On Modeling Wood Formation Using Parametric And Semiparametric Regressions For Count Data

Henri H. Cuny, Tristan Senga Kiessé

ON MODELING WOOD FORMATION USING PARAMETRIC AND SEMIPARAMETRIC REGRESSIONS FOR COUNT DATA

Preprint (revised version) submitted in *Communications in Statistics - Simulation and Computation*, 2013

Henri E. Cuny

INRA, UMR1092, Laboratoire d Etude des Ressources Foret Bois (LERFoB),

Centre INRA de Nancy, F-54280 Champenoux, France.

henri.cuny@nancy.inra.fr

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.

tristan.sengakiesse@univ-nantes.fr

ABSTRACT

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.

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

1. INTRODUCTION

Wood is the most abondant biological component of the biosphere and plays a key role in ecosystem functioning, representing for example one of the strongest sink of CO2 (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 product 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 that divide to produce the new wood cells inwards (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 and destined to become a functional tracheid undergoes a differentiation program in order to acquire the particular morphological and physiological characteristics of this specialised cell type (Figure 1) (Plomion et al., 2001): (1) firstly, its diameter enlarges mainly in the radial direction under water turgor pressure; (2) secondly, its thick, rigid, and waterproof secondary wall mainly composed of cellulose, hemicellulose and lignin is deposited. Under temperate climate, wood formation present 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 (Figure 2).

Figures 1 and 2 about here

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 homogeneous 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 successive samples are related to growth or within-tree growth variability, making the characterization 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 number 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 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 ω an unknown discrete nonparametric one. The semiparametric estimation is realized in two steps: a first approximation $\widehat{l}$ of l followed by a discrete nonparametric kernel regression of $\omega = m/\widehat{l}$ using a discrete version of Nadaraya-Watson estimator; in this way, the nonparametric estimate plays the role of a correction coefficient of the parametric 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 appropriated for estimating count regression function by using some *associated kernels* which 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 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 paper 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 regresion 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 which are related to data-driven bandwidth selection procedures are postponed to Appendix.

2. MATHEMATICAL MODELS

In this section, we first present generalized linear and additive models used for a parametric and semi-parametric regression, respectively, of function m; then we detail the semiparametric kernel regression methodology which 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 semi-parametric 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 non-linear 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 non-linear and non-monotonic 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 non-linear 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 which 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 semi-parametric additive formulation of GAMLSS can be given by

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

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

2.2. Semiparametric kernel regression

2.2.1. Estimator

Let us consider the sequence of independent and identically distributed (i.i.d.) 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 ϵ_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

$$m(x) = l(x; \Theta) \times \omega(x), \quad x \in \mathbb{N}, \tag{2}$$

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

$$\widehat{m}_n(x) = \widehat{l}(x) \times \widehat{\omega}_n(x) = \widehat{l}(x) \sum_{i=1}^{n} \frac{\{Y_i/\widehat{l}(X_i)\}K_{x,h}(X_i)}{\sum_{j=1}^{n} K_{x,h}(X_j)}, \ x \in \mathbb{N}; \tag{3}$$

see Abdous et al. (2012). The bandwidth $h = h(n) > 0$ is an arbitrary sequence of smoothing parameters that fulfills $\lim_{n\to\infty} h(n) = 0$; and, the discrete associated kernel $K_{x,h}(\cdot)$ of random variable $\mathcal{K}_{x,h}$ is a probability

mass function (p.m.f.) with support $\mathcal{S}_x$ (containing x) included in $\mathbb{N}$ satisfying the following hypotheses:

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

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 have its variance such that $\lim_{h\to 0} \mathrm{Var}(\mathcal{K}_{x,h}) \in \mathcal{V}(0)$ instead of (H2), where $\mathcal{V}(0)$ 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 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 $\mathcal{B}\{x + 1, (x + h)/(x + 1)\}$ on support $\mathcal{S}_x = \{0, 1, \cdots, x + 1\}$.

