Projet ECLID

Extrêmes CLimatiques et Dendrochronologie

Fondation MAIF

Coordinateur : P. Yiou

Laboratoire des Sciences du Climat et de l'Environnement,
UMR CEA/CNRS/UVSQ

Durée du projet : 36 mois

Rapport No. 3 (du 1.1.2010 au 15.12.2010)

Rédigé le 24 décembre 2010

Table des matières

1 Avancement des travaux **2**

2 Mission à l'étranger **3**

3 Conférences **4**

4 Publications **6**

4.1 Extracting hidden trends in tree rings with a semi-parametric Bayesian hierarchical model 6

4.2 Bayesian variables selection for Generalized Additive Models applied to climatic reconstructions 36

1 Avancement des travaux

Durant cette dernière année de projet, nous avons suivi deux voies complémentaires pour le projet ECLID. Conformément aux engagements pris lors du dernier comité de suivi du projet, Ophélie Guin s'est rendue à Zurich afin de discuter avec le groupe de Paléoclimatologie de D. Frank sur l'utilisation de données dendroclimatiques, et la confrontation de son approche avec les techniques traditionnelles. Ceci a permis d'obtenir des données ayant une bonne couverture spatiale, et d'appliquer les techniques développées pendant le projet à une base de données dendroclimatiques très complète. Le traitement des données est en soi très lourd en temps de calcul car le nombre de séries est très grand (plusieurs centaines) et les méthodes probabilistes utilisées demandent des simulations d'un grand nombre de variables aléatoires.

Ophélie a aussi continué les développements méthodologiques autour de la modélisation statistique de la croissance des arbres. Ces travaux ont principalement porté sur la sélection des variables climatiques indépendantes qui expliquent au mieux les variations dendrochronologiques, dans un cadre bayésien. Dans ce cadre, un modèle additif généralisé de dépendance entre la croissance d'un cerne et des variables climariques est paramétré pour répondre à un modèle additif généralisé. La difficulté qui a été résolue, face à un trop grand nombre de variables potentielle pour expliquer une épaisseur de cerne, a été de se placer dans un cadre bayésien pour la sélection de variable. Cette nouvelle méthode a été testée sur des données simulées et appliquée à des mesures de densité de cernes d'arbres (*Pinus halepensis Mil l.*) enregistrées sur la côte Méditerranéenne française.

Un article a été soumis au Journal of the American Statistical Association, au cours de cette année. Un autre est en préparation pour une soumission prochaine.

Les applications climatiques et environnementales ont demandé plus de temps que prévu (notamment à cause de la lourdeur des calculs). Ceci explique un retard de deux mois dans l'avancement de la thèse d'Ophélie. Ce retard dans la soutenance sera financé par un autre projet portant sur la dendroclimatologie et la paléoclimatologie (projet ANR MOPERA), auquel le projet ECLID apporte une expertise en statistiques. Nous prévoyons une soutenance de thèse à la fin du mois de mars 2011 (i.e. avec un retard de 2 mois par rapport à la date prévue de fin du projet). Le manuscrit de thèse et la soutenance publique constitueront le rapport final et définitif du projet.

2 Mission à l'étranger

Les 15 et 16 avril 2010, Ophélie Guin s'est rendue au Swiss Federal Research Institute (WSL) à Zurich afin d'y rencontrer David Frank, spécialiste de dendroclimatologie. Le but de la rencontre était d'avoir l'avis de ce spécialiste sur les techniques statistiques dévelopées dans le projet, et leur application sur la base de données qu'il développe.

Ophélie a tout d'abord fait une présentation de nos travaux, et en particulier du modèle que nous avons mis en place pour extraire un signal commun à partir d'un jeu de cernes d'arbres (voir section 4), devant l'équipe de Paléoclimatologie. Cette présentation a été bien accueillie par l'équipe et a permis une discussion scientifique intéressante. L'équipe de Paléoclimatologie étant principalement composée de dendrochronologues, les questions ont été d'ordre technique afin de s'assurer de la compréhension de notre modèle statistiques. Ensuite, la discussion a dévié sur une question fondamentale : est-ce que notre modèle est meilleur que les techniques classiques ? Que peut-il apporter de plus à la dendrochronologie ?

Ophélie a longuement discuté des techniques classiques de la dendrochronologie avec David Frank, ce qui lui a permis d'éclaircir certains points techniques sur les techniques souvent employées en "boîtes noires" par les dendroclimatologues. Il a également montré comment se servir du logiciel ARSTAN, couramment employé par les dendrochronologues.

La conclusion de cette visite a été de se lancer dans une comparaison poussée des résultats que l'on obtient à l'aide notre modèle bayésien hiérarchique avec ceux obtenus par des méthodes classiques de la dendrochronologie. Le but est de définir précisément les avantages et les inconvénients de chacune des méthodes les unes par rapport aux autres. Pour cela, David Frank nous a fourni les données qui ont été utilisées pour l'article de Büntgen et al. (2006), qui reconstruisait les températures d'été pour les Alpes européennes. Ce jeu de données ayant une taille importante (1270 ans et 180 séries de largeurs de cernes d'arbres) il a fallu un certain temps pour le faire tourner sur notre modèle et obtenir des résultats. Ces résultats viennent juste d'être obtenus et nous devont les interpréter dans les semaines qui viennent.

3 Conférences

- Extraction de signaux temporels à l'aide d'un modèle Bayésien Hiérarchique - 6ème Rencontres Statistiques de Rochebrune, 29 mars au 1 avril 2010 - Oral

  Résumé : Les statistiques sont devenues une composante essentielle des reconstructions climatiques, qui sont elles-mêmes extrêmement importantes pour la quantification du réchauffement climatique global. Ainsi, ces quelques dernières années, il y a eu un effort de recherche scientifique important afin de combiner de manière spatiale et temporelle différents proxies (c'est-à-dire des mesures indirectes du climat). Généralement, les statisticiens ne travaillent pas directement avec des mesures de proxies brutes mais passent par une étape de pré-traitement, appelée standardisation et mise en place dans le but d'extraire un signal climatique pertinent pour chaque proxy. Ce papier s'intéresse tout particulièrement à cette étape pour l'un des proxy les plus utilisé : les mesures de cernes d'arbres. En revenant aux données brutes, on cherche à améliorer l'analyse statistique des mesures de cernes d'arbres dans l'espoir d'améliorer les reconstructions climatiques.
  L'un des principes de base de la dendroclimatologie est que les cernes d'arbres sont supposés contenir des informations sur le climat passé. D'un point de vu statistique, ce problème d'extraction peut être vu comme la recherche d'une variable cachée qui représente un signal commun à une collection de séries de mesures de cernes d'arbres. En comparaison avec les études dendroclimatologiques passées, nous proposons un modèle bayésien hiérarchique semi-paramétrique qui offre la possibilité de capturer les hautes et les basses fréquences contenues dans les cernes d'arbres. Notre modèle est testé sur des données simulées et appliqué à des mesures de densité de cernes d'arbres (*Pinus halepensis Mill.*) enregistrées sur la côte Méditerranéenne française.

- Sélection bayésienne de variables pour les modèles d'état dans le cadre de reconstructions climatiques - 42ème Journées de Statistique, 24 au 28 mai 2010, Marseille (France) - Oral

  Résumé : De nombreuses variantes sur la sélection de variables pour un modèle de régression sous l'approche bayésienne ont été proposées dans la littérature. Dans cette présentation, nous adaptons cette méthode de sélection de variables à un modèle d'état, ce qui revient à ajouter une équation à notre modèle de régression. On applique cette méthode à un problème de reconstruction climatique. En effet, afin de comprendre si le réchauffement climatique actuel est plus important que la variabilité climatique naturelle, il est nécessaire d'avoir de longues séries climatiques. Des mesures directes de température ou de précipitations manquent, en particulier pour les périodes les plus anciennes, et il est nécessaire d'utiliser des proxies climatiques afin de reconstruire des chronologies passées. Un proxy bien connu est la croissance des cernes d'arbres. Afin de comprendre les relations existantes entre ce proxy et des variables climatiques on essaie d'expliquer la croissance des cernes d'arbres avec la meilleure combinaison possible de séries de températures et de précipitations.

4 Publications

4.1 Extracting hidden trends in tree rings with a semi-parametric Bayesian hierarchical model

Cet article a été soumis à Journal of the American Statistical Association.

Résumé : Les statistiques sont devenues une composante essentielle des reconstructions climatiques, qui sont elles-mêmes extrêmement importantes pour la quantification du réchauffement climatique global. Ainsi, ces quelques dernières années, il y a eu un effort de recherche scientifique important afin de combiner de manière spatiale et temporelle différents proxies (c'est-à-dire des mesures indirectes du climat). Généralement, les statisticiens ne travaillent pas directement avec des mesures de proxies brutes mais passent par une étape de pré-traitement, appelée standardisation et mise en place dans le but d'extraire un signal climatique pertinent pour chaque proxy. Ce papier s'intéresse tout particulièrement à cette étape pour l'un des proxy les plus utilisé : les mesures de cernes d'arbres. En revenant aux données brutes, on cherche à améliorer l'analyse statistique des mesures de cernes d'arbres dans l'espoir d'améliorer les reconstructions climatiques. L'un des principes de base de la dendroclimatilogie est que les cernes d'arbres sont supposés contenir des informations sur le climat passé. D'un point de vu statistique, ce problème d'extraction peut être vu comme la recherche d'une variable cachée qui représente un signal commun à une collection de séries de mesures de cernes d'arbres. En comparaison avec les études dendroclimatologiques passées, nous proposons un modèle bayésien hiérarchique semi-paramétrique qui offre la possibilité de capturer les hautes et les basses fréquences contenues dans les cernes d'arbres. Notre modèle est testé sur des données simulées et appliqué à des mesures de densité de cernes d'arbres (*Pinus halepensis Mill.*) enregistrées sur la côte Méditerranéenne française.

