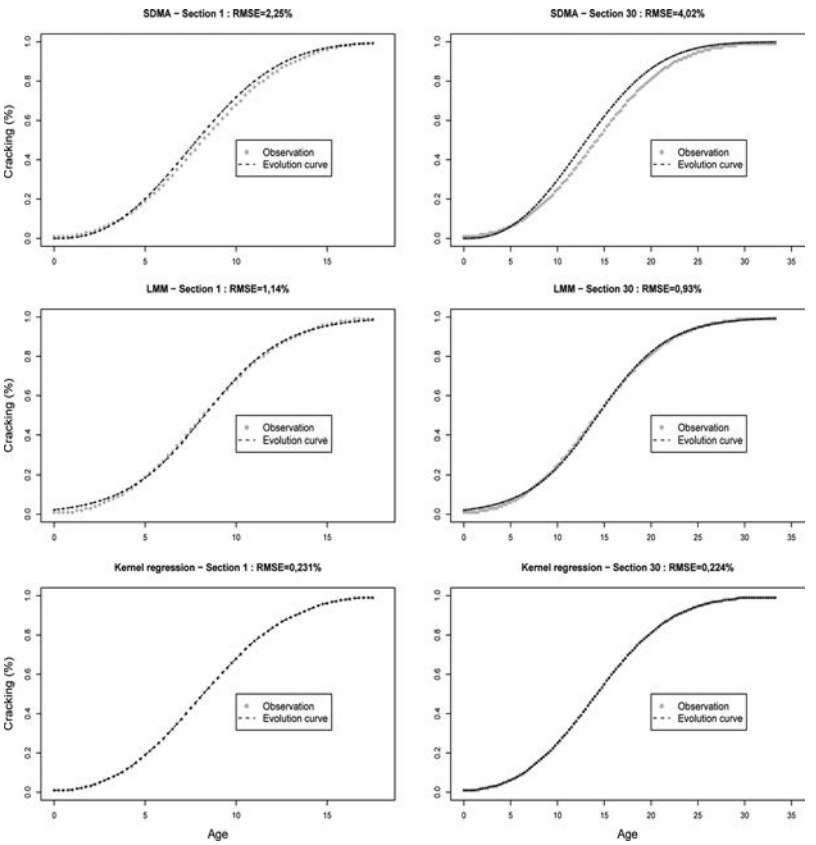

Figure 2. Comparison of evolution curves using SDAM, LMM, and discrete tringular kernel regression from simulated fatigue data of pavement sections 1 and 30.

Table 3
Means of RMSE (in %) calculated from bootstrap samples of simulated fatigue data

		$\widehat{RMSE}^{(sec)}$		
Section	N	SDAM	LMM	NDKR*
	100	4.49	3.01	0.379
1	300	4.39	2.87	0.399
	600	4.45	2.86	0.383
	100	4.76	2.93	0.390
13	300	4.43	2.93	0.373
	600	4.36	2.89	0.373
	100	3.29	2.15	0.216
30	300	3.21	2.14	0.214
	600	3.25	2.11	0.203
	100	4.71	2.95	0.393
41	300	4.72	2.89	0.378
	600	4.80	2.94	0.397

*With $a = 2$

which converges as N is increasing; see Efron and Tibshirani (1986) for bootstrap methods. All the results coming from the application of the SDAM, LMM, and discrete nonparametric regression estimator $\tilde{m}_n$ on the 100 sections are not presented here. Finally, the evolution curves are obtained using the linear interpolation procedure between the discrete points.

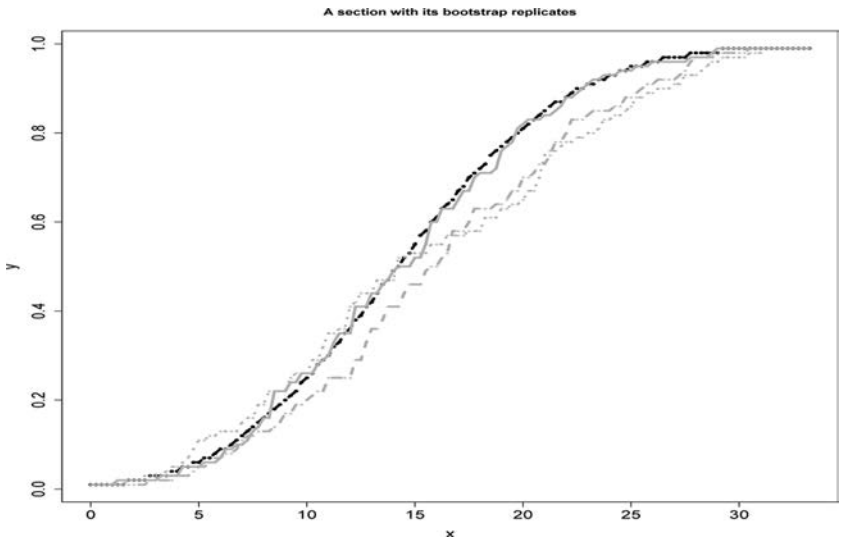

Figure 3. An example of section (in black with dash lines) with its boostrap replicates (in grey).

