On Modeling Wood Formation Using Parametric and Semiparametric Regressions for Count Data

HENRI E. CUNY & TRISTAN SENGÀ KIESSÉ

1INRA, UMR1092, Laboratoire d’Etude des Ressources Forêt Bois (LERFoB), Centre INRA de Nancy, Champenoux, France
2L Université Nantes Angers Le Mans (LUNAM), Chaire Génie Civil Eco-construction, Institut de Recherche en Génie Civil et Mécanique GeM UMR-CNRS 6183, Saint-Nazaire, France

Understanding how wood develops has become an important problematic of plant sciences. However, studying wood formation requires the acquisition of count data difficult to interpret. Here, the annual wood formation dynamics of a conifer tree species were modeled using generalized linear and additive models (GLM and GAM); GAM for location, scale, and shape (GAMLSS); a discrete semiparametric kernel regression for count data. The performance of models is evaluated using bootstrap methods. GLM was useful to describe the wood formation general pattern but had a lack of fitting, while GAM, GAMLSS, and kernel regression had a higher sensibility to short-term variations.

Keywords  Bootstrap resampling methods; Count regression function; Discrete kernel; Generalized linear and additive models; Optimal bandwidth selections; Wood formation.

Mathematics Subject Classification Primary 62G05; Secondary 62F10; 62G08

1. Introduction

Wood is the most abundant biological component of the biosphere and plays a key role in ecosystem functioning, representing, for example, one of the strongest sink of CO$_2$ (Tans and White, 1998), which is a major contributor to climate change. Wood also plays a crucial economical role, being one of the most important products of the world trade (FAO, 2007). Understanding how wood develops therefore implies crucial issues, and the study of wood formation has become an innovative and fast-growing field in plant sciences over the last decade (Gricar et al., 2011).

Wood derives from the cambium, a thin layer of cells between the wood and the bark which divides to produce the new wood cells inward (Lachaud et al., 1999). Conifers represent the easiest case of wood formation because one type of cells, called tracheids, composed 90% to 95% of their wood. A cell newly produced by division in the cambium

Received August 6, 2013; Accepted December 11, 2013

Address correspondence to Tristan Senga Kiessé, L Université Nantes Angers Le Mans (LUNAM), Chaire Génie Civil Eco-construction, Institut de Recherche en Génie Civil et Mécanique GeM UMR-CNRS 6183, 58 rue Michel Ange F- 44600, Saint-Nazaire, France; E-mail: tristan.sengakiesse@univ-nantes.fr

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.
and destined to become a functional tracheid undergoes a differentiation program in order to acquire the particular morphological and physiological characteristics of this specialized cell type (Fig. 1) (Plomion et al., 2001): (1) first, its diameter enlarges mainly in the radial direction under water turgor pressure; (2) second, its thick, rigid, and waterproof secondary wall mainly composed of cellulose, hemicellulose, and lignin is deposited. Under temperate climate, wood formation presents an alternation between active and inactive periods related to the alternation of the hot and cold seasons. All the tracheids formed during the active period (growing season) are organized in juxtaposed radial files and formed an annual tree ring, which adds to the tree rings formed the previous years (Fig. 2).

Detailed analyses of wood formation need repeated sampling (generally at a weekly time step) of the developing tree ring during the growing season (Rossi et al., 2006). One of the main problem of this method is that growth is not homogenous along and around tree stem, so that the number of wood cells produced at a given time can considerably vary according to the position of the sampling (Wodzicki and Zajaczkowski, 1970). Therefore, it is difficult to know whether the wood cell number variations observed between the

**Figure 1.** Diagram showing the development of a tracheid, from its production by division of a cambial cell to its mature and functional state.

**Figure 2.** Anatomical transverse section of wood cut from a sample collected in October 2008 on the stem of a silver fir tree. The section is 6 μm in thick, stained with cresyl violet acetate and observed under light microscope (×100).
successive samples are related to growth or within-tree growth variability, making the
categorization of the wood cell number variation in the successive phases of tracheid
differentiation during the season difficult from a simple description of the raw cell count
dataset. Modeling these count data could be a good way to highlight the growth signal by
smoothing the variations resulting from within-tree variability.