Extracting hidden trends in tree rings with a semi-parametric Bayesian hierarchical model

Ophélie Guin[1], Philippe Naveau[1], Jean-Jacques Boreux[2]

[1]Laboratoire de Sciences du Climat et de l'Environnement, IPSL-CNRS, France

`ophelie.guin@lsce.ipsl.fr` and `naveau@lsce.ipsl.fr`

[2]University of Liège, Arlon, Belgium

`jj.boreux@ulg.ac.be`

December 3, 2010

Abstract

Statistics have become an essential component in the field of climate reconstructions, which is an important topic in quantifying global warning amplitude. In the last few years, there has been an important statistical research effort to spatially and temporally combine different climate proxies (i.e. indirect measurements). Still, it is unfrequent for the statistician to work directly with raw proxy measurements. Typically, a preprocessing step, often called standardization, is implemented to extract the relevant climatic signal in each proxy. This paper focuses on this preprocessing stage for the most used climate proxy, tree ring measurements. By going back to the

data source, we focus on improving the statistical analyses of the original tree ring measurements, and this could ultimately improve climate reconstructions.

One basic premise of dendroclimatology is that tree rings are assumed to contain hidden information about past climate. From a statistical perspective, this extraction problem can be understood as the search of a hidden variable, which represents a common signal within a series of tree ring measurements. Compared to past dendroclimatology studies, we propose a semi-parametric Bayesian hierarchical model that offers the possibility to capture hidden low and high frequencies in tree rings. Our new model is tested on simulated data and applied to tree rings density measurements (*Pinus halepensis Mill.*) recorded in French Mediterranean.

1 Dendrochronology and statistical climatology

Recently there have been a strong interest among statisticians, politicians and even bloggers concerning the role of statistics within the scientific climate change debate, e.g. see the transcript of the ASA discussion “Statisticians Comments on Status of Climate Change Science", March 2010
`http://magazine.amstat.org/blog/2010/03/01/climatemar10/`,
or the recent JASA comments of the article by Li et al. (2010). One key issue to understanding past and recent climate changes is to derive, study and apply efficient statistical procedures to reconstruct past records of temperatures and precipitation. Direct measurements of such climatological variables are missing whenever the instrumental record length is shorter than the period of interest. The so-called proxies, i.e. indirect measurements, offer the material to reconstruct past chronologies in such situations.

Tree ring measurements may be the most well known and common climate proxy. Since the work of Douglass (1920, 1936), there has been an active and extensive research activity dedicated to the field of dendrochronology (dendron = tree and chronos = time) that study tree rings to analyze temporal and spatial patterns of processes in the physical and social sciences. Journals like Tree-Ring Research (formerly Tree-Ring Bulletin) and Dendrochronologia, numerous books (e.g. Cook and Kairikukstis, 1990; Gornitz, 2009) and thousands of articles show the vitality and the importance of tree rings in many fields, e.g. forest ecology, climatology, archaeology and botany. Within the realm of reconstructions studies, dendroclimatology focuses on identifying links between tree rings information and climate variables. Implicitly it is assumed that a climatic signal can be hidden into tree ring growths. To illustrate the importance of dendrochronology in climatology, we recall the important and actively commented papers of Mann et al. (1999) and Esper et al. (2002) that used some tree ring data to reconstruct Northern Hemispheric annual temperatures for the last millennium. One heated point of discussion in the global climate warming debate was the statistical analysis of tree ring data in these two papers (Committee on Surface Temperature Reconstructions for the Last 2000 Years, 2006; Mann et al., 2008). To integrate the information from different sources, Li et al. (2010) studied a Bayesian hierarchical model to reconstruct past temperatures and they assessed their method via synthetic data generated from a global climate model. The recent paper by McShane and Wyner (2010) proposed a different temperatures reconstruction which behaves similarly to past reconstructions but has much wider standard errors. Smith (2010) highlighted the sensitivity of paleoclimatic reconstructions to the time period of observational data and to the selection of proxies. These articles underline the difficulty of analyzing proxies and reconstructing past climate variables. In contrast to this recent research our goal is neither to propose a new reconstruction of past temperatures neither to develop a novel way to combine different proxies. By focusing on a single proxy (tree rings), our main scope is to propose a novel statistical scheme to extract the most relevant climatological information from tree rings. In other words we believe that improving the statistical analysis of raw tree ring data, the building block of most reconstruction studies, could eventually lead to better reconstructions. In addition, the method proposed here could be applied to other proxies used in environmental sciences.

One major advantage of dendrochronology over other dating techniques is that annual ring formation makes the time sampling, one ring per year, constant in zones that have a distinct dormant season related to cold weather (most tropical tree species, not studied here, may not produce distinctive annual growth rings (Stahle, 1999)). A recurrent difficulty associated with the temporal scale resides in the tree lifetime heterogeneity. Figure 1 shows the lifetime of the fourteen trees that are used in our applications. The x-axis corresponds to the years and the y-axis to the tree label. Each individual tree has a different lifetime and some trees like 4 has a short record while others like 1 contains more information. Typically the number of sampled trees diminishes as one go back in time. Finding older trees becomes more and more arduous for the field experimenter. This classical issue in paleo-studies implies that the assessment of uncertainty can be non-trivial and should vary in time.

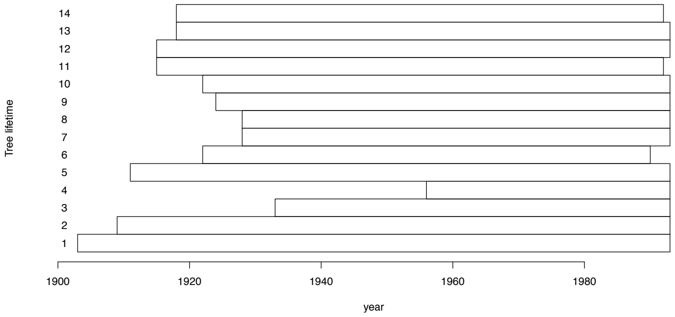

Figure 1: The lifetime of the fourteen trees that has been used in our application. The x-axis corresponds to the years and the y-axis to the tree label.

Another statistical difficulty in dendroclimatology concerns the delicate choice of the explanatory variables and their time scales. Should the tree ring growth be correlated to the average of daily precipitation over the summer months, the largest number of consecutive days without rain during one year, a function of seasonal temperatures or any other choice? The number of possibilities is nearly endless and depends on the tree species and the region of interest. Hence the dendrochronologue expertise is invaluable to pre-select possible meaningful explanatory variables and this sometimes allows the statistician to view a tree ring reconstruction problem as a variable selection problem within an inverse regression procedure. In this paper we decouple tree ring analysis from the selection problem by treating a different statistical question. Given tree rings measurements from a given site, how should one extract a hidden common signal from this tree ring data set? Our underlined assumption is that the common signal shared by all the trees from a particular site should be due to an environmental factor, possibly climatic but not necessarily. The clear advantage of this inquiry is that the extraction of the common component does not depend on an arbitrary choice of explanatory variables and therefore, the common signal extraction is clearly decoupled of the selection problem and so can be interpreted independently. This leaves the possibility that the extracted signal may be linked to non-climatic variables. The main drawback is that the interpretation of the extracted signal remains an open question. This issue will be discussed in Section 3.2.

A classical decomposition to represent yearly individual tree ring growths is the following additive model, often called the linear aggregate model (Cook, 1990; Buckley, 2009),

$$\text{individual tree-growth} = G_t + F_t + D_t + \text{unexplained variability} \tag{1}$$

where t represents a year, G_t corresponds to the age-related trend due to normal physiological aging processes (see Figure 2), F_t to the climatically-related environmental signal and D_t to disturbance factors, either within the forest stand or outside of it (e.g., insect outbreaks or fires). In most studies, the site of interest is selected in order to minimize the possibility of internal and external ecological processes affecting tree growth. In this paper, we follow this hypothesis and D_t is set to zero. Concerning G_t, Figure 2 displays the idealized tree age effect curve over time. The juvenile stage with a rapid growth is followed by a mature stage with a fairly constant growth rate and finally a senescent phase terminates the life cycle of the tree. These phases are difficult to capture in actual tree growth time series. Given a set of trees from the same species, site and environmental surroundings, the variability among individual age effect components has to be taken into account in order to discriminate between the distinct environmental signal shared by trees of different ages and each individual's own age effect. Although idealized, the scheme in Figure 2 provides important *a priori* information about the age effect. It corresponds to a smooth (low frequency) signal and we expect a rather concave shape. These two pieces of information are rather vague and can be sharpen according to the tree species and region under study. In this paper the frequency information has been used to guide some of our prior distributions choice within our Bayesian modeling. The concavity of the age effect curve has not been imposed *a priori* and serves rather, as an yardstick to discuss our data analysis.

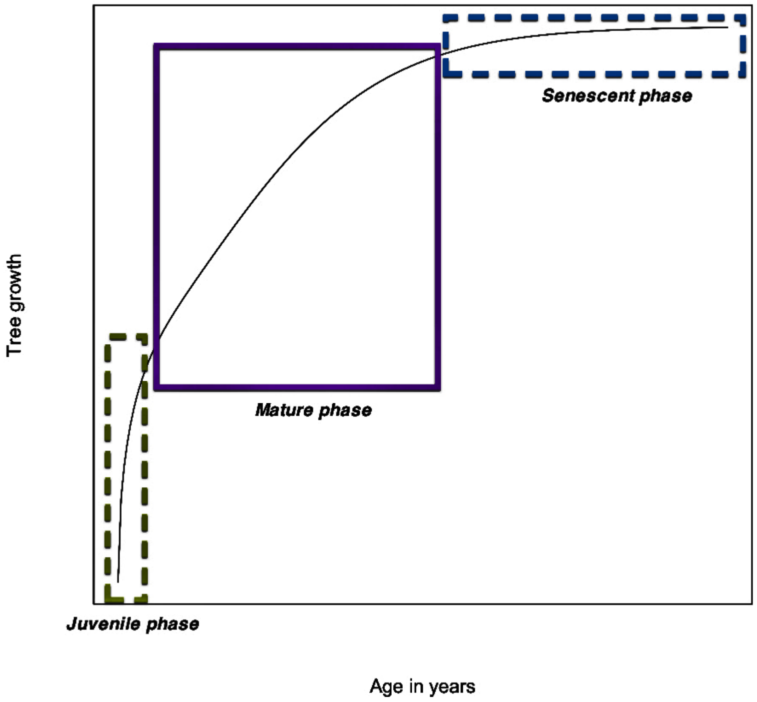

Figure 2: Idealized tree age effect behavior over time. After a juvenile phase (youth) with an accelerating rate of growth, the tree enters a mature phase with a roughly constant rate of growth, follows by a senescent phase with a decelerating rate of growth. In practice it is difficult to statistically identify with three phases because of changing environmental and internal factors.