Table 4
Some characteristics of sections 1 and 30 with respect to the age (in years)

Section	Age	Mean	Variance	Coefficient of variation
	0–5	0.067	0.003	0.859
1	5–15	0.638	0.062	0.391
	>15	0.980	0.011×10^{-2}	0.011
	0–5	0.027	0.029×10^{-2}	0.629
30	5–25	0.529	0.087	0.559
	>25	0.979	0.018×10^{-2}	0.014

4.2. *Study with Respect to the Age*

Here, the bootstrap samples generated from the sections 1 and 30 are studied by three parts with respect to the age, then the average error $\overline{RMSE}^{(*)}$ is calculated on each part. These parts are defined such that they correspond to similar behaviors of fatigue cracking data of the two chosen sections. Thus, the first part corresponds

Table 5
Means of RMSE (in %) calculated from bootstrap samples of simulated fatigue data

		$\widehat{RMSE}^{(sec)}$		
Age*	N	SDAM	LMM	NDKR**
		Section 1		
	100	2.80	2.50	0.265
0–5	300	2.80	2.49	0.261
	600	2.88	2.52	0.260
	100	5.25	3.12	0.500
5–15	300	5.36	3.19	0.447
	600	5.42	3.17	0.471
	100	1.07	1.44	0.142
>15	300	1.19	1.40	0.114
	600	1.18	1.39	0.116
		Section 30		
	100	1.61	1.95	0.102
0–5	300	1.51	1.87	0.488
	600	1.55	1.81	0.102
	100	3.95	2.45	0.258
5–25	300	3.95	2.42	0.256
	600	3.99	2.41	0.255
	100	0.92	0.88	0.075
>25	300	0.95	0.88	0.411
	600	0.93	0.91	0.070

*Age in years. **With $a = 2$

to small fatigue cracking values (about <20%) ang goes from 0–5 years for the two sections. At the opposite, the last part corresponds to large fatigue cracking values (about >90%) and is over 15 years for the section 1 and 25 years for the section 30. In the middle, there is an intermediary part from 5–15 years for the section 1, and from 5–25 years for the section 30. Some characteristics of the fixed sections with respect to the age are given in Table 4.

By applying each model, the estimations are better on the section parts corresponding to small and large fatigue cracking data in comparison with the intermediary part (refer to Table 5). Then, looking at the performances of the three models, the nonparametric kernel regression estimator outperforms the two other models on each of the three parts. About the LMM and SDAM, they are closed in mean (in the sense of the RMSE) for small and large fatigue cracking values which correspond to a small data dispersion (Table 4). The LMM is better than the SDAM on the middle part where both the mean and the dispersion of the data are largest. These results confirm the global perfomance of these three approaches, as it has been arleady pointed out in the previous paragraph. Moreover, through the varying results between the middle part and the two other parts, this study seems pointed out that the performance of each model and the dispersion of the data are connected.

5. Discussion

This article is concerned with a discrete associated kernel approach in comparison to a survival data analysis method and a logistic regression for the modelisation of discrete pavement condition data. A comparative simulation study is provided leading to the following concluding remarks.

The purely discrete nonparametric approach outperforms the two parametric approaches, in term of fitting the discrete observed data of pavement condition. Indeed, the nonparametric estimation techniques are well known to be impartial to special types of the underlying density function. Moreover, the nonparametric associated kernel regression estimator takes into account the correlation between the observations on a same section through both the behaviour of the associated kernel and the role of the smoothing parameter. However, there are some limitations to this discrete associated kernel method as the fact that we can not expect it to involve covariates.

The parametric approaches may have the advantages both to include ordinal covariables and examine their influence in the modelisation of the phenomenon, which may lead to more detailed analysis. In addition, some confidence intervals of the parameters are provided as in the LMM. In these approaches some various tests are also available to evaluate their accuracy in terms of fitting and prediction. At last, the parametric approaches, based on some known probability density functions (e.g., Weibull, logistic) contrary to discrete nonparametric kernel procedure, are useful to detail the "true form" of the underlying function to estimate.

Finally, the nonparametric and parametric procedures are complementary and can be used depending on the researched purpose of the user. In perspective, it would be interesting to investigate the NDKR results depending on the choice of discrete kernel and of bandwidth. Moreover, a generalized linear mixed model or a discrete semiparametric procedure recently proposed by Abdous et al. (2012) might be compared to NDKR. In this way, some parametric, nonparametric, and

semiparametric procedures will be available for fitting pavement deterioration, and using diagnostic checks will allow to choose the appropriate approach according to the data set. Some works are in progress in this direction.

References

Abdous, B., Kokonendji, C. C., Senga Kiessé, T. (2012). On semiparametric regression for count explanatory variables. *J. Statist. Plan. Infer*. 142:1537–1548.

Efron, B., Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* 1:54–77.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* 58:13–30.

Khraibani, H., Lorino, T. Lepert, Ph., Marion, J.-M. (2010). Non linear mixed effects model for the evaluation and prediction of pavement deterioration. *J. Transport. Eng*. 138:149–156.

Kokonendji, C. C., Senga Kiessé, T., Demétrio, C. G. B. (2009). Appropriate kernel regression on a count explanatory variable and applications. *Adv. Applic. Statist.* 12:99–125.

Kokonendji, C. C., Senga Kiessé, T. (2011). Discrete associated kernels method and extensions. *Statist. Methodol.* 8:497–516.

Lepert, Ph., Leroux, D., Savard, Y. (2003). Use of pavement performance models to improve efficiency of data collection procedures. *3rd Int. Symp. Maint. Rehab. Pavements Technologi. Control*. University of Minho, Guimaraes, Portugal.

Lindstrom, M. J., Bates, D. M. (1990). Nonlinear mixed-effects models for repeated measures data. *Biometrics* 46:673–687.

R Development Core Team. (2008). *A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at http://www.r-project.org.

Senga Kiessé, T., Zocchi, S. S., Kokonendji, C. C. (2009). The R package for discrete triangular distributions. Version 1.2R. Available via http://cran.r-project.org.