In this work, we are concerned in finding a suitable methodology to model the num-
ber of cells weekly counted in the enlargement and thickening phases of five silver fir
(*Abies alba* Mill.) trees during the season 2007. Thus, let *Y* be a response variable in \( \mathbb{R} \),
*X* be a count explanatory variable in \( \mathbb{N}^d \), and the conditional mean
\( m(\cdot) = \mathbb{E}(Y | X = \cdot) \) an
unknown count regression function (c.r.f.) to estimate. A wide range of structured models
could be considered in modeling count data, for instance, generalized (linear additive)
models or generalized (partial linear) models discussed in Pardinas and Sperlich (2010).
First, we propose to focus on Generalized Linear and Additive Models (GLMs and GAMs,
respectively) for parametric and semiparametric regressions, respectively, of function *m*,
useful in the situation of a regression model for which the error term is not normally
distributed, as for binomial response variables or count data (Hastie and Tibshirani, 1990;
McCullagh and Nelder, 1989). More precisely, because the observations are repeated on
the same trees during the year, we are concerned with mixed effects versions of GLMs and
GAMs, denoted as GLMM and GAMM, respectively. Mixed models are recommended
to reduce the bias of the estimations in the case of repeated measurements since random
effects allow to take into account the correlation between observations (Zuur et al., 2009).
Then, we investigate a discrete semiparametric regression for count explanatory variables,
which requires to assume that c.r.f can be expressed as \( m = l \times \omega \) with *l* a parametric
function and \( \omega \) an unknown discrete nonparametric one. The semiparametric estimation
is realized in two steps: a first approximation \( \hat{l} \) of *l* followed by a discrete nonparametric
kernel regression of \( \omega = m / \hat{l} \) using a discrete version of Nadaraya–Watson estimator; in
this way, the nonparametric estimate plays the role of a correction coefficient of the para-
metric estimate (Abdous et al., 2012; Senga Kiessé and Rivoire, 2011). The purpose of
this semiparametric model is to improve the performance of parametric model *l*; later, in
the applications, we will illustrate this semiparametric approach using (parametric) model *l*
having the worst performance as departure. This approach focuses on the discrete character
of variables, and is thus appropriate for estimating count regression function by using some
associated kernels that are also discrete (Kokonendji and Senga Kiessé, 2011). Finally,
for comparison with other competitive generalized structured models, we apply GAMs
for location, scale, and shape (GAMLSS) proposed by Rigby and Stasinopoulos (2005)
as semiparametric regression-type models. They were introduced as a way of overcoming
some of the limitations associated with GLMs and GAMs. Note that for other works on
basic nonparametric or semiparametric regression models, one can refer to Dai and Sperlich
(2010), Lombardia and Sperlich (2008), or Alberts and Karunamuni (2003).

The remainder of this article is organized as follows. The mathematical models applied
are recalled in Section 2. For semiparametric kernel estimator, some examples of discrete
kernels are given. Moreover, some data-driven bandwidth selection procedures well known
in continuous case but not available until now for our discrete semiparametric regression
estimator are adapted. In particular, a new theoretical expression of optimal bandwidth
parameter is given for a type of discrete associated kernels used. The model application
on wood formation data and the results are provided in Section 3, bootstrap resampling
methods are used for a simulational assessment of the performance of models. Section
4 contains some concluding remarks. Finally, all mathematical details that are related to
data-driven bandwidth selection procedures are postponed to the Appendix.
2. Mathematical Models

In this section, we first present generalized linear and additive models used for a parametric and semiparametric regression, respectively, of function $m$; then we detail the semiparametric kernel regression methodology that consists by an improvement of the parametric estimate by using a nonparametric correction factor.

2.1. Generalized Models

An important statistical development over the last 30 years has been the advance in regression analysis provided by GLMs and GAMs. GAMs are semiparametric extensions of GLMs, which are themselves mathematical extensions of classical linear models (Hastie and Tibshirani, 1990). By applying a mathematical transformation to the response variable according to the real distribution of the errors, GLMs generalize the linear model to nonlinear responses and variables that can have other than a normal distribution, including the binomial, Poisson, or gamma probability distributions. GAMs are GLMs in which the linear predictor depends, in part, on a sum of smooth functions of predictors. The strength of GAMs is their ability to deal with highly nonlinear and nonmonotonic relationships between the response and the set of explanatory variables. At last, in GAMLSS the exponential family distribution assumption for the response variable is relaxed and replaced by a general distribution family, including highly skew and/or kurtotic continuous and discrete distributions (Rigby and Stasinopoulos, 2005). The ability of this tool to handle nonlinear data structures can aid in the development of ecological models that better represent the underlying data, and hence increase our understanding of ecological systems (Guisan et al., 2002).

Concerning GLM with the standard Poisson model as an appropriate choice, it consists here by a linear predictor $l(\cdot; \Theta)$ with parameter $\Theta = (\theta_0, \theta_1, \theta_2)$, based on a combination of realization $x \in \mathbb{N}$ of the explanatory variable $X$ with a logarithmic link such that we have

$$y_i = l(x_i; \Theta) + e_i = \theta_0 + \theta_1 x_i + \theta_2 \log x_i + e_i,$$

where $e_i$ represents the residuals. The linear and logarithm parts in $l(\cdot; \theta)$ allow to describe the tendency of data. About GAM, we have

$$y_i = \theta_0 + s(x_i) + e_i, \quad x_i \in \mathbb{N},$$

with $s(\cdot)$ a nonparametric function that can be modeled by splines. Furthermore, the mixed effects versions of GLM and GAM (GLMM and GAMM, respectively) are available for describing relationship between a response variable and covariates in the repeated measurements data that are grouped according to a cluster factor. At last, we consider GAMLSS models defined as follows (Rigby and Stasinopoulos, 2005). Assume $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ be a vector of four distribution parameters where each of which can be a function to explanatory variables. For $k = 1, 2, 3, 4$, let $g_k$ be known monotonic link functions relating the distribution parameters to explanatory variables. The semiparametric additive formulation of GAMLSS can be given by

$$g_k(\theta_k) = X_k \beta_k + \sum_{j=1}^{J_k} d_{jk}(x_{jk}),$$

(1)
where \( d_{jk} \) is an unknown function of the explanatory variable. Under some conditions, GAMLSS in (1) can be extended to parametric linear, nonlinear parametric, or nonlinear semiparametric additive model.