One of the main dendroclimatologist interests resides in finding the component F_t in Equation (1). This quest leads to the so-called *standardization* problem and remains an object of active research (Melvin and Briffa, 2008; Nicault et al., 2010). Basically individual trees at an environmentally homogenous site can have their own physiological aging process G_t, see Figure 1. They can also share a common element due to the local environment. Standardization aims at calculating a dimensionless yearly index that reflects this hidden common environmental chronology. The most popular standardization approach proposed by dendroclimatologists can be summarized by the following steps (e.g., Melvin and Briffa, 2008). First an age-related trend is estimated and removed individually for each measurement to eliminate the age-affect G_t. This is classically done by implementing an univariate parametric regression (e.g., negative exponential curve (Fritts et al., 1969)) or a semi-parametric one (Cook and Peters, 1981; Barefoot et al., 1974). Second each measurement is divided by the corresponding fitted value obtained from the regression. This produces the so-called tree indices that should have a mean of approximately equal to one. Third the so-called chronology time series, i.e. the standardized dimensionless index, is calculated as the arithmetic mean of all tree indices for a year. The underlining model beneath this series of statistical steps is similar to a multiplicative model, i.e. instead of working directly with the raw tree-growth measurements, their logarithms are modeled by (1). The inference aspect of this standardization approach is not clear. Each step is made independently of the previous one. Consequently, calculating valid estimates and confidence intervals of the final output, the dimensionless index, remains challenging. The common hidden variable of interest should make the inference fully multivariate. In other words, univariate techniques have been used at each step while the problem is multivariate by nature and inferences made of each step are decoupled from each other. This later issue leads to another drawback. By construction, the classical standardization scheme takes out all the low frequency information contained in tree rings. This due to the removal of the age-effect. Individually an univariate regression cannot make the distinction between two low frequency components, see G_t and F_t in (1). Only, by treating the full set of trees jointly, one can hope to discriminate between a common smooth climate signal and other individual ones. For the practitioner, this drawback is very important. It implies that the classical standardization scheme is only adapted to capture annual variability but not decadal or centennial trends from tree rings. This is also true for other standardization based on ARMA modeling (Guiot, 1987). Recently Boreux et al. (2009) proposed and studied a Bayesian hierarchical model to extract hidden signal but again, it was under the hypothesis that smooth trends have already been removed by a preprocessing of individual tree rings. The Regional Curve Standardization (RCS) and the Adaptive Regional Growth Curve (Nicault et al., 2010) are attempts to preserve low frequency climatic information contained into tree rings. The former is based on producing a global biological growth trend obtained by averaging ring widths that have been aligned according to their biological age (not their chronological age). This requires a large number of trees. Another assumption here is that this structural form is the same for each tree and does not vary in time. Coming back to (1), this means that G_t comes from an unique profile that has been shifted according to the tree age. This is rather strong limitation because individual growth rate trees can differ according to soil conditions, competition and other factors governing productivity. To circumvent this issue, Nicault et al. (2010) proposed to regress tree rings according to cambial age, initial and maximum productivities using a neural network. The initial and maximum productivities are defined as the average of the first 10 rings and the maximum value during the first 50 years over an individual smoothed growth profile, respectively. Hence the computation of the predictors is tailored to the application at hand and may be difficult to generalize to other cases without an expert in dendrochronoloy. In addition, the inference properties of the method are not clear to us because tree rings seem to be used as predictant and as data for building the predictors.

To summarize our objectives, we aim to propose and study a multivariate model and global inference scheme capable of extracting hidden individual and common trends. Essential elements of our analysis are the modeling of varying uncertainties due to tree lifetime heterogeneity, bypassing the need of parametric forms for either individual or common trends and taking into account the prior information given by dendroclimatologists. To exemplify and discuss our approach, we have analyzed a set of fourteen *Pinus halepensis Mill.* Figure 3 localizes the site with geographical coordinates (5˚28’E, 43˚4’N) named “Les Pennes-Mirabeau" and situated along the French Mediterranean coast where tree ring measurements were studied by Nicault et al. (2001). This region is climatically characterized by a Mediterranean climate with clear summer droughts. Nicault et al. (2001) identified possible relationships between tree growth measurements and climatic factors in the same geographical region and with the same tree species. Hence this past study provides a referential for our extraction procedure and has been beneficial for discussing and interpreting our approach. Figure 4 displays fourteen *Pinus halepensis Mill* tree ring density time series

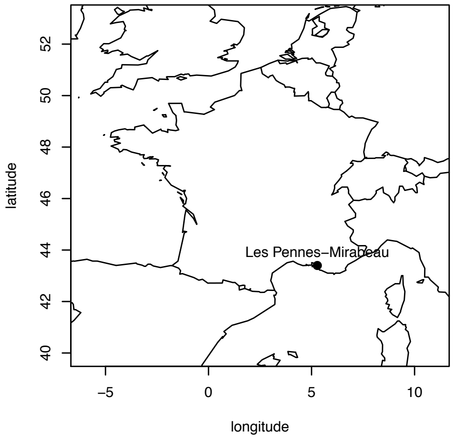

Figure 3: The “Les Pennes-Mirabeau" site located in the South of France where *Pinus halepensis Mill* tree ring densities series shown in Figure 4 were recorded.

(in mg/cm3) from the “Les Pennes-Mirabeau" site. The group of fourteen time series illustrates the difficulty of finding a common signal; each time series having its own time length (see Figure 1), its own growth trend and a large variability. To conclude this short description of the data set, we would like to add that the choice of studying ring density profiles over other dendrochronological variables like tree ring growths is rather arbitrary. For this site, our method has also been applied to tree ring growth measurements, to its logarithm

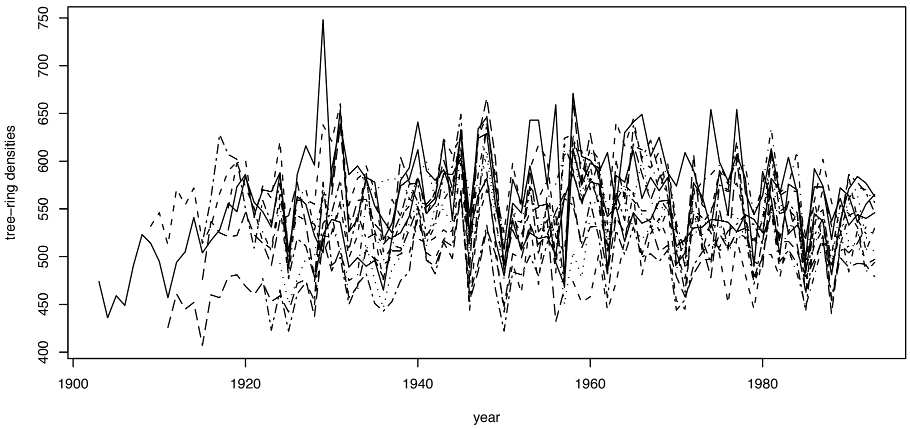

Figure 4: Fourteen *Pinus halepensis Mill* tree ring density time series (in mg/cm3) from the “Les Pennes-Mirabeau" site located in Figure 3. The x-axis (years) covers the period $1903 - 1993$ and each time series has a different length, see Figure 1.

and to the wood density logarithm. The extracted hidden common signal for each random variable type appears to be very similar. Consequently we only study one type: tree ring density profiles.

2 Model description and its inference

During the last two decades, Bayesian Hierarchical Models (BHM) have blossomed in climate sciences. One appealing idea in BHMs is to probabilistically decompose a complex climatic process and its relationships to observations in several simple components throughout a hierarchy of layers. BHMs handle efficiently the uncertainty assessment of each layer by clearly identifying prior and posterior distributions of underlining processes. For an introduction to such models, see e.g. Gelman et al. (2003) and the forthcoming book of Cressie and Wikle (2011). Examples of BHM applied to climate issues could be as follows. Berliner et al. (2000) studied long-lead predictions of Pacific Sea Surface Temperatures via Bayesian Dynamic Modeling. Cooley et al. (2005) implemented a BHM to infer glacial retreats in Bolivia using lichen growths as a proxy. Schliep et al. (2010) estimated extreme precipitation from regional climate models by combining BHM and extreme value theory. Tebaldi et al. (2010) characterized uncertainties of future climate change projections using BHM and Sahu et al. (2007) studied space-time ozone modeling for assessing trends. Haslett et al. (2006) investigated the problem of reconstructing prehistoric climates from lake sediment cores.

Schematically, uncertainty in BHM is spread over different layers, usually three. The base level, called the *data layer*, characterizes observations, e.g. tree ring density profiles in our case. The second level in the hierarchy, called the *process layer*, models latent processes that drive the growth of such rings, tree-to-tree and regional variations. In this second layer, one can start incorporating temporal processes, e.g. individual age effects and the hidden common environmental factor. The third level, called the *parameter layer*, consists of the information concerning prior parameters distributions that control the process layer.

In dendrochronology, Hooten and Wikle (2007) investigated with a BHM shifts in the spatio-temporal growth dynamics of shortleaf pine. These authors did not work with raw tree measurements but with chronology indices, i.e. already preprocessed and standardized data. They linked these chronologies with drought information like the Palmer Drought Severity Index. Concerning the standardization issue and BHM, Boreux et al. (2009) extracted an inter-annual high frequency signal from detrended tree ring series and consequently, smooth trends were also overlooked. Compared to these past studies, our goal is to add the flexibility of modeling non-parametric trends that can capture low frequency changes for the age effect and higher frequency variations for the hidden common environmental signal.