Example 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 $\mathcal{K}_{a;x,h}$ defined on support $\mathcal{S}_{a,x} = \{x, x \pm 1, \cdots, x \pm a\}$ and whose p.m.f. is given by

$$\Pr(\mathcal{K}_{a;x,h} = z) = \frac{(a + 1)^h - |z - x|^h}{P(a, h)}, \quad \forall z \in \mathcal{S}_{a;x},$$

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

$$\Pr(\mathcal{K}_{a;x,h} = x) = 1 - 2h\mathrm{A}(a) + O(h^2) \;\text{ and }\; \mathrm{Var}(\mathcal{K}_{a;x,h}) = 2h\mathrm{V}(a) + O(h^2),$$

with $\mathrm{A}(a) = a\log(a + 1) - \sum_{k=1}^{a} \log(k)$ and $\mathrm{V}(a) = \{a(2a^2 + 3a + 1)/6\}\log(a + 1) - \sum_{k=1}^{a} k^2 \log(k)$. Thee expansions are useful for establishing an expression of optimal bandwidth (see 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 $\widehat{m}_n$. The first consists by finding an optimal value h_{cv} by minimizing the score function

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \widehat{m}_{n,-i}(X_i; h)\}^2 = CV_1(h) - 2CV_2(h) + (1/n)\sum_{i=1}^{n} Y_i^2.$$

The second consists by finding a theoretical expression of optimal bandwidth parameter $\widehat{h}_{opt}$ obtained by

minimizing the asymptotic part of mean integrated squared error (MISE) of $\widehat{m}_n$ such that

$$\text{MISE}\{\widehat{m}_n(x)\} = \sum_{x\in\mathbb{N}} \text{Var}\{\widehat{m}_n(x)\} + \sum_{x\in\mathbb{N}} \text{Bias}^2\{\widehat{m}_n(x)\}. \tag{4}$$

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, $\widehat{h}_{opt}$ is an approximate of the true h_{opt}. However, this optimal bandwidth is available only for estimator using discrete associated kernels satisfying (H_1)-(H_2). Thus, a theoretical expression $\widehat{h}_{opt}$ is not available for binomial kernel that does not fulfill (H2) but only for discrete triangular kernels. Moreover, we will see that $\widehat{h}_{opt}$ cannot be directly used since it requires to know the true function f. Therefore, in the following section, the estimator $\widehat{m}_n$ 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 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 paper.

Remark .

(i) First, it would be of some interest to compare, for example, $\text{MISE}(\widehat{m}_{n,h_{cv}})$ and $\text{MISE}(\widehat{m}_{n,\widehat{h}_{opt}})$. To derive such a result, some information about $\sup_{H_n} |CV_1 - 2CV_2 + (1/n)\sum_{i=1}^n Y_i^2 - \text{MISE}|$, for some sufficiently large region H_n, would be necessary. That will be the subject of a forthcoming paper.

(ii) Secondly, a common problem with cross-validation methodology is that the criteria CV is not always consistent depending on discrete kernel $\text{K}_{x,h}$ and sample $(X_i)_{i=1,2,\ldots,n}$. In this situation, for example, the function $h \mapsto CV(h)$ 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 semi-parametric 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 5 silver fir trees in 2007 presented in Figure 3.

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 5 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 diametercell enlargement and wall thickening of one silver fir tree are presented in Figure 3 (a) and (b). The observations in the two same phases of the 5 silver fir trees are plotted in Figure 3 (c) and (d).

Figure 3 about here

3.2. Mixed effect models

For wood cell count data in Figure 3 (c) and (d) which present 5 trees i observed on 31 weeks j, we apply the mixed model such that $m(\cdot) = \mathbb{E}(Y_{ij}|X_{ij} = \cdot), i = 1, 2, \ldots, 5, \quad j = 1, 2, \ldots, 31$, 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 Θ of GLM is replaced by $\Theta_i = (\theta_0 + a_{0i}, \theta_1 + a_{1i}, \theta_2 + a_{2i})$ where θ_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 5 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 $\mathrm{RMSE}_i = \sqrt{\{\sum_{j=1}^{n}(y_{ij} - \widehat{y}_{ij})^2\}/n}, \quad i = 1, 2, \ldots, 5$, where $\widehat{y}_{ij}$ 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 $\overline{\mathrm{RMSE}} = (1/5)\sum_{i=1}^{5} \mathrm{RMSE}_i$.

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 $i = 5$ (Figure 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 $i = 5$.

Table 1 and Figure 4 about here

For thickening phase, the conclusions are similar as previously. In particular, let us present the following remarks. In Figure 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).

Figure 5 about here

In the following in order to deepen our study by using bootstrap methods, a comparison is realized with GAMLSS which is an other competitive generalized structured models. We apply GLM, GAM, GAMLSS and $\widetilde{m}_n$ 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 Figure 3 (a) and (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, the performance of models is also evaluated by using the Akaike information criterion (AIC). For the number N of bootstrap samples, the average RMSE and AIC with degrees of freedom (df), rounded to integers, are presented in Table 2.