2.2. **Semiparametric Kernel Regression**

2.2.1. **Estimator.** Let us consider the sequence of independent and identically distributed \((\text{iid})\) random variables \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\) defined on \(\mathbb{N} \times \mathbb{R}\) such that \(Y_i = m(X_i) + \epsilon_i\), with \(\epsilon_i\) the residuals assumed to have zero mean and finite variance. In the semiparametric approach, the distribution of \(Y_i\) is assumed to be a modified parametric function given by

\[
m(x) = l(x; \Theta) \times \omega(x), \quad x \in \mathbb{N},
\]

where \(l(x; \Theta)\) is a function relative to the parameter \(\Theta\) and \(\omega(x) > 0\) is a nonparametric function. The discrete semiparametric regression estimator of \(m\) in (2) results from a parametric estimation \(\hat{l}(x) = l(x; \hat{\Theta})\) of \(l\) followed by a nonparametric kernel estimation \(\hat{\omega}_n(x)\) of \(\omega(x) = m(x)/\hat{l}(x)\) such that we have

\[
\hat{m}_n(x) = \hat{l}(x) \times \hat{\omega}_n(x) = \hat{l}(x) \sum_{i=1}^{n} \frac{Y_i/\hat{l}(X_i)}{\sum_{j=1}^{n} K(x; h)} K(x; h), \quad x \in \mathbb{N};
\]

see Abdous et al. (2012). The bandwidth \(h = h(n) > 0\) is an arbitrary sequence of smoothing parameters, which fulfills \(\lim_{n \to \infty} h(n) = 0\); and the discrete associated kernel \(K(x; h)\) of random variable \(K_{x,h}\) is a probability mass function (p.m.f.) with support \(S_x\) (containing \(x\)) included in \(\mathbb{N}\) satisfying the following hypotheses:

- (H1): \(\lim_{h \to 0} E(K_{x,h}) = x\) and (H2): \(\lim_{h \to 0} \text{Var}(K_{x,h}) = 0\).

2.2.2. **Examples of discrete kernel.** We present two examples of discrete kernels for which more details are available in Kokonendji and Senga Kiessé (2011), and references therein. The first discrete kernel satisfies only the assumption (H1) and has its variance such that \(\lim_{h \to 0} \text{Var}(K_{x,h}) \in \mathcal{N}(0)\) instead of (H2), where \(\mathcal{N}(0)\) is a neighborhood of 0. This kernel might be particularly useful for smoothing small or moderate samples sizes (refer to Kokonendji and Senga Kiessé, 2011; Zougab et al., 2012, 2013). The second kernel fulfills (H1)–(H2).

**Example 2.1.** For \(x \in \mathbb{N}\) and \(h \in (0, 1)\), the first kernel is the binomial one \(B(x; h)\), which follows the binomial distribution \(\mathbb{B}(x+1, (x+h)/(x+1))\) on support \(S_x = \{0, 1, \ldots, x+1\}\).

**Example 2.2.** For \((x, a) \in \mathbb{N} \times \mathbb{N}\) and \(h > 0\), the second kernel is a discrete symmetric triangular one with random variable \(K_{a,x,h}\) defined on support \(S_{a,x} = \{x, x \pm 1, \ldots, x \pm a\}\) and whose p.m.f. is given by

\[
\text{Pr}(K_{a,x,h} = z) = \frac{(a + 1)^h - |z - x|^h}{P(a, h)}, \quad \forall z \in S_{a,x},
\]
with $P(a, h) = (2a + 1)(a + 1)^h - 2 \sum_{k=1}^{a} k^h$ the normalizing constant. In addition, we propose the following expansions of modal probability and variance of this kernel such that

$$\Pr(K_{a,x,h} = x) = 1 - 2hA(a) + O(h^2) \quad \text{and} \quad \Var(K_{a,x,h}) = 2hV(a) + O(h^2),$$

with $A(a) = a \log(a + 1) - \sum_{k=1}^{a} \log(k)$ and $V(a) = \{a(2a^2 + 3a + 1)/6\} \log(a + 1) - \sum_{k=1}^{a} k^2 \log(k)$. These expansions are useful for establishing an expression of optimal bandwidth (see the Appendix).

2.2.3. Bandwidth choices. We adapt two data-driven bandwidth selection methods from continuous regression but not developed until now for discrete regression estimator $\hat{m}_n$. The first consists by finding an optimal value $h_{cv}$ by minimizing the score function

$$\text{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \hat{m}_{n,-i}(X_i; h) \right)^2 = \text{CV}_1(h) - 2\text{CV}_2(h) + (1/n) \sum_{i=1}^{n} Y_i^2.$$

The second consists by finding a theoretical expression of optimal bandwidth parameter $\hat{h}_{opt}$ obtained by minimizing the asymptotic part of mean integrated squared error (MISE) of $\hat{m}_n$ such that

$$\text{MISE}\{\hat{m}_n(x)\} = \sum_{x \in \mathbb{N}} \text{Var}\{\hat{m}_n(x)\} + \sum_{x \in \mathbb{N}} \text{Bias}^2\{\hat{m}_n(x)\}.$$

Indeed, the problem with using MISE to bandwidth selection is that we are not able to provide a theoretical expression of optimal bandwidth $h_{opt}$ minimizing MISE in discrete associated kernel estimation. Thus, $\hat{h}_{opt}$ is an approximate of the true $h_{opt}$. However, this optimal bandwidth is available only for estimator using discrete associated kernels satisfying (H1)–(H2). Thus, a theoretical expression $\hat{h}_{opt}$ is not available for binomial kernel that does not fulfill (H2) but only for discrete triangular kernels. Moreover, we will see that $\hat{h}_{opt}$ cannot be directly used since it requires to know the true function $f$. Therefore, in the following section, the estimator $\hat{m}_n$ in (3) is just applied with binomial kernel using the optimal $h$-value $h_{cv}$. The mathematical details of these two methods are postponed to the Appendix.