Denote $\mathbf{y}_j = (y_j(t_1), ..., y_j(t_n))^T$ the tree ring measurement vector produced by tree j over the period of interest $(t_1, \ldots, t_n)$. Equation (1) provides the foundation of our data layer that can be expressed with the common notations used by the Bayesian community as

$$\mathbf{y}_j|\mathbf{g}_j, \mathbf{f}, \beta_j, \sigma^2 \sim \mathbf{g}_j + \beta_j \mathbf{f} + \sigma^2 \mathcal{N}_n(\mathbf{0}_n, \mathbf{I}_n), \text{ with } j = 1, \ldots, p, \tag{2}$$

where the unknown $\mathbf{f} = (f(t_1), \ldots, f(t_n))^T$ represents the hidden common signal, see F_t in (1), the unknowns $\mathbf{g}_j = (g_j(t_1), \ldots, g_j(t_n))^T$ correspond to the individual age effect for each tree j, see G_t in (1), $\mathbf{0}_n = (0, \ldots, 0)^T$ and $\mathbf{I}_n$ denotes the identity matrix of size n. Measurement uncertainty is modeled as a zero mean Gaussian vector with covariance $\sigma^2\mathbf{I}_n$ and each tree record $[\mathbf{y}_j|\mathbf{g}_j, \mathbf{f}, \beta_j, \sigma^2]$ is supposed to be mutually independent of each other. In our application shown in Figure 4, the number of tree p is equal to fourteen and the time period is defined as $t_1 = 1903$ and $t_n = 1993$. The tree length variation displayed in Figure 1 implies that $\mathbf{g}_j$ starts or ends with a series of missing values for most trees.

To go one step further in our Bayesian hierarchy, we need to define the process layer, i.e. to set priors for $\mathbf{g}_j, \mathbf{f}, \beta_j$ and σ^2. In contrast to past dendrochronological studies that imposed a parametric form for $\mathbf{g}_j$ or $\mathbf{f}$ or both, we opt to describe both functions as semi-parametric splines viewed within a BHM framework.

Splines modeling was formulated by Reinsch (1967) and developed by many author (e.g., Eubank, 1999; Wand and Jones, 1995; Fan and Gijbels, 1996). Within the Bayesian framework, Kimeldorf and Wahba (1970) demonstrated that specific forms of spline smoothing correspond to Bayesian estimates under a class of improper Gaussian prior distributions on function spaces. For the classical non-parametric regression problem $\mathbf{y} = \mathbf{f} + \sigma^2\mathcal{N}(\mathbf{0}, \mathbf{I})$, Wahba (1978) proposed and studied a particular partially improper Gaussian prior for the trend $\mathbf{f}$

$$\mathbf{f}|\tau^2 \sim \mathcal{N}_n(0, \tau^2\mathbf{K}^-) \tag{3}$$

where $\tau^2 = \sigma^2/\lambda$ and $\lambda \geq 0$ is the smoother parameter of the classical penalized sum of squares criterion $\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int (f''(\mathbf{x}))^2 d\mathbf{x}$ that is minimized over all functions

$f(\mathbf{x})$ such that the integral exists. In (3), $\mathbf{K}^-$ refers to a generalized inverse of a matrix $\mathbf{K}$, with the understanding that an eigenvalue of zero for $\mathbf{K}$ gives an eigenvalue of $+\infty$ for $\mathbf{K}^-$. In the case of smoothing splines $\mathbf{K}$ is linked to the penalty $\int(f''(\mathbf{x}))^\mathbf{2}\mathbf{dx} = \mathbf{f^TKf}$. Hastie and Tibshirani (1990, 2000) showed that this prior covariance matrix $\mathbf{K}^-$ is equal to $\mathbf{B\Omega B}^T$ evaluated at the data. Let n_u the number of unique value of $\mathbf{x}$, the basis matrix $\mathbf{B}$ consist of the vector of $n_u + 2$ cubic B-splines basis functions $b(\mathbf{x})$ (de Boor, 1978) evaluated at the n_u sample values x_i and the penalty matrix $\mathbf{\Omega}$ has elements $\Omega_{ij} = \int b''_i(x)b''_j(x)dx$. Priors for the smoothing parameter or the variances σ^2 and τ^2 belongs to the parameter layer of the Bayesian hierarchy and they have to be fixed. Hastie and Tibshirani (1990, 2000) suggested to use proper inverse gamma priors for the variance components $\sigma^2 \sim \mathcal{IG}(a_\sigma, b_\sigma)$ and $\tau^2 \sim \mathcal{IG}(a, b)$.

Following the work of Wahba (1978) and Hastie and Tibshirani (1990, 2000), priors of our model defined by (2) can take their roots in (3) and consequently we assume the same type of priors for $\mathbf{g}_j$ and $\mathbf{f}$

$$\mathbf{f}|\tau_0^2 \sim \mathcal{N}_n(0, \tau_0^2\mathbf{K}^-) \text{ and } \mathbf{g}_j|\tau_j^2 \sim \mathcal{N}_n(0, \tau_j^2\mathbf{K}^-), \text{ for all } j = 1, \dots, p.$$

At this stage, our model is too versatile and associated with identifiability issues. For example, if all $\mathbf{g}_j$ are proportional to $\mathbf{f}$, it is impossible to distinguish $\mathbf{f}$ from $\mathbf{g}_j$. Additional constraints are needed and these have been be chosen according to basic tree ring characteristics. From Figure 2 we know that the individual age effect function $\mathbf{g}_j$ should be very smooth because individual tree growth is a rather slow and cumulative process. In contrast, we assume that the hidden signal shared by all trees $\mathbf{f}$ should capture environmental variabilities that correspond to rapid (yearly or decadal) changes. This means that the frequency range of $\mathbf{g}_j$ is assumed to be distinct from the one of $\mathbf{f}$. To illustrate this difference, Figure 5 displays simulations that mimic this phenomenon. The top and middle panels represent a simulated common signal $\mathbf{f}$ and simulated individual tree growth signals $\mathbf{g}_j$, respectively. In this idealized example, one can see that the functions $\mathbf{g}_j$ do not reproduce the rapid variations seen in $\mathbf{f}$. To test the resilience of our method, a slow positive trend was also included into $\mathbf{f}$ here and this adds difficulties to separate $\mathbf{f}$ from $\mathbf{g}_j$, see Section 3.1. The smoothness information can be translated into informative prior choice of the smoothness parameters τ_j^2 for $j = 0, \ldots, p$. For comparison and interpretation reasons, we substitute τ_j^2 by a parameter that lives on the interval $[0, 1]$

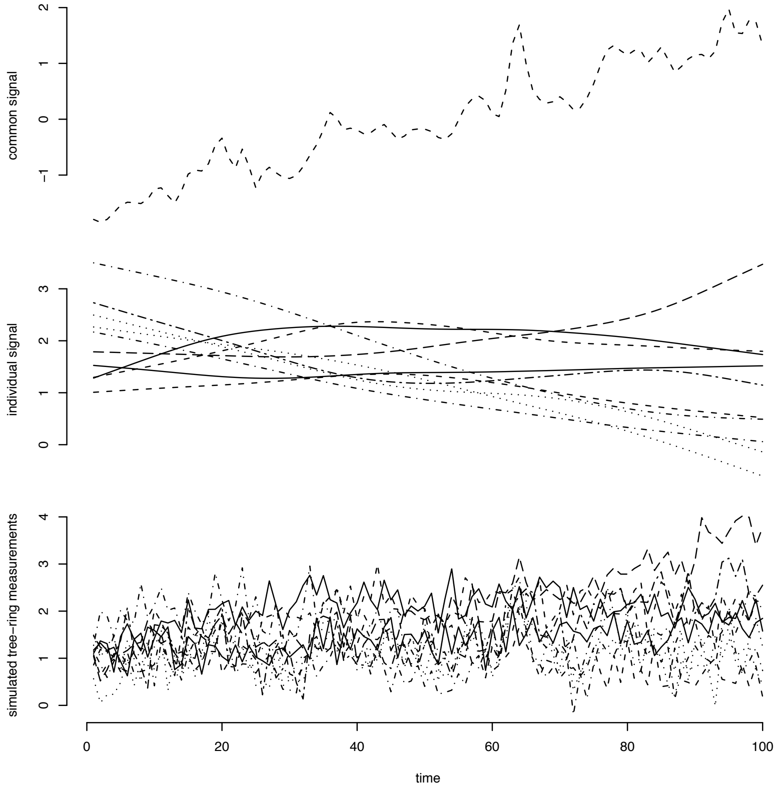

Figure 5: Simulations of tree ring measurements from the additive model (2). The top panel corresponds to the common signal $\mathbf{f}$, the second panel to individual growth tree effect signals $\mathbf{g}_j$ and the bottom panel to simulated tree ring series $\mathbf{y}_j$, respectively. Our objective is to find $\mathbf{f}$ and $\mathbf{g}_j$ from the $\mathbf{y}_j$'s.

$$\varphi_j = \frac{\sigma^2}{\tau_j^2 + \sigma^2}, \text{ for all } j = 0, \ldots, p.$$

If φ_j takes a value near one, then it means that the curve is very smooth. For the tree data analyzed in paper and after discussions with experts in dendrochronology, we set a strongly informative beta prior for $\varphi_j \sim \text{Beta}(100, 1)$ for $j = 1, \ldots, p$, see the dotted line in Figure 6. For the parameter describing the smoothness of $\mathbf{f}$, $\varphi_0 \sim \text{Beta}(2, 10)$ is also an informative but with a wider range. The choice insures a kind of orthogonality in the sense that the priors φ_0 and φ_j for $j \neq 0$ do not overlap, see Figure 6. The priors for φ_j may seem very strong but this corresponds to the clear information about the age effect frequency for the tree species studied in our example. To improve identifiability of the common signal, the function $\mathbf{f}$ is constrained to have a zero mean and unit variance. As in any dendroclimatology studies, the hidden signal $\mathbf{f}$ is dimensionless and should be interpreted as such. Concerning the parameter β_j that reflects the contribution of the common factor $\mathbf{f}$ to the growth of tree j, we assume that it is positive and it follows a truncated Normal with a rather non-informative variance of 10.

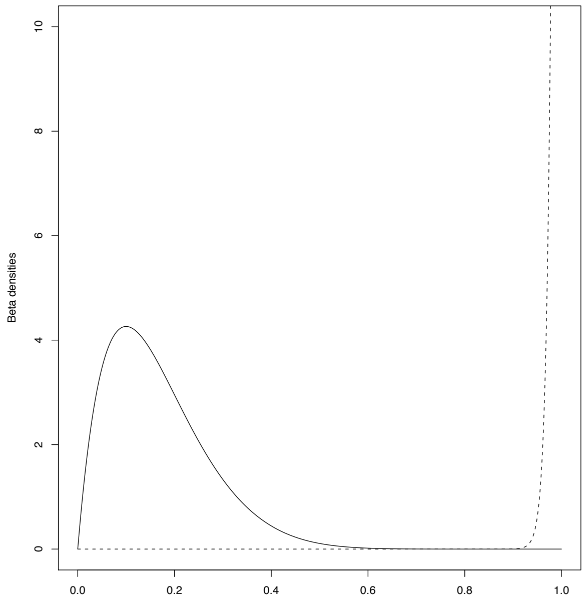

Figure 6: Informative Beta prior densities for the smoothness parameter $\varphi_0 \sim \text{Beta}(2, 10)$ (solid line) of the function $\mathbf{f}$ and for $\varphi_j \sim \text{Beta}(100, 1)$ (dotted line) of the function $\mathbf{g}_j$ for $j = 1, \ldots, p$. A value near one (near zero) corresponds to a smooth (jagged) curve.