Table 2

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 Akaike information criterion, 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 problem 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 which 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 non-independence 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 modelling not only of the mean but other parameters of the distribution of Y as, linear and/or non-linear, parametric and/or additive non-parametric 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.

A. APPENDIX

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 $\widehat{m}_n$ in (3). For this, we express the discrete semiparametric estimator $\widehat{m}_n$ of the c.r.f. m as

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

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

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \widehat{m}_{n,-i}(X_i;h)\}^2 &= \frac{1}{n}\sum_{i=1}^{n}\widehat{m}^2_{n,-i}(X_i;h) - \frac{2}{n}\sum_{i=1}^{n}\widehat{m}_{n,-i}(X_i;h)Y_i + \frac{1}{n}\sum_{i=1}^{n}Y_i^2 \\ &\equiv CV_1 - 2CV_2 + \frac{1}{n}\sum_{i=1}^{n}Y_i^2, \end{aligned} \tag{5}$$

where

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

is the leave-one-out kernel estimator of $\widehat{m}_n(X_i;h)$; thus, we have $h_{cv} = \arg\min CV(h)$ with $CV(h) = CV_1 - 2CV_2$. By calculating the expectation of (5), 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 $\mathbb{E}[\sum_{x\in\mathbb{N}}\{\widehat{m}_n(x) - m(x)\}^2 f(x)]$.

Indeed, we directly see that the terms

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

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

$$CV_2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}U_{X_i,h}(X_j)Y_iY_j\frac{l(X_i;\widehat{\Theta}_{-i})}{l(X_j;\widehat{\Theta}_{-i})}$$

since we firstly have

$$\mathbb{E}(CV_2) = \mathbb{E}\Big\{\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\frac{l(X_i;\widehat{\Theta}_{-i})}{l(X_j;\widehat{\Theta}_{-i})}U_{X_i,h}(X_j)Y_iY_j\Big\} = \mathbb{E}\Big\{\sum_{j\neq 1}^{n}\frac{l(X_1;\widehat{\Theta}_{-1})}{l(X_j;\widehat{\Theta}_{-1})}U_{X_1,h}(X_j)Y_1Y_j\Big\}$$

and, secondly,

$$\begin{aligned}\mathbb{E}\Big\{\sum_{x\in\mathbb{N}}\widehat{m}_n(x)m(x)f(x)\Big\} &= \mathbb{E}\Big\{\sum_{i=1}^{n}\frac{\widehat{l}(x)}{\widehat{l}(X_i)}U_{x,h}(X_i)Y_i\mathbb{E}(Y_i|X_i=x)f(x)\Big\}\\ &= \mathbb{E}\Big[\mathbb{E}\Big\{\sum_{i=1}^{n}\frac{\widehat{l}(x)}{\widehat{l}(X_i)}U_{x,h}(X_i)Y_i^2f(x)\Big\}\Big|X_i=x\Big]\\ &= \mathbb{E}\Big\{\sum_{i=1}^{n}\frac{\widehat{l}(X_1)}{\widehat{l}(X_i)}U_{X_1,h}(X_i)Y_i^2\Big\}.\end{aligned}$$

Note that the estimate $\widehat{\Theta}_{-i}$ is calculated as $\widehat{\Theta}$ 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 $\widehat{h}_{opt}$ for estimator $\widehat{m}_n$ with discrete associated kernels satisfying (H1′)-(H2′).

A.2. Minimization of asymptotic part of mean integrated squared error

This approach required to calculate bias and variance of estimator $\widehat{m}_n$ since MISE in (4). Let us assume $l_0(x) = l(x;\Theta_0)$ be a fixed parametric start in (2), i.e. $m = l_0\omega$, such that $\widehat{\Theta}$ converges to Θ_0 in probability. For $x \in \mathbb{N}$, the discrete semiparametric estimator $\widehat{m}_n$ in (3) using discrete triangular kernels admits the following bias:

$$\text{Bias}\{\widehat{m}_n(x)\} = \frac{h}{2}\text{V}(\mathcal{K}_{a;x,h})W(x) + O\left(\frac{1}{n}+h^2\right) + o(h),$$

with

$$\sum_{x\in\mathbb{N}}\{W(x)\text{V}(\mathcal{K}_{a;x,h})\}^2 = \sum_{x\in\mathbb{N}}\Big[\text{V}(\mathcal{K}_{a;x,h})\left\{l_0(x)\omega^{(2)}(x) + 2l_0(x)\omega^{(1)}(x)\left(f^{(1)}/f\right)(x)\right\}\Big]^2 < \infty;$$

then, for its variance, we have

$$\text{Var}\{\widehat{m}_n(x)\} = \frac{\text{Var}(Y|X=x)}{nf(x)}\{1 - h\text{A}(\mathcal{K}_{a;x,h})\}^2 + o\left(\frac{1}{n}\right) + O(h^2).$$