A Bayesian approach is proposed in Zougab et al. (2014) as an alternative approach to bandwidth selection in the context of nonparametric count regression using binomial kernel; this approach permits also the variance estimation of the model error. An extension of this approach in the context of semiparametric count regression using binomial kernel would provide an improvement of the smoothing quality; this requires a thorough job in a separate article.

Remark 2.1. (i) First, it would be of some interest to compare, for example, $\text{MISE}(\hat{m}_{n,h_{cv}})$ and $\text{MISE}(\hat{m}_{n,\hat{h}_{opt}})$. To derive such a result, some information about $\sup_{H_n} |\text{CV}_1 - 2\text{CV}_2 + (1/n) \sum_{i=1}^{n} Y_i^2 - \text{MISE}|$, for some sufficiently large region $H_n$, would be necessary. That will be the subject of a forthcoming article.

(ii) Second, a common problem with cross-validation methodology is that the criteria CV is not always consistent depending on discrete kernel $K_{x,h}$ and sample $(X_i)_{i=1,2,...,n}$. In this situation, for example, the function $h \mapsto \text{CV}(h)$ is only increasing and does not alternate the phases of decreasing and increasing that allow to find a minimum $h$-value. It would be interesting to adapt a generalized cross-validation procedure in forthcoming
works for our discrete semiparametric regression, for example, see the works of Tong and Yao (1998) for regression estimation based on dependent data.

3. Applications

In this section, we consider the data of the number of cells weekly counted in the diameter enlargement and wall-thickening phases of five silver fir trees in 2007 presented in Fig. 3.
### Table 1

| Phase         | Tree | GLMM | GAMM | Semip. regr. | GLMM | GAMM |Semip. regr. |
|---------------|------|------|------|--------------|------|------|------------|
| Diam. enlarg. |      | 6.421| 5.632| 4.584       | 5.213| 4.635|4.794       |
|               |      | 8.942| 11.580| 6.603       | 40.584| 4.695|9.603       |
|               |      | 12.434| 8.132| 6.559       | 24.664| 6.101|16.654      |
|               |      | 11.045| 9.179| 5.705       | 49.383| 9.210|32.710      |
|               |      | 17.824| 15.181| 7.705       | 49.383| 9.210|32.710      |
|               |      | 11.333| 9.941| 5.821       | 34.041| 34.251|25.900      |
| Wall thick.   |      | 22.582| 18.897| 9.014       | 22.582| 18.897|9.014       |
|               |      | 40.584| 29.495| 11.897      | 40.584| 29.495|11.897      |
|               |      | 24.664| 18.601| 7.943       | 24.664| 18.601|7.943       |
|               |      | 49.383| 33.210| 12.654      | 49.383| 33.210|12.654      |
|               |      | 34.041| 29.296| 16.630      | 34.041| 29.296|16.630      |
|               |      | 34.251| 25.900| 11.628      | 34.251| 25.900|11.628      |

### 3.1. Data

Small samples of wood called microcores (2 mm diameter, 15–20 mm length) were collected weekly from April to November 2007 at breast height on the stems of five silver fir trees using a specifically designed puncher, the Trephor tool (Vitzani, Belluno, Italy) (Rossi et al., 2006), and following an ascending spiral pattern. For each sample, the number of cells in the enlargement and thickening phases was counted along three radial files (Cuny et al., 2012, 2013). The number of cells weekly counted in the wood formation phases of diameter cell enlargement and wall thickening of one silver fir tree are presented in Fig. 3(a) and 3(b). The observations in the two same phases of the five silver fir trees are plotted in Fig. 3(c) and 3(d).

### 3.2. Mixed Effect Models

For wood cell count data in Fig. 3(c) and 3(d), which present five trees $i$ observed on 31 weeks $j$, we apply the mixed model such that $m(\cdot) = E(Y_{ij}|X_{ij} = \cdot), i = 1, 2, \ldots, 5, j = 1, 2, \ldots, 31$, where the response variable $Y_{i,j}$ is the number of cells measured at time $j$, the function $m$ relates the response variable to other covariates $X_{i,j}$ varying with each tree and time. Thus, for GLMM as example, the parameter $\Theta$ of GLM is replaced by $\Theta_i = (\theta_0 + a_{0i}, \theta_1 + a_{1i}, \theta_2 + a_{2i}, \theta_3 + a_{3i})$, where $\theta_k, k = 1, 2, 3$, are the fixed effects and $a_{ki}$ are random effects related to each individual $i$. The different types of generalized models mentioned are implemented on the R packages “lme4” (Bates et al., 2011) and “gamm4” (Wood, 2011). We first apply the mixed effects models GLMM and GAMM on data of five trees presented in the graphs (c) and (d) in comparison with discrete semiparametric binomial kernel regression estimator using $h = h_{cv}$. The parametric estimates provided by GLMM are used as departure for the semiparametric procedure, since we will see that GLMM has the worst performance.