To compute the posteriors of the latent vectors and model parameters, we use Gibbs sampler and Metropolis-Hasting algorithms. Explicitly posterior distribution for some functions can be derived (Hastie and Tibshirani, 1990, 2000)

$$\mathbf{f}|\boldsymbol{\beta}, \mathbf{G}, \lambda_0 \mathbf{Y}, \sigma^2 \sim \mathcal{N}_n(\mathbf{B}(\mathbf{B}^T\mathbf{R}\mathbf{B} + \lambda_0\boldsymbol{\Omega})^{-1}\mathbf{B}^T\mathbf{s}, \sigma^2\mathbf{B}(\mathbf{B}^T\mathbf{R}\mathbf{B} + \lambda_0\boldsymbol{\Omega})^{-1}\mathbf{B})$$

with

$$\mathbf{s} = \sum_{j=1}^{p} \beta_j(\mathbf{y}_j - \mathbf{g}_j),\ \lambda_0 = \varphi_0/(1-\varphi_0),\ \mathbf{R} = \sum_{j=1}^{p} \beta_j^2 \mathbf{I}$$

and

$$\mathbf{g}_j|\beta_j, \mathbf{f}, \lambda_j \mathbf{y}_j, \sigma^2 \sim \mathcal{N}_n(\mathbf{B}(\mathbf{B}^T\mathbf{B} + \lambda_j\boldsymbol{\Omega})^{-1}\mathbf{B}^T\mathbf{d}, \sigma^2\mathbf{B}(\mathbf{B}^T\mathbf{B} + \lambda_j\boldsymbol{\Omega})^{-1}\mathbf{B})$$

with $\mathbf{d} = \mathbf{y}_j - \beta_j\mathbf{f}$ and $\lambda_j = \varphi_j/(1-\varphi_j)$. It is also possible to show that β_j follows a truncated normal posterior distribution and σ^2 have an inverse gamma posterior distribution. These parameters are estimated with Gibbs sampler. The parameters φ_0 and φ_j don't have standard posterior distributions so we use Metropolis-Hasting algorithm to estimate them. The Bayesian inference was carried out with the open source R statistical software.

3 Data analysis

3.1 Simulations results

From the $\mathbf{y}_j$'s displayed in the bottom panel of Figure 5, the posterior probability density functions (pdf) of our model parameters are obtained. The dotted lines in Figure 7 indicates the posterior median of $\mathbf{g}_1$, $\mathbf{g}_2$ and $\mathbf{g}_3$ (the same type of graphs can be obtained for $\mathbf{g}_j$ with $j \geq 4$). Compared to the original $\mathbf{g}_1$, $\mathbf{g}_2$ and $\mathbf{g}_3$ (solid lines), the 95% credibility intervals (gray area) appear to capture reasonably well the "true" age effect variations. In particular the smoothness of the estimated $\mathbf{g}_j$'s is comparable to the original one. This is not surprising and it is mainly due to the strong informative prior put on the parameter φ_j, i.e. a prior mode around one. The right panels confirm this by showing the posterior pdf of φ_j centered around the value 0.995. Minor edge effects seem to be present and the 95% credibility intervals (gray area) reflects this by getting wider around both edges. More interestingly, the original shape difference between $\mathbf{g}_1$ and $\mathbf{g}_3$, the latter first increases and then plateaus in time and the former does the opposite, indicates that such variations among individual hidden smooth profiles can be found with our approach. From a dendrochronological side, this distinguishes our method from the RCS one that postulates an unique biological age trend for all trees (see Section 1).

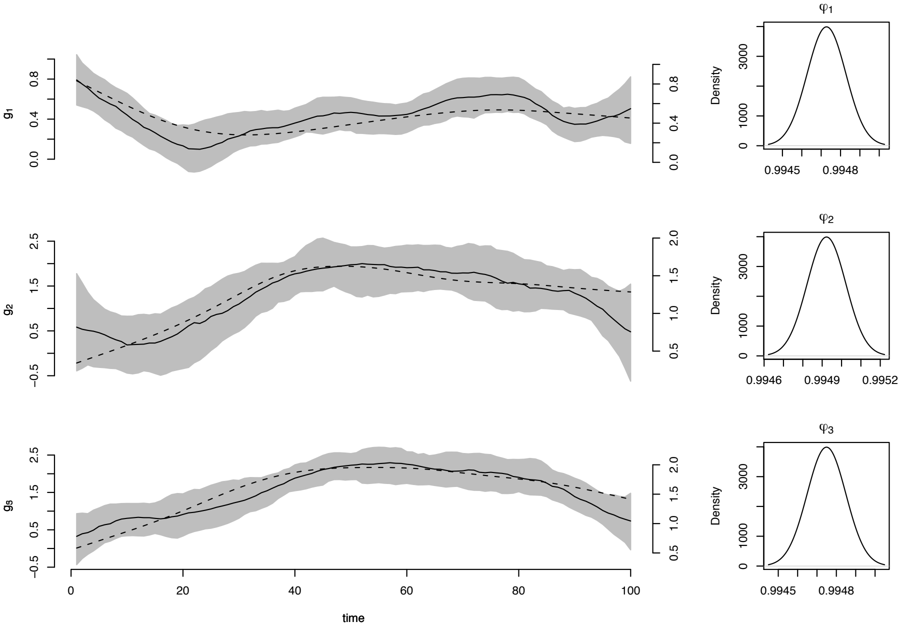

Figure 7: Posterior information about the tree age effect $\mathbf{g}_j$ for $j = 1, 2, 3$ obtained from the simulated tree series shown in the bottom panel of Figure 5. The solid and dotted lines in the left panels correspond to the true $\mathbf{g}_j$ and the estimated posterior median, respectively. The gray gray area represents the 95% credibility intervals. The right panels display the posterior pdf of the smoothness parameter the posterior median and 95% credibility intervals of φ_j for $j = 1, 2, 3$.

Concerning the main element of our modeling, the temporal evolution of the hidden signal shared by all trees $\mathbf{f}$ is represented by a solid line in the right panel of Figure 8. The posterior median (dotted line) and the 95% credibility intervals (gray area) adequately follow the behavior of the true $\mathbf{f}$. As expected from the choice of our φ_0 prior, quicker variations than the ones observed in the posteriors pdf of $\mathbf{g}_j$'s can be seen in the posterior of $\mathbf{f}$. The right panel of Figure 8 corroborates this point, the posterior pdf of the smoothness parameter φ_0 takes its values around 0.15. It is worthwhile to notice that the increasing slow trend in $\mathbf{f}$ is also captured by its posterior. This implies that, although the prior and posterior of φ_0 is dedicated to a high frequency range, smooth trends in $\mathbf{f}$ can be detected via our approach. This is due to the combination of two items: the variability among the $\mathbf{g}_j$'s and the number of trees. If all $\mathbf{g}_j$ had the same shape, say slowly increasing, then it would be impossible to capture an increasing trend in $\mathbf{f}$. This variability among $\mathbf{g}_j$'s should increase with the number trees, especially if the trees have different ages and therefore span different age related curves. May be counterintuitively, this means that having a wide range of tree age effect profiles could be an advantage to detect smooth trend in $\mathbf{f}$. But only if the statistical extraction is truly multivariate and performs with well-chosen priors for the smoothness parameters of the $\mathbf{g}_j$'s.

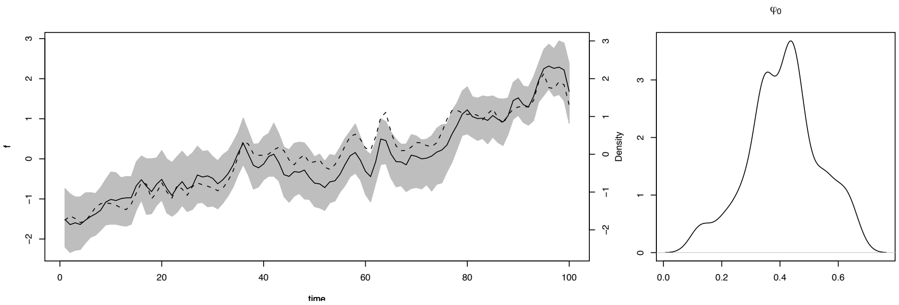

Figure 8: Left panel: posterior median and 95% credibility intervals of the common signal shared by all trees $\mathbf{f}$ obtained from the simulated tree series shown in the bottom panel of Figure 5. The solid corresponds to the true $\mathbf{f}$. Right panel: posterior pdf of the smoothness parameter φ_0.

Different sensitivity analysis concerning the influence of the noise level and the number of tree on the inference quality were also performed and are available upon request. In a nutshell, the noise level σ^2 can influence the analysis if the noise ratio becomes too large. Concerning the number of trees, around 10 trees in our simulations were necessary to derive reasonable results like in figures 8 and 7. However this remark about a minimal number of trees is only valid within the framework of our simulations and it should not be directly transposed to real data because the shapes of $\mathbf{f}$ and $\mathbf{g}_j$ and the variance σ^2 strongly depend on the tree species and the site characteristic.

3.2 Analysis of 14 tree ring density series of *Pinus halepensis Mill*

Our model and inference scheme have been applied to the fourteen tree density series shown in Figure 4. The posterior median (solid line) and their associated 95% credibility intervals (gray area) of the three individual age trends $\mathbf{g}_1$, $\mathbf{g}_2$ and $\mathbf{g}_3$ are shown in the right panels of Figure 9. As in our simulation study, the curves are smooth by construction (prior choice of the smoothness parameter) and display a variety of shape (increasing or decreasing depending on the period and the tree).