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 :

$$\begin{aligned}\mathrm{MISE}\{\widehat{m}_n(x)\} = \sum_{x\in\mathbb{N}} \mathbb{E}\{\widehat{m}_n(x) - m(x)\}^2 &= \mathrm{AMISE}(h) + o\Big(\frac{1}{n} + h^2\Big) + O(h^2) \\ &\equiv \mathrm{MISE}(h),\end{aligned}$$

where AMISE(h) is the main term in the sum of integrated variance and squared bias of $\widehat{m}_n$. The bandwidth $\widehat{h}_{opt} = \arg\min_{h>0} \mathrm{AMISE}\{\widehat{m}_n(x)\}$ comes by solving the following equation $\mathrm{d}\{AMISE(h)\}/\mathrm{dh} = 0$ which is equal to

$$h\mathrm{V}^2(a)\sum_{x\in\mathbb{N}}\{W(x)\}^2 - \sum_{x\in\mathbb{N}}\mathrm{A}(a)\{1 - h\mathrm{A}(a)\}\frac{\mathrm{Var}(Y|X=x)}{nf(x)} = 0.$$

That leads to expression

$$\widehat{h}_{opt}(a; n, f) = \frac{A(a)\sum_{x\in\mathbb{N}}\mathrm{Var}(Y|X=x)/f(x)}{\mathrm{A}^2(a)\sum_{x\in\mathbb{N}}\mathrm{Var}(Y|X=x)/f(x) + n\mathrm{V}^2(a)\sum_{x\in\mathbb{N}}W^2(x)} \tag{6}$$

such that $\widehat{h}_{opt} \to 0$ when $n \to \infty$. The following asymptotic relationship can be pointed out: $\widehat{h}_{opt} \sim k_0 n^{-1}$ with

$$k_0 = \frac{\mathrm{A}(a)\sum_{x\in\mathbb{N}}\mathrm{Var}(Y|X=x)/f(x)}{\mathrm{V}^2(a)\sum_{x\in\mathbb{N}}W^2(x)}.$$

Acknowledgements. We 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 thank also the Associate Editor and anonymous referee for hints which improved the paper.

References

[1] 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.

[2] Alberts, T., Karunamuni, R.J. (2003). A semiparametric method of boundary correction for kernel density estimation. *Statistics & Probability Letters* 61: 287–298.

[3] Bates, D., Maechler, M., Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-42. http://CRAN.R-project.org/package=lme4.

[4] 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.

[5] 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*, DOI: 10.1093/jxb/ert057.

[6] 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.

[7] FAO, 2007. *Global wood and wood products - Flow trends and perspectives*. Advisory Committee on Paper and Wood Products, Shanghai, China, 6 June, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy.

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

[9] 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.

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

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

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

[13] 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.

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

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

[16] Pardinas, R., Sperlich, S. (2010). Feasible Estimation in Generalized Structured Models. Statistics and Computing 20: 367-379

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

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

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

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

[21] 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

[22] Wodzicki, T. J., Zajaczkowski, S. (1970). Methodical problems in studies on seasonal production of cambial xylem derivatives. Acta societatis botanicorum poloniae 39: 509–520

[23] Wood, S. (2011). *gamm4: Generalized additive mixed models using mgcv and lme4*. R package version 0.1-5. http://CRAN.R-project.org/package=gamm4. http://CRAN.R-project.org/package=gamm4.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

FIGURES AND TABLES

Table 1: RMSE resulting from estimations of the number of wood cells weekly counted in the diameter enlargement and wall thickening phases of 5 silver fir trees during 2007, by applying GLMM, GAMM and semiparametric regression estimator with binomial kernel