In order to evaluate the performance of the models, we just apply the root mean square error (RMSE), which is a descriptive measure of degree-of-fit defined as $\text{RMSE}_i = \sqrt{\left(\sum_{j=1}^n (y_{ij} - \hat{y}_{ij})^2\right)/n}, i = 1, 2, \ldots, 5$, where $\hat{y}_{ij}$ is the adjustment of the $j$th measurement $y_{ij}$ of a tree $i$ and $n = 31$ the total number of observations. The average RMSE is also presented in Table 1 such that $\text{RMSE} = (1/5) \sum_{i=1}^5 \text{RMSE}_i$. 

For the diameter enlargement phase, the GLMM and GAMM as well as the semiparametric kernel estimator were suitable to exhibit the unimodal tendency of the seasonal dynamics for all trees except the tree $i = 5$ (Fig. 4). For this last tree, GLMM and GAMM described a unimodal curve while semiparametric regression provided a bimodal curve. In general, the estimated curves provided by GLMM were over-smoothed, while those provided by GAMM and semiparametric binomial regression reflected more data variations. Moreover, the modal value was globally under-estimated and moved from a few days by GLMM; and, this parametric model over-estimated the null values observed at left boundary points. Finally, concerning GAMM and semiparametric regression, they allowed to highlight similar tendencies, except for tree $i = 5$.

For thickening phase, the conclusions are similar as previously. In particular, let us present the following remarks. In Fig. 5, for individual tree $i = 2$, the semiparametric binomial regression clearly pointed out some estimations with a bimodal tendency in contrast to GLMM and GAMM. Moreover, GLMM largely over-estimated the values observed at right boundary points. The corresponding results are given in Table 1. Finally, it appeared that the nonparametric correction provided by estimator using binomial kernel
allowed to clearly improve the parametric estimates resulting from GLMM and was better than GAMM in term of performance (Table 1). In the following in order to deepen our study by using bootstrap methods, a comparison is realized with GAMLSS, which is another competitive generalized structured model. We apply GLM, GAM, GAMLSS, and $\tilde{m}_n$ on re-sampled data of one tree presented in the graphs (a) and (b). Note that for GAMLSS we proceed by fitting smoothing cubic splines to the data and for semiparametric kernel model we omit to present $h_{cv}$-values.

3.3. Bootstrap Methods

These methods are applied for a robust evaluation, which consists by re-sampling the number of wood cells weekly counted in the diameter enlargement or wall-thickening phase of one silver fir tree during 2007 presented in Fig. 3(a) and 3(b). In addition, compared to the previous section, we deepen our study by comparing to GAMLSS. Thus, from the data of this tree we draw $N = \{25, 50, 150, 200, 250\}$ bootstrap samples on whom we apply the different models studied. In addition with RMSE applied in previous section,
Table 2

RMSE and AIC resulting from estimations of the number of wood cells weekly counted in the diameter enlargement and wall-thickening phases of five silver fir trees during 2007, by applying GLM, GAM, GAMLSS, and semiparametric regression estimator with binomial kernel.

| Model            | Criterion | 25  | 50   | 150  | 200  | 250  |
|------------------|-----------|-----|------|------|------|------|
|                  | RMSE      |     |      |      |      |      |
| Diameter enlargement phase |           |     |      |      |      |      |
| GLM              | RMSE      | 29.750 | 29.771 | 29.610 | 29.996 | 29.883 |
| with df = 28     | AIC       | 230.279 | 230.353 | 229.990 | 230.830 | 230.580 |
| GAM              | RMSE      | 24.071 | 25.169 | 25.104 | 24.493 | 25.073 |
| with df = 10     | AIC       | 216.758 | 219.506 | 219.283 | 217.676 | 219.286 |
| GAMLSS           | RMSE      | 22.263 | 22.363 | 23.113 | 23.147 | 23.215 |
| with df = 10     | AIC       | 211.913 | 212.004 | 214.334 | 214.439 | 214.569 |
| Semip. regr.     | RMSE      | 23.644 | 22.542 | 22.836 | 23.558 | 23.150 |
| with df = 27     | AIC       | 249.835 | 246.737 | 247.526 | 247.892 | 248.505 |
| Wall-thickening phase |           |     |      |      |      |      |
| GLM              | RMSE      | 77.939 | 77.823 | 77.806 | 77.777 | 77.736 |
| with df = 28     | AIC       | 290.034 | 289.949 | 289.928 | 289.904 | 289.875 |
| GAM              | RMSE      | 67.676 | 66.178 | 67.546 | 66.746 | 67.202 |
| with df = 10     | AIC       | 281.179 | 279.561 | 281.008 | 280.210 | 280.630 |
| GAMLSS           | RMSE      | 61.661 | 63.869 | 63.188 | 63.827 | 63.611 |
| with df = 10     | AIC       | 275.351 | 277.436 | 276.856 | 277.487 | 277.279 |
| Semip. regr.     | RMSE      | 58.435 | 59.813 | 58.538 | 59.443 | 58.779 |
| with df = 27     | AIC       | 305.868 | 307.267 | 305.933 | 306.935 | 306.201 |

looking first at the descriptive measure of degree-of-fit, GAMLSS and semiparametric binomial kernel regression have closed performance; they are both better than GLM and GAM. Then, by taking also into account the number of parameters via the AIC, the smallest values of this measure are provided by GAMLSS (df = 10), followed by GAM (df = 10), GLM (df = 28), and at last semiparametric kernel regression (df = 27). From here, GAMLSS is the most interesting model.