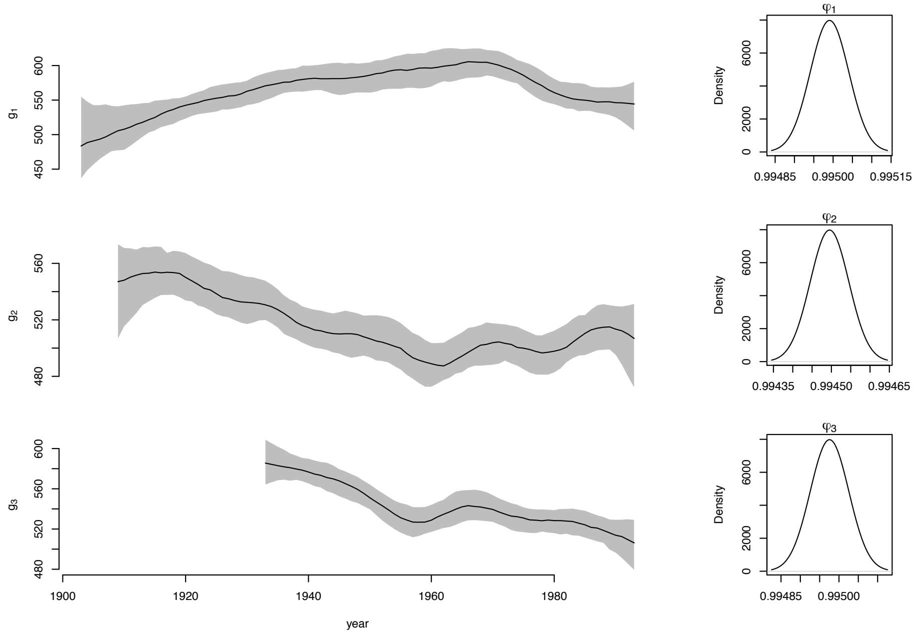

Figure 9: Left panels: posteriors of the three individual age effect trends $\mathbf{g}_1$, $\mathbf{g}_2$ and $\mathbf{g}_3$ obtained from our analysis of the fourteen tree density series shown in Figure 4. Black lines correspond to posterior medians and gray areas to 95% credibility intervals. Right panels: posterior pdfs of the smoothness parameters φ_1, φ_2 and φ_3.

To put our approach into perspective with respect to the RCS method, Figure 10 compares posterior median of individual age effect profiles $\mathbf{g}_j$ that have been aligned according to their biological age (not their chronological age) with the classical global biological trend obtained by averaging ring widths in function of their biological age (gray line). The line thickness is proportional to the posterior median coefficient β_j. Although a majority of curves follow a similar shape (an early increase, then one (or two) peak followed by a decrease), this figure emphasizes the variability among age effect profiles. In particular, the peak of the RCS biological curve occurs after about 40-50 years. In terms of $\mathbf{g}_j$, this peak date (when available) varies greatly from one tree to another. This tends to indicate that the added flexibility of our modeling approach allows to improve individual age-related growth variability. A strong message from Figure 10 resides in the large variability among the different age effect shapes. Each tree has its own trend and associated uncertainty. And having this information could help dendrochronologists to interpret local tree behaviors.

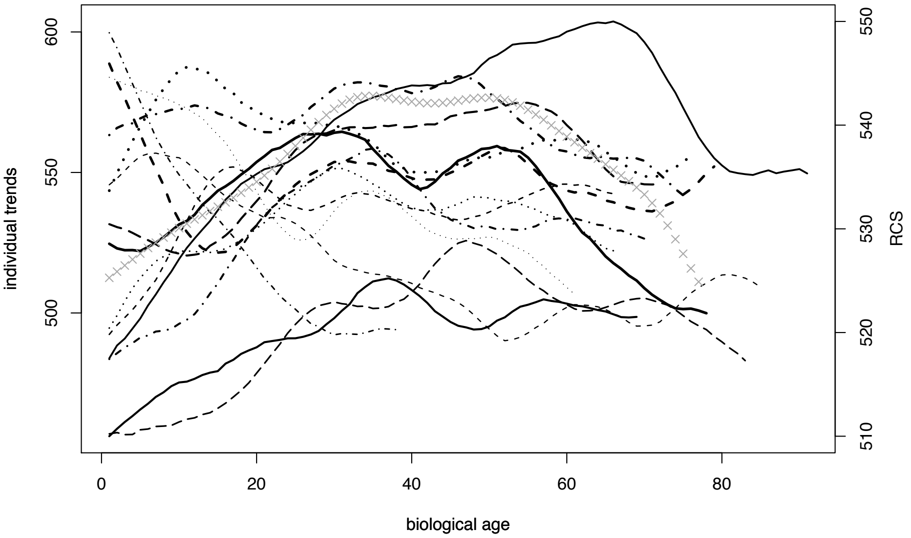

Figure 10: Posterior median of individual age effect profiles $\mathbf{g}_j$ that have been aligned according to their biological age (not their chronological age). The line thickness is proportional to the posterior median coefficient β_j. The gray line represents the classical global biological trend obtained by averaging ring widths in function of their biological age.

To better understand the limits of our approach, we had left two other sites called “Rognac" and “Gardanne" out of our analysis. These two places have similar environmental and climatic characteristics than the original site “Les Pennes Mirabeau", see Figure 3 and the same species of tree has been sampled. The same variable, tree ring density series, has been modeled independently for each site. Figure 11 compares the extracted signal $\mathbf{f}$ for the three sites. Overall there is a reasonable agreement among the three posterior medians for $\mathbf{f}$. The 95% credibility interval computed from the fourteen tree density series of “Les Pennes Mirabeau" seems to contain most of the data points from the two other curves. None of the curves appears to have a centennial trend. Prior to 1920, the 95% credibility interval becomes wider because the number of tree decreases around this epoch, see Figure 1 and minor edge effects can also occur.

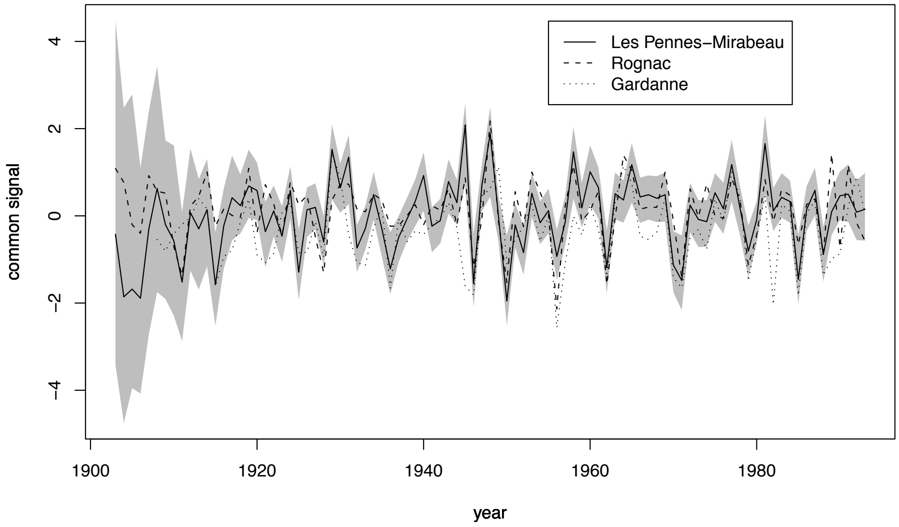

Figure 11: Posterior median of the common signal $\mathbf{f}$ obtained from trees measured at the site of “Les Pennes Mirabeau" (solid line), the site of “Rognac" (dashed line) and the site of “Gardanne" (dotted line). The three sites belong to the same climatological region and have the same tree species. The 95% credibility interval is computed from the fourteen tree density series shown in Figure 4.

To conclude this analysis, we briefly investigate potential links between our extracted signals and climatic variables. Inspired by the work of Nicault et al. (2001), we focus on one explanatory variable: the sum of Summer daily precipitation recorded over the period $1947 - 1993$ in Marseille, (latitude = +43:18:18, longitude = +05:23:48 and altitude = 75) from the European Climate Assessment & Dataset (ECA&D) (`http://eca.knmi.nl/`). One goal of dendroclimatology is to reconstruct climatic variables from tree rings. To perform this task, we calibrate our relationships on the period $1961 - 1993$ and we leave out the period $1947 - 1960$ in order to assess the quality of our predictions. Basic linear modeling indicates a clear link between our extracted $\mathbf{f}$ and the logarithm of observed rainfall (a correlation of 0.63). To a lesser degree, a linear relationship between $\mathbf{f}$ and and Spring daily temperatures seems also plausible, via a correlation of 0.51, but this won't be explored here. To visualize if it is possible to bring out relevant rainfall information from the signal $\mathbf{f}$ over the validation period $1947 - 1960$, Figure 12 compares the reconstructed log-precipitation (black line) obtained by inverting the linear relationship calibrated over $1961 - 1993$ with the measured log-rainfall (grey line), both shares a correlation coefficient of 0.54 over the validation period. On this graph, we have also added two other re-

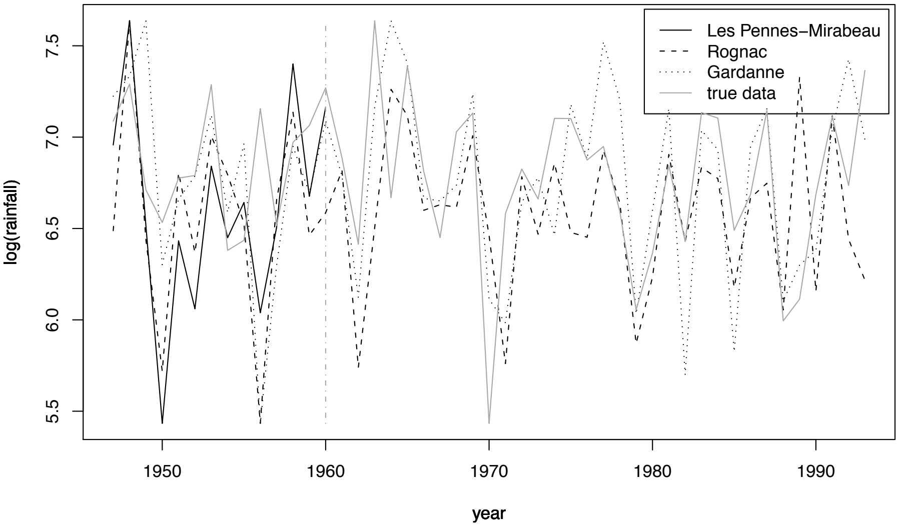

Figure 12: Rainfall reconstruction. The grey line represents the logarithm of observed total Summer precipitation recorded during 1947-1993 in Marseille, source ECA&D (`http://eca.knmi.nl/`). During the period 1961-1993, a linear estimation between log(rainfall) and the signal $\mathbf{f}$ was implemented for the site of "Les Pennes-Meribeau". For this site, the reconstruction, i.e. inverting a linear relationship, was done for the early time period 1947-1960 (solid black line). This relationship calibrated for the site “Les Pennes-Meribeau" was also applied to two $\mathbf{f}$s from two other sites “Rognac" (dotted line) and “Gardanne" (dashed line).