Phase	Model	Tree					$\overline{\text{RMSE}}$
		1	2	3	4	5	
Diam. enlarg.	GLMM	6.421	8.942	12.434	11.045	17.824	11.333
	GAMM	5.632	11.580	8.132	9.179	15.181	9.941
	Semip. regr.	4.584	4.653	6.603	5.559	7.705	5.821
Wall thick.	GLMM	22.582	40.584	24.664	49.383	34.041	34.251
	GAMM	18.897	29.495	18.601	33.210	29.296	25.900
	Semip. regr.	9.014	11.897	7.943	12.654	16.630	11.628

Table 2: $\overline{\text{RMSE}}$ and $\overline{\text{AIC}}$ resulting from estimations of the number of wood cells weekly counted in the diameter enlargement and wall thickening phases of 5 silver fir trees during 2007, by applying GLM, GAM, GAMLSS and semiparametric regression estimator with binomial kernel

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

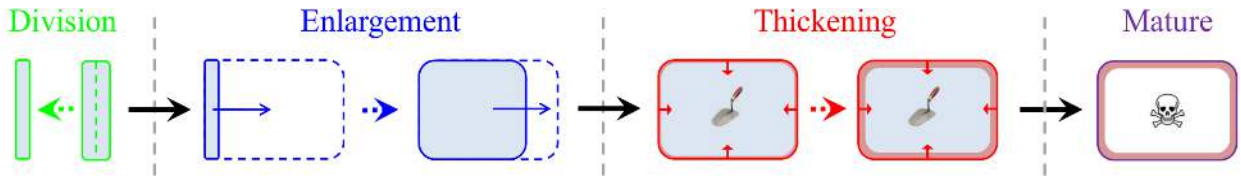

Figure 1: Diagram showing the development of a tracheid, from its production by division of a cambial cell to its mature and functional state