4. Discussion

Detailed analyses of wood formation need repeated sampling (generally at a weekly time step) of the developing tree ring during the growing season (Rossi et al., 2006). One of the main problems encountered when studying wood formation is that the wood cell number variations related to growth, which interest the biologists, are confused with the wood cell number variations related to the heterogeneous growth along and around the stem. So, the objective is to find a model that smooths the data while maintaining a suitable fit. In this study, the application of GL(M)M, GA(M)M, GAMLSS, and semiparametric binomial
kernel regression on the data of five trees offered different solutions to cope with this trade-off between degree of smoothing and goodness of fit, so that these different methods should be intended to different purposes. The GL(M)M method, which allows for a high degree of smoothing, may be sufficient when the will is to describe the wood formation general pattern, because it highlights the general shape of cell number variations during the season. But its lack of fitting makes it unsuitable when the will is to assess in detail the dynamics of the wood formation process. For example, one of the central aim of wood formation studies is to understand how climate influences wood formation, and this needs to relate climatic and wood formation data at a fine temporal resolution (Gricar et al., 2011). In this case, the GA(M)M, GAMLSS, and semiparametric kernel procedures could offer great perspectives because of their higher sensibility to the high-frequency variations in the dataset. The incorporation of mixed effects in GA(M)M is also a good point because it allows to accurately account for the nonindependence of weekly cell count data (Zuur et al., 2009), providing a more robust description of wood formation dynamics. Finally, the discrete semiparametric kernel regression and GAMLSS offer interesting alternative to GL(M)M and GA(M)M. Indeed, GAMLSS is known in literature to have a systematic part expanded to allow modeling not only of the mean but other parameters of the distribution of \( Y \) as linear and/or nonlinear, parametric and/or additive nonparametric functions of explanatory variables, and/or random effects (Rigby and Stasinopoulos, 2005). Concerning semiparametric kernel regression, the balance wished between degree of smoothing and goodness of fit can be obtained by playing on the value of the bandwidth parameter \( h > 0 \) according to the purpose of the study.

Appendix: Bandwidth Choices

A.1. Cross-validation Procedure

An \( h \)-value denoted \( h_{cv} \) can be selected by the well-known cross-validation procedure adapted to the estimator \( \hat{m}_n \) in (3). For this, we express the discrete semiparametric estimator \( \hat{m}_n \) of the c.r.f. \( m \) as

\[
\hat{m}_n(x) = \sum_{i=1}^{n} U_{x,h}(X_i)Y_i \times \frac{\hat{l}(x)}{\hat{l}(X_i)}
\]

with \( U_{x,h}(X_i) = K_{x,h}(X_i)/ \sum_{j=1}^{n} K_{x,h}(X_j) \). The optimal bandwidth selection by the cross-validation method consists by the minimization of the terms depending on \( h \) in the criterion

\[
\frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \hat{m}_{n,-i}(X_i; h) \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_{n,-i}^2(X_i; h) - \frac{2}{n} \sum_{i=1}^{n} \hat{m}_{n,-i}(X_i; h)Y_i + \frac{1}{n} \sum_{i=1}^{n} Y_i^2
\]

\[
\equiv CV_1 - 2CV_2 + \frac{1}{n} \sum_{i=1}^{n} Y_i^2,
\]

where

\[
\hat{m}_{n,-i}(X_i; h) = \sum_{j \neq i}^{n} \frac{Y_j K_{X_i,h}(X_j)}{\sum_{j \neq i}^{n} K_{X_i,h}(X_j)} \times \frac{\hat{l}(X_i)}{\hat{l}(X_j)}
\]
is the leave-one-out kernel estimator of \( \hat{m}_n(X_i; h) \); thus, we have \( h_{cv} = \arg\min CV(h) \) with \( CV(h) = CV_1 - 2CV_2 \). By calculating the expectation of (A.1), we show that it is an unbiased estimator of the following version of mean integrated squared error (MISE) weighted by \( p.m.f \ f \) given by \( E\left[\sum_{x \in \mathbb{N}}(\hat{m}_n(x) - m(x))^2 f(x)\right] \).