constructed log-precipitation time series computed from the $\mathbf{f}$s derived from our other two sites “Rognac" and “Gardanne". These sites were not used during the calibration period and Figure 12 displays reconstructed variations over the entire period $1947 - 1993$ for which rainfall data are available and can be compared to. Visual inspections and correlation coefficients of 0.30 (Rognac) and 0.53 (Gardanne) indicate that the reconstructed log(rainfall) for Gardanne reproduced more efficiently the observed log(rainfall) time series than the one derived from the Rognac site. Overall this short reconstruction exercise reveals that our extraction method applied at *Pinus halepensis Mill* tree ring density series recorded at three different sites produces hidden common signals correlated with environmental factors like precipitation.

4 Discussions

From a statistical perspective, the extraction of a common signal shared by all trees remains difficult because ring growth variations result from complex interactions between climatic and non-climatic factors. The common signal could be viewed as a representation of the regional environmental pressure affecting trees over a studied area. To make a very limited number of statistical assumptions, we opted for combining two semi-parametric spline models, one for the common hidden signal and one for individual age effects, within a multivariate Bayesian hierarchical model. Identifiability issues imposed to have a strong informative prior on the individual age effect frequency. This was not too stringent because past dendrochronological studies provide such information. The advantage of our approach is that the variability among individual age effect components is less constrained than with the classical RCS technique that forces a “one size fits all" age effect curve for all trees. One consequence of our age effect modeling could be that the common signal is better decoupled from the later. We have also chosen to dissociate the extraction problem from the selection problem. In other words, how to extract a common signal is viewed differently from the question, how to explain such an extracted signal with climatic or non-climatic covariates. This strategy may reduce the chances to make false connections between tree ring information and supposed explanatory variables.

Our analysis of *Pinus halepensis Mill* tree ring density series gave encouraging results in terms of extraction and reconstructions. Similar extracted signals were found at three different sites and this seems to confirm the regional environmental factor of the extracted signal, most likely linked with Summer precipitation. Finally we are convinced that it would be of interest to incorporate a spatial component in our model because the common signal should correspond to a specific spatial environmental scale. But this was not possible with the three sites we had for this analysis. Integrating a spatial component in a semi-parametric context is non trivial for many reasons. Climate variables like precipitation are associated with large and small spatial patterns while trees may record local variations at a much finer spatial scale. In addition weather stations and tree ring measurements are not placed at the same locations. Those spatial discrepancies translate into complex problems in terms of spatial sampling and change of support within a semi-parametric BHM framework. It would be of interest to followed some of the ideas developed in Hooten and Wikle (2007) and Li et al. (2010) to integrate a spatial dimension into our analysis.

Finally an interesting new strategy in climate reconstructions consists in producing an ensemble of simulated proxy records from forward-process based models (Hughes et al., 2010; Hughes and Ammann, 2009; Guiot et al., 2009). In this context, the capabilities of our semi-parametric BHM at simulating synthetic tree-rings densities could be explored. A possibility could be to constrain the function $\mathbf{f}$ with environmental factors. A main difficulty resides in the strong heterogeneity among the individual age effect, see Figure 10, and dynamical tree ring growth models could tested from the extracted signal represented in Figure 10.

Acknowledgements

Part of this work has been supported by the MAIF foundation, the EU-FP7 ACQWA Project (www.acqwa.ch) under Contract Nr 212250, and by the ANR-MOPERA project . The authors would also like to credit the contributors of the R project.

References

Barefoot, A., Woodhouse, L., Hafley, W., and Wilson, E. (1974). Developing a dendrochronology for winchester england. *Journal of the Institute of Wood Science*, 6:34–40.

Berliner, L., Wikle, C., and Cressie, N. (2000). Long-lead prediction of pacific ssts via bayesian dynamic modeling. *J. Climate*, 13:3953–3968.

Boreux, J.-J., Naveau, P., Guin, O., Perreault, L., and Bernier, J. (2009). Extracting a common high frequency signal from northern quebec black spruce tree-rings with a bayesian hierarchical model. *Climate of the Past*, 5(4):607–613.

Buckley, B. (2009). *Encyclopedia of Paleoclimatology and Ancient Environments. Encyclopedia of Earth Sciences Series.*, chapter Dating, dendrochronology. Dordrecht, Netherlands: Springer.

Committee on Surface Temperature Reconstructions for the Last 2000 Years (2006). *Surface Temperature Reconstructions for the Last 2000 Years*. National Research Council.

Cook, E. (1990). *Methods of Dendrochronology: Applications in the Environmental Sciences.*, chapter A conceptual linear aggregate model for tree rings, pages 98–104. Kluwer Academic Publ., Dordrecht.

Cook, E. and Kairikukstis, L. (1990). *Methods of dendrochronology : applications in the environmental sciences / edited by E.R. Cook and L.A. Kairiukstis*. Kluwer Academic Publishers, Dordrecht, Netherlands, Boston.

Cook, E. and Peters, K. (1981). The smoothing spline : a new approach to standardizing forest interior tree-ring width series for dendroclimatic studies. *Tree-ring Bulletin*, 41:45–53.

Cooley, D., Naveau, P., and Jomelli, V. (2005). A bayesian hierarchical extreme value model for lichenometry. *Environmetrics*, 16:1–20.

Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. Wiley.

de Boor, C. (1978). *A practical Guide to Splines*. Applied Mathematical Sciences.

Douglass, A. (1920). Evidence of climatic effects in the annual rings of trees. *Ecology*, 1:24–32.

Douglass, A. (1936). Climatic cycles and tree-growth. *Carnergie Institution of Washington publication*, 289(3).

Esper, J., Cook, E. R., and Schweingruber, F. H. (2002). Low frequency signals in long tree-ring chronologies for reconstructing past temperature variability. *Science*, 295.

Eubank, R. (1999). *Nonparametric regression and spline smoothing*. Marcel Dekker.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall.

Fritts, H., Mosimann, J., and Bottorff, C. (1969). A revised computer program for standardizing tree-ring series. *Tree-Ring Bulletin*, 29:15–20.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis*. Chapman and Hall, 2nd edition.

Gornitz, V., editor (2009). *Encyclopedia of Paleoclimatology and Ancient Environments. Encyclopedia of Earth Sciences Series.* Dordrecht, Netherlands: Springer.

Guiot, J. (1987). *Methods of dendrochronology - 1*, chapter Standardization and selection of the chronologies by the ARMA analysis. International Institute for Applied Systems Analysis, Laxenburg, Austria and Polish Academy of Sciences-System Research Institute, Warsaw, Poland.

Guiot, J., Wu, H. B., Garreta, V., Hatté, C., and Magny, M. (2009). A few prospective ideas on climate reconstruction: from a statistical single proxy approach towards a multi-proxy and dynamical approach. *Climate of the Past*, 5(4):571–583.

Haslett, J., Salter-Townshend, M., Wilson, S. P., Bhattacharya, S., Whiley, M., Allen, J. R. M., Huntley, B., and Mitchell, F. J. G. (2006). Bayesian palaeoclimate reconstruction. *J. R. Statist. Soc. A*, 169, Part 3, pp.(3):1–36.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (2000). Bayesian backfitting. *Statistical Science*, 15(3):196–223.

Hooten, M. and Wikle, C. (2007). Shifts in the spatio-temporal growth dynamics of shortleaf pine. *Environmental and Ecological Statistics*, 14(3):207–227.

Hughes, M. K. and Ammann, C. M. (2009). The future of the past—an earth system framework for high resolution paleoclimatology: editorial essay. *Climatic Change*, 94:247–259.

Hughes, M. K., Guiot, J., and Ammann, C. M. (2010). An emerging paradigm: Process-based climate reconstructions. *PAGES news*, 18:87–89.

Kimeldorf, G. and Wahba, G. (1970). A correspondance between bayesian estimation of stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41:495–502.

Li, B., Nychka, D. W., and Ammann, C. M. (2010). The value of multiproxy reconstruction of past climate. *Journal of the American Statistical Association*, 105(491):883–895.

Mann, M., Bradley, R., and Hughes, M. (1999). Northern hemisphere temperatures during the past millennium : inferences, uncertainties and limitations. *Geophysical Research Letters*, 2:759–762.

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., and Rutherford, S. (2008). Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proc. Natl. Acad. Sci.*, 105:13252–13257.

McShane, B. B. and Wyner, A. J. (2010). A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable? *Annals of Applied Statistics*, In press.

Melvin, T. and Briffa, K. R. (2008). A "signal-free" approach to dendroclimatic standardisation. *Dendrochronologia*, 26:71–86.

Nicault, A., Guiot, J., Edouard, J., and Brewer, S. (2010). Preserving long-term fluctuations in standardisation of tree-ring series by the adaptative regional growth curve (argc). *Dendrochronologia*, 28(1):1 – 12.

Nicault, A., Rathgeber, C., Tessier, L., and Thomas, A. (2001). Observations sur la mise en place du cerne chez le pin d'alep (pinus halepensis mill.) : confrontation entre les mesures de croissance radiale, de densité et les facteurs climatiques. *Annals of forest science*, 58:769–784.

Reinsch, C. (1967). Smoothing by spline functions. *Numer. Math.*, 10:177–138.

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2007). High-resolution space-time ozone modeling for assessing trends. *Journal of the American Statistical Association*, pages 1–14.

Schliep, E., Cooley, D., Sain, S., and Hoeting, J. (2010). A comparison study of extreme precipitation from six different regional climate models via spatial hierarchical modeling. *Extremes*, 13:219–239.