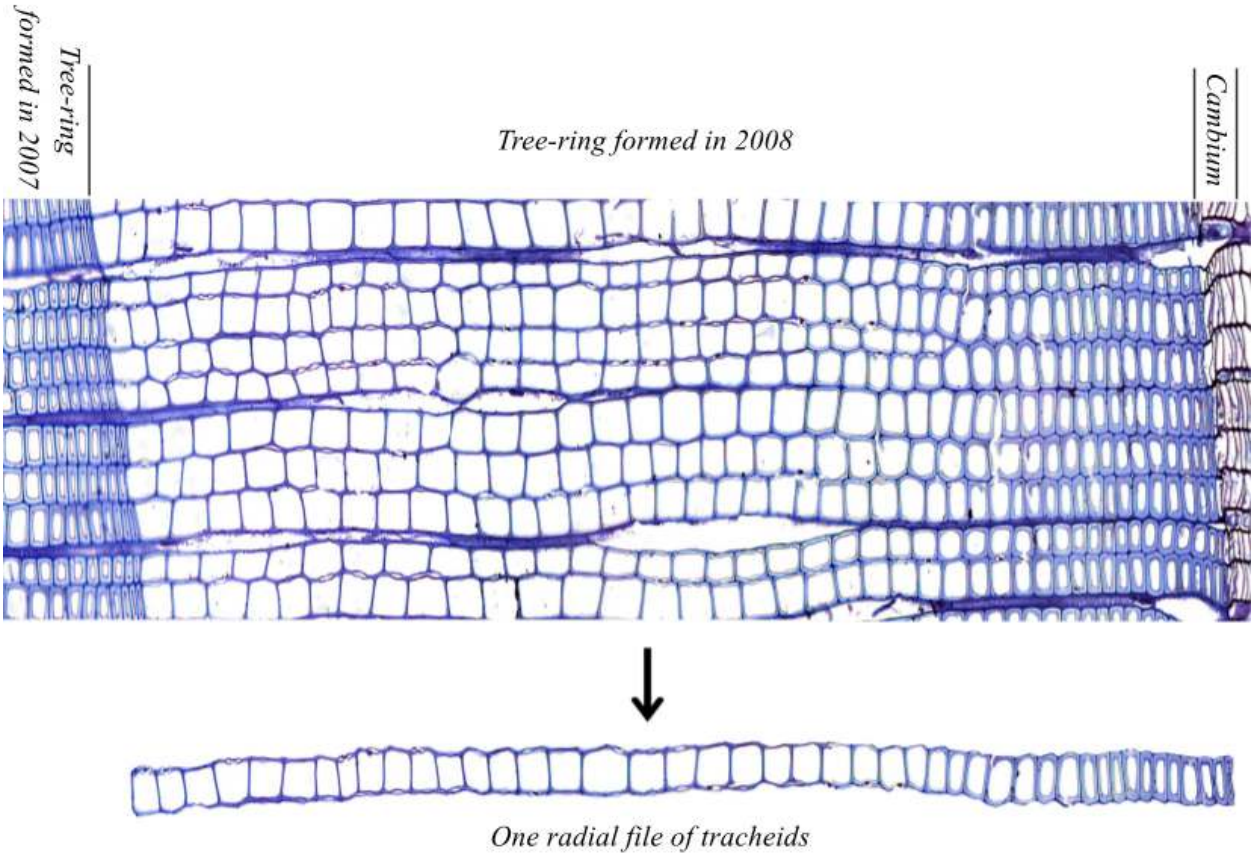

Figure 2: Anatomical transverse section of wood cut from a sample collected in October 2008 on the stem of a silver fir tree. The section is 6 μm in thick, stained with cresyl violet acetate and observed under light microscope ($\times$100)

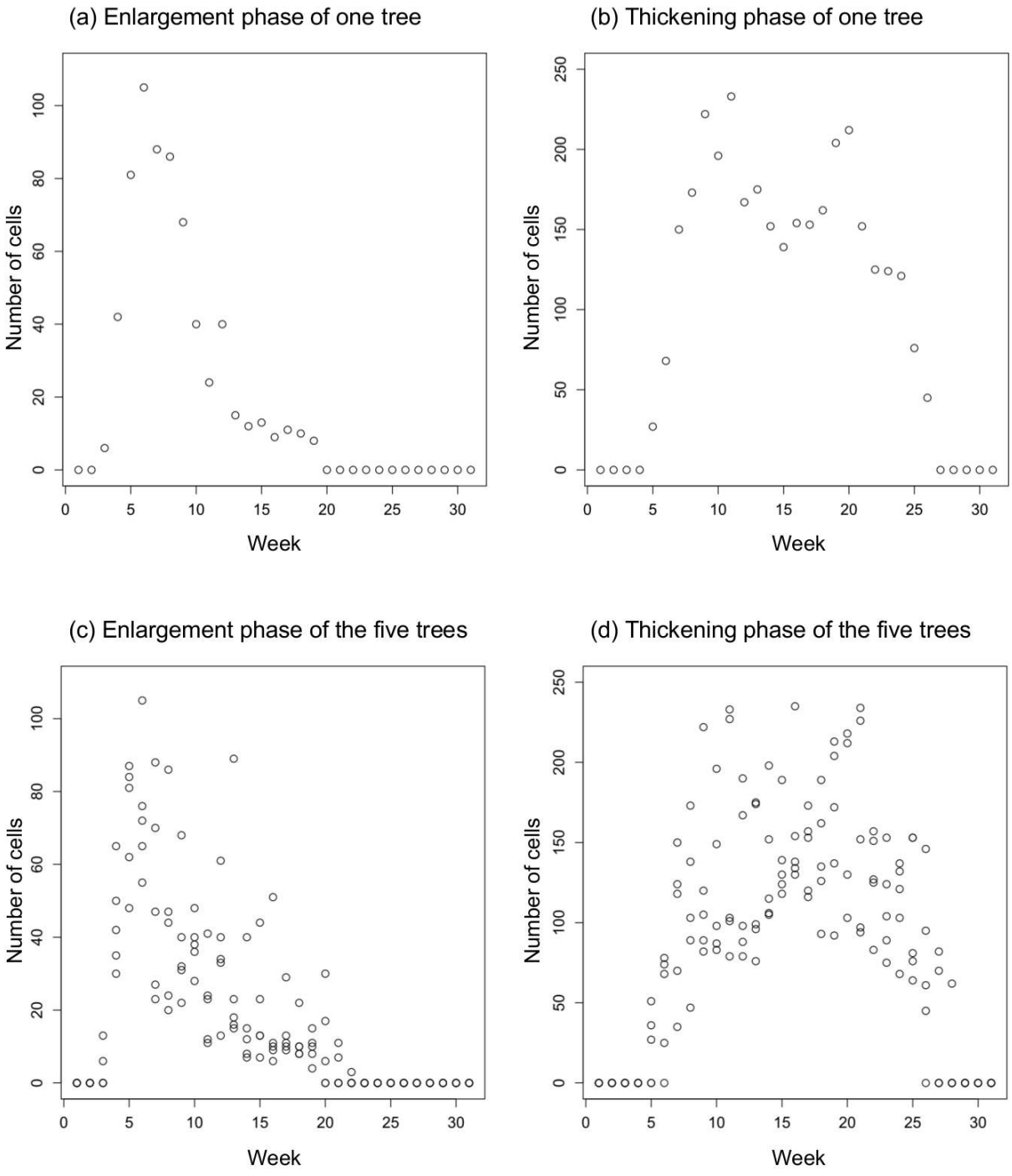

Figure 3: Number of wood cells (×10) weekly counted in the diameter enlargement and wall thickening phases of five silver fir trees during 2007