Indeed, we directly see that the terms

\[
CV_1 = \frac{1}{n} \sum_{i=1}^{n} \left\{ \sum_{j \neq i}^{n} l(X_i; \hat{\Theta}_{-i}) U_{X_i,h}(X_j) Y_j \right\}^2
\]

and \((1/n)\sum_{i=1}^{n} Y_i^2\) are, respectively, unbiased asymptotic estimates of \( E\{\sum_{x \in \mathbb{N}} \hat{m}_n(x)^2 f(x)\} \) and \( E\{\sum_{x \in \mathbb{N}} m(x)^2 f(x)\} \). Then, we show that \( E\{\sum_{x \in \mathbb{N}} \hat{m}_n(x)m(x)f(x)\} \) is asymptotically approximated by

\[
CV_2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i}^{n} U_{X_i,h}(X_j) Y_i Y_j \frac{l(X_i; \hat{\Theta}_{-i})}{l(X_j; \Theta_{-i})}
\]

since we first have

\[
E(CV_2) = E\left\{\frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i}^{n} l(X_i; \hat{\Theta}_{-i}) U_{X_i,h}(X_j) Y_i Y_j \right\} = E\left\{\sum_{j \neq i}^{n} l(X_i; \hat{\Theta}_{-i}) U_{X_i,h}(X_j) Y_i Y_j \right\}
\]

and, second,

\[
E\left\{\sum_{x \in \mathbb{N}} \hat{m}_n(x)m(x)f(x)\right\} = E\left\{\sum_{i=1}^{n} \frac{\hat{f}(x)}{l(X_i)} U_{x,h}(X_i) Y_i E(Y_i | X_i = x) f(x)\right\}
\]

\[
= E\left[ E\left\{\sum_{i=1}^{n} \frac{\hat{f}(x)}{l(X_i)} U_{x,h}(X_i) Y_i^2 f(x)\right\} | X_i = x\right]
\]

\[
= E\left\{\sum_{i=1}^{n} \frac{\hat{f}(x)}{l(X_i)} U_{X_i,h}(X_i) Y_i^2\right\}.
\]

Note that the estimate \( \hat{\Theta}_{-i} \) is calculated as \( \hat{\Theta} \) by excluding \( X_i \). One can refer to Hardle and Marron (1985) for bandwidth selection in nonparametric continuous regression.

In the next part, we establish a new theoretical expression of optimal parameter \( \hat{h}_{\text{opt}} \) for estimator \( \hat{m}_n \) with discrete associated kernels satisfying (H1′)–(H2′).

### A.2. Minimization of Asymptotic Part of Mean Integrated Squared Error

This approach is required to calculate bias and variance of estimator \( \hat{m}_n \) since MISE in (4). Let us assume \( l_0(x) = l(x; \Theta_0) \) be a fixed parametric start in (2), that is, \( m = l_0(x) \), such that \( \hat{\Theta} \) converges to \( \Theta_0 \) in probability. For \( x \in \mathbb{N} \), the discrete semiparametric estimator \( \hat{m}_n \) in (3) using discrete triangular kernels admits the following bias:

\[
\text{Bias}\{\hat{m}_n(x)\} = \frac{h}{2} V(K_{u;x,h}) W(x) + O\left(\frac{1}{n} + h^2\right) + o(h),
\]
with
\[ \sum_{x \in \mathbb{N}} \{ W(x) V(K_{a; x, h}) \}^2 = \sum_{x \in \mathbb{N}} \left[ V(K_{a; x, h}) \left\{ l_0(x) \omega^{(2)}(x) + 2 l_0(x) \omega^{(1)}(x) \left( f^{(1)}/f \right)(x) \right\} \right]^2 < \infty; \]
then, for its variance, we have
\[ \text{Var}\{ \hat{m}_n(x) \} = \frac{\text{Var}(Y|X = x)}{n f(x)} \left( 1 - h A(K_{a; x, h}) \right)^2 + o \left( \frac{1}{n} \right) + O(h^2). \]

The conditional variance \( \text{Var}(Y|X = x) \) is finite, the function \( f > 0 \) is the p.m.f. of the regressor \( X \); \( \omega^{(1)} \), \( f^{(1)} \), and \( \omega^{(2)} \) are respectively finite differences of first and second order (Abdous et al., 2012). Hence, we give the following expression of MISE:
\[ \text{MISE}\{ \hat{m}_n(x) \} = \sum_{x \in \mathbb{N}} \mathbb{E}\{ \hat{m}_n(x) - m(x) \}^2 = \text{AMISE}(h) + o \left( \frac{1}{n} + h^2 \right) + O(h^2) \]
\[ \equiv \text{MISE}(h), \]
where \( \text{AMISE}(h) \) is the main term in the sum of integrated variance and squared bias of \( \hat{m}_n \). The bandwidth \( \hat{h}_{\text{opt}} = \arg \min_{h > 0} \text{AMISE}\{ \hat{m}_n(x) \} \) comes by solving the following equation \( d\{ \text{AMISE}(h) \}/dh = 0 \), which is equal to
\[ h V^2(a) \sum_{x \in \mathbb{N}} \left\{ W(x) \right\}^2 - \sum_{x \in \mathbb{N}} A(a) \left[ 1 - h A(a) \right] \frac{\text{Var}(Y|X = x)}{n f(x)} = 0. \]
That leads to expression
\[ \hat{h}_{\text{opt}}(a; n, f) = \frac{A(a) \sum_{x \in \mathbb{N}} \text{Var}(Y|X = x)/f(x)}{A^2(a) \sum_{x \in \mathbb{N}} \text{Var}(Y|X = x)/f(x) + n V^2(a) \sum_{x \in \mathbb{N}} W^2(x)} \]
(6)
such that \( \hat{h}_{\text{opt}} \to 0 \) when \( n \to \infty \). The following asymptotic relationship can be pointed out: \( \hat{h}_{\text{opt}} \sim k_0 n^{-1} \) with
\[ k_0 = \frac{A(a) \sum_{x \in \mathbb{N}} \text{Var}(Y|X = x)/f(x)}{V^2(a) \sum_{x \in \mathbb{N}} W^2(x)}. \]

Acknowledgments

The authors thank E. Cornu, E. Farré, C. Freyburger, P. Gelhaye, and A. Mercanti for fieldwork and monitoring; M. Harroué for sample preparation in the laboratory; and M. Dassot for his help with editing the figures. The authors also thank the associate editor and anonymous referee for hints that improved the article.

References

Abdous, B., Kokonendji, C. C., Senga Kiessé, T. (2012). On semiparametric regression for count explanatory variables. Journal of Statistical Planning and Inference 142:1537–1548.
Alberts, T., Karunamuni, R. J. (2003). A semiparametric method of boundary correction for kernel density estimation. Statistics & Probability Letters 61:287–298.
Bates, D., Maechler, M., Bolker, B. (2011). lme4: Linear Mixed-Effects Models Using S4 Classes. R Package Version 0.999375-42. Available at: http://CRAN.R-project.org/package=lme4.
Cuny, H., Rathgeber, C. B. K., Lebourgeois, F., Fortin, M., Fournier M., (2012). Life strategies in intra-annual dynamics of wood formation: Example of three conifer species in a temperate forest in north-east France. *Tree Physiology* 32:612–625.

Cuny, H., Rathgeber, C. B. K., Senga Kiessé, T., Hartman, F. P., Barbeito, I., Fournier M., (2013). Generalised additive models reveal the intrinsic complexity of the wood formation dynamics. *Journal of Experimental Botany* 64(7): 1983–1994.

Dai, J., Sperlich, S. (2010). Simple and effective boundary correction for kernel densities and regression with an application to the world income and Engel curve estimation. *Computational Statistics & Data Analysis* 54:2487–2497.

FAO. (2007). *Global Wood and Wood Products - Flow Trends and Perspectives*. Shanghai, China: Advisory Committee on Paper and Wood Products; Rome, Italy: Food and Agriculture Organization of the United Nations (FAO).

Gricar, J., Rathgeber, C. B. K., Fonti, P. (2011). Monitoring seasonal dynamics of wood formation. *Dendrochronologia* 29:123–125.

Guisan, A., Edwards, T. C., Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling* 157:89–100.

Hardle, W., Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics* 13:1465–1481.

Hastie, T. J., Tibshirani, R. J. (1990). *Generalized Additive Models*. 4th ed. London: Chapman and Hall.

Kokonendji, C. C., Senga Kiessé, T. (2011). Discrete associated kernel method and extensions. *Statistical Methodology* 8:497–516.

Lachaud, S., Catesson, A. M., Bonnemain, J. L. (1999). Structure and functions of the vascular cambium. *Comptes Rendus de l’Académie des Sciences - Series III - Sciences de la Vie* 322:633–650.

Lombardia, M. J., Sperlich, S. (2008). Semiparametric Inference in Generalized Mixed Effects Models. *Journal of the Royal Statistical Society: Series B* 70:913–930.

McCullagh, P., Nelder, J. A. (1983). *Generalized Linear Models*. 1st ed. London: Chapman and Hall.

Pardinas, R., Sperlich, S. (2010). Feasible estimation in generalized structured models. *Statistics and Computing* 20:367–379.

Plomion, C., Leprovos, G., Stokes, A. (2001). Wood formation in trees. *Plant Physiology* 127:1513–1523.

Rigby, R. A., Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics* 54:507–554.

Rossi, S., Anfodillo, T., Menardi, R. (2006a). Trephor: A new tool for sampling microcores from tree stem. *IAWA Journal* 27:89–97.

Tans, P. P., White, J. W. C. (1998). In balance, with a little help from the plants. *Science* 281:183–184.

Tong, H., Yao, Q. (1998). Cross-validatory bandwidth selection for regression estimation based on dependent data. *Journal of Statistical Planning and Inference* 68:387–415.

Wodzicki, T. J., Zajaczkowski, S. (1970). Methodical problems in studies on seasonal production of cambial xylem derivatives. *Acta Societatis Botanicorum Poloniae* 39:509–520.

Wood, S. (2011). *gamm4: Generalized Additive Mixed Models Using mgcv and lme4. R Package Version 0.1-5*. Available at: http://CRAN.R-project.org/package=gamm4. http://CRAN.R-project.org/package=gamm4.

Zougab, N., Adjabi, S., Kokonendji, C. C. (2012). Binomial kernel and Bayes local bandwidth in discrete functions estimation. *Journal of Nonparametric Statistics* 24:783–795.

Zougab, N., Adjabi, S., Kokonendji, C. C. (2013). Adaptive smoothing in associated kernel discrete functions estimation using Bayesian approach. *Journal of Statistical Computation and Simulation* 83:2219–2231.

Zougab, N., Adjabi, S., Kokonendji, C. C. (2014). Bayesian approach in nonparametric count regression with binomial kernel. *Communications in Statistics-Simulation and Computation* 43:1052–1063.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R. Statistics for Biology and Health*. New York: Springer.