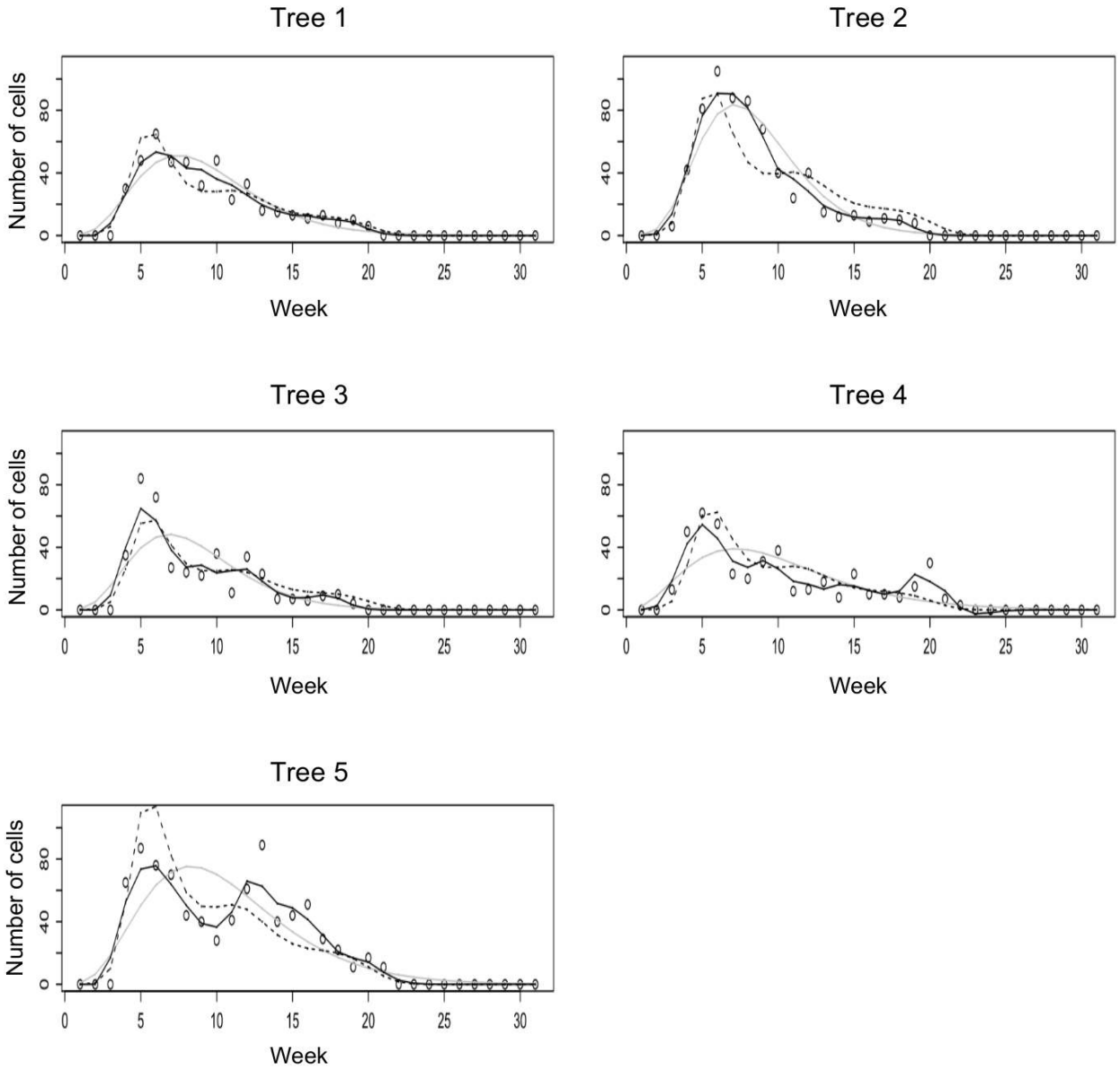

Figure 4: Individual estimations of the number of wood cells (×10, circles) weekly counted in the diameter enlargement phase of 5 silver fir trees during 2007 by applying GAMM (black dotted lines), GLMM (grey lines) and semiparametric binomial kernel regression (black lines)

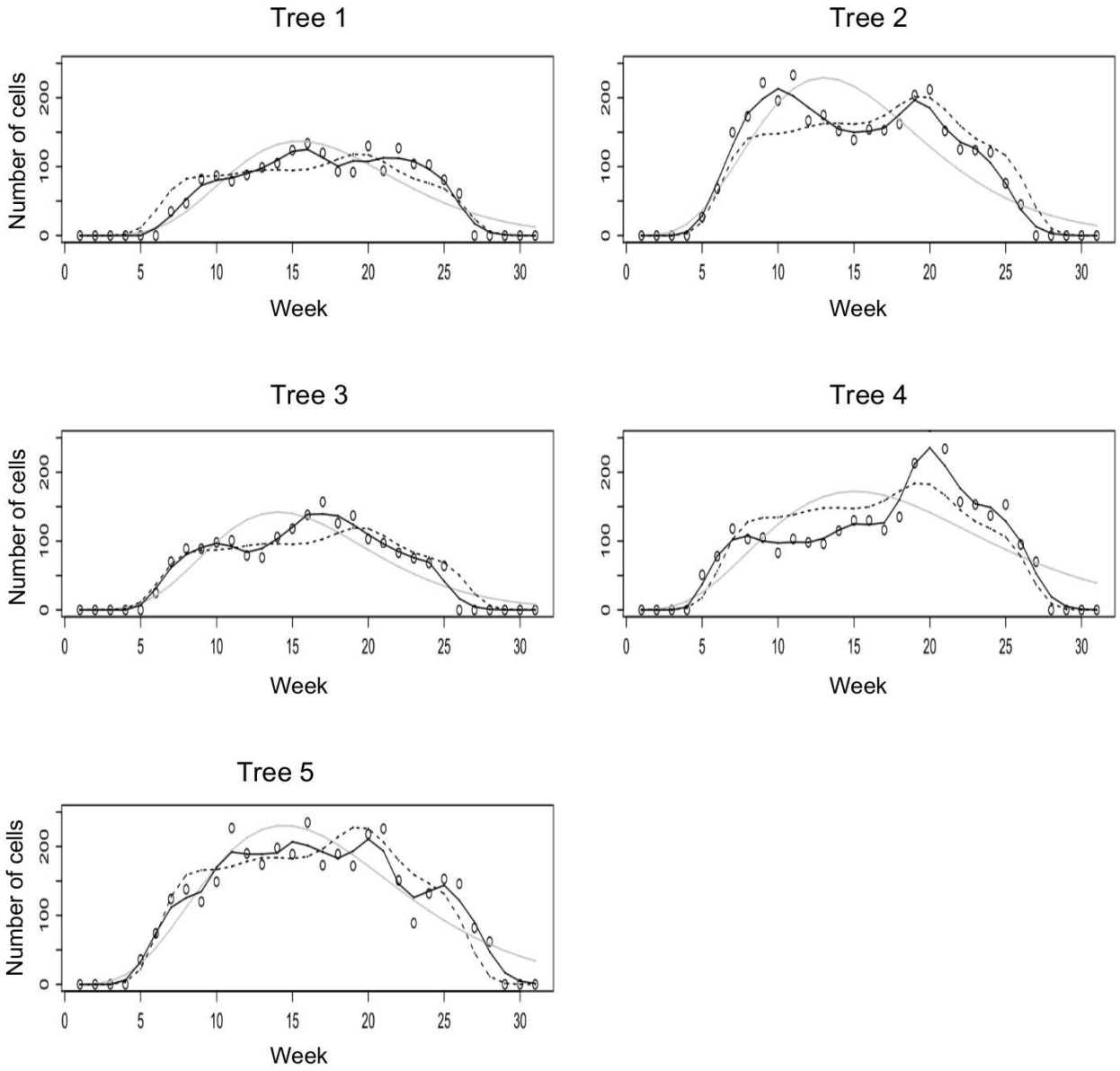

Figure 5: Individual estimations of the number of wood cells (×10, circles) weekly counted in the thickening phase of 5 silver fir trees during 2007 by applying GAMM (black dotted lines), GLMM (grey lines) and semiparametric binomial kernel regression (black lines)