Smith, R. (2010). Understanding sensitivities in paleoclimatic reconstructions. Technical report, Technical report.

Stahle, D. W. (1999). Useful strategies for the development oftropical tree-ring chronologies. *IAWA Journal*, 20(3):249–253.

Tebaldi, C., Smith, R. L., and Sanso, B. (2010). *BAYESIAN STATISTICS 9*, chapter Characterizing Uncertainty of Future Climate Change Projections Using Hierarchical Bayesian Models. Oxford University Press,.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society*, 40(3):364–372.

Wand, M. and Jones, M. (1995). *Kernel smoothing*. Chapman and Hall.

4.2 Bayesian variables selection for Generalized Additive Models applied to climatic reconstructions

Cet article est en cours d'achèvement et devrait être soumis courant janvier.

Résumé : Afin de comprendre si le réchauffement climatique actuel est plus important que la variabilité climatique naturelle, il est nécessaire d'avoir de longues séries de températures ou de précipitations. Seulement, les mesures directes manquent, en particulier pour les périodes les plus anciennes, et il est nécessaire d'utiliser des proxies afin de reconstruire des chronologies passées. Ce papier s'intéresse tout particulièrement à l'un des proxy les plus utilisé : les mesures de cernes d'arbres. On cherche donc à identifier les relations existant entre le climat et les cernes d'arbres, c'est-à-dire à choisir des variables climatiques expliquant au mieux la croissance des arbres. Est-ce que les différentes mesures sur les cernes d'arbres sont corrélées à la moyenne journalière des précipitations durant les mois d'été ? Aux températures saisonnières ? Etc... Le nombre de possibilité et sans fin et dépend de l'espèce des arbres et de leur région géographique. D'un point de vu statistique, ce problème peut être vu comme un problème de sélection de variables. En modélisant les relations cernes-climat à l'aide d'un modèle bayésien additif généralisé, nous proposons dans ce papier une méthode de sélection de variables afin de déterminer quels facteurs climatiques influencent le plus la croissance des arbres et comment. Notre méthode est testée sur des données simulées et appliquée à des mesures de densité de cernes d'arbres (*Pinus halepensis Mill.*) enregistrées sur la côte Méditerranéenne française.

Bayesian variables selection for Generalized Additive Models applied to climatic reconstructions

Ophélie Guin[1], **James Merleau**[2], **Philippe Naveau**[1]

[1]Laboratoire de Sciences du Climat et de l'Environnement, IPSL-CNRS, France
[2]Institut de Recherche d'Hydro-Québec (IREQ), Montréal, Canada

November 29, 2010

Abstract

1 Climatic reconstruction and tree-rings

In order to understand past and recent climate changes it is necessary to construct long temperatures and precipitation series. But, direct measurements of such climatological variables are missing whenever the instrumental record length is shorter than the period of interest. Proxies, i.e. indirect measurements, offer the material to reconstruct past chronologies in such situations. So, one key to understand climate it is to derive, study and apply efficient statistical procedures to identify link between proxies information and temperatures or precipitation. One of the the most well-known and common climate proxy is tree-ring measurements. Since the work of Douglass (1920, 1936), there has been an active and extensive research activity dedicated to the field of dendrochronology (dendron = tree and chronos = time) that study tree-ring to analyze temporal and spatial patterns of processes in the physical and cultural sciences. Journals like Tree-Ring Research (formerly Tree- Ring Bulletin) and Dendrochronologia, numerous books (e.g. Cook and Kairikukstis 1990) and thousands of articles show the vitality and the importance of tree-rings in many fields, e.g. forest ecology, climatology, archaeology and botany. To illustrate the importance of dendrochronology in climatology, we recall the importance and heavily commented papers of Mann et al. (1999) and Esper et al. (2002) that used tree-ring data to reconstruct Northern Hemispheric annual temperatures for the last millennium. One heated point of discussion in the global climate warming debate was the statistical analysis of tree-ring data in these two papers (Committee on Surface Temperature Reconstructions for the Last 2000 Years, 2006; Mann et al. 2008). The recent paper by ? provides a statistical blueprint to combine different proxies to reconstruct past temperatures, but it does not focus on the statistical tree-ring analysis. This will the main object of our paper.

Indeed one of the statistical difficulty in dendroclimatology concerns the delicate choice of the explanatory variables and their time scales. Should the tree-ring growth be explained by the average of daily precipitation over the summer months, the largest number of consecutive days without rain during one year, a function of seasonal temperatures or any other choice? The number of possibilities is endless and depends on the tree specie and the region of interest at hand. For example it is well documented that precipitation in Arizona have a strong influence on ... Hence the dendrochronologue expertise is invaluable to preselect possible meaningful explanatory variables and this sometimes allows the statistician to view a tree-ring reconstruction problem as a variable selection problem within an inverse regression procedure.

The most common statistical models used by dendrochronologists are called "correlation functions" and "response functions" (Blasing et al., 1984; Fritts et al., 1971). The term "function" indicates a sequence of coefficients computed between the tree-ring chronology and the monthly climatic variables, which are ordered in time from the previous-year growing season to the current-year one. In correlation functions the coefficients are univariate estimates of Pearson's product moment correlation (e.g. Morrison 1983), while in response function the coefficients are multivariate estimates from a principal component regression model (Briffa and E.R., 1990; Morzukh and Ruark, 1991).

Interpretation of correlation and response functions is favored by an accurate assessment of statistical significance, so that appropriate ecophysiological hypotheses (e.g. Biondi 1993, Biondi 1997) and paleoclimatic reconstructions (e.g. Biondi 2000, Biondi 1999) can be generated. In response functions, normal significance levels of coefficients are misleading because error estimates are underestimated (Cropper, 1985; Morzukh and Ruark, 1991), hence some coefficients can erroneously pass the significance test. This usually causes a grater number of significance coefficients in response functions than in correlation functions (e.g. Villalba et al. 1994). As a solution, bootstrapped error estimates can be used to obtain more accurate results (Efron, 1979; Efron and Tibshirani, 1986; Guiot, 1990, 1991). Correlation functions can also be incorrectly tested for significance, as explained by Biondi (1997), and it is therefore desirable to compute bootstrapped confidence intervals for correlation functions as well.

An implicit assumption of these statistical techniques is that climate-tree growth relationships can be represented by linear models. This assumption holds well in areas such as the American Southwest, where trees are strongly limited by cool season precipitation, so that linear regression can be used successfully to reconstruct climate. But, in areas where multiple climatic parameters control tree growth, and/or where climate response varies with species and site characteristics, non-linear methods appear more suitable to identify relationships between tree growth and climate. For example, experimental data (Fritts, 1976; Kramer and Kozlowski, 1979; Gates, 1980; Evans et al., 2006) suggest that the dependance of the growth rate function on temperature may be subdivided into three segments : rising growth rate with increasing temperatures below growth-optimal temperature range, relatively constant rates within an optimal range of temperatures and decreasing growth rates above that temperatures range (Figure 1). To circumvent this issue, some methods were proposed like response surface (Graumlich, 1993) or Neural network (Woodhouse, 1999), but they have not been successfully tested.

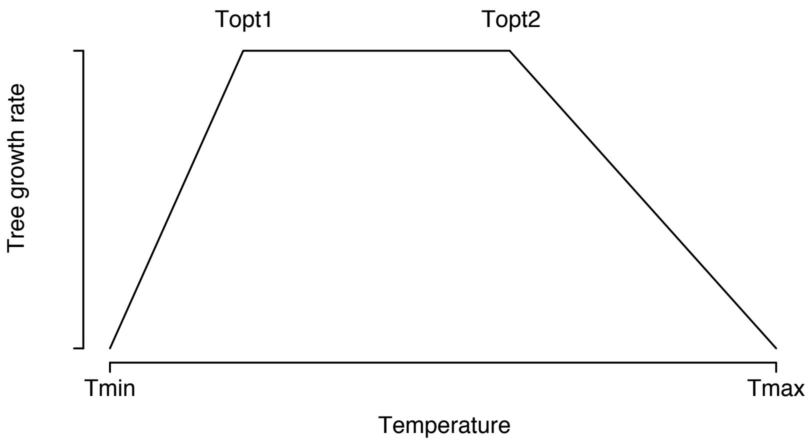

Figure 1: Tree growth rates function on temperature. The first segment corresponds to the period of rising growth rate with increasing temperatures below growth-optimal temperature range (Topt1), the second segment to a relatively constant rates within an optimal range of temperatures [Topt1, Topt2] and the third segment to the period of decreasing growth rates above that temperatures range.

In the light of this reflections, we suppose tree-ring growths are not a simple linear regression of climatic variables but linear regression of climatic data functions. In this paper, we propose to model relation between tree-rings and climatic factors with a Generalized Additive Model (GAM). So, from a statistical point of view, we must solve variable selection problem for GAM. *[Bayesian choice].*

2 Bayesian Generalized Additive Models

First, we consider a simple one component smoothing problem with data (x_1, y_1), $(x_2, y_2), ..., (x_n, y_n)$. Here y_i are the response values (a tree-ring in our context) and x_i is an input or predictor (temperature, precipitation, etc...). We consider the following model

$$y_i = f(x_i) + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

A smoothing spline is a popular model for representing $f(\mathbf{x})$, and can be derived as the minimizer of the following penalized sum of squares criterion

$$J(f) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int (f''(\boldsymbol{x}))^2 d\boldsymbol{x} \tag{1}$$

over all functions $f(\boldsymbol{x})$ such that the integral exists. The constant $\lambda \geq 0$ is a smoothing parameter, with larger values resulting in smoother curves since the curvature is then more penalized. For a given value of λ, the solution of this minimization problem, $\hat{\boldsymbol{f}}$, is a natural cubic spline, with knots at each of the unique values of x_i.

Another characterization of $\hat{\boldsymbol{f}}$ can be obtained through a Bayesian formulation of the problem. With the distributional hypothesis given above for the observations and with the following prior distribution on $\boldsymbol{f}$ (Wahba, 1978; Hastie and Tibshirani, 1990), the Bayesian model can be written as

$$\boldsymbol{y}|\sigma^2 \sim \mathcal{N}_n(\boldsymbol{f}, \sigma^2\mathbf{I}) \text{ and } \mathbf{f}|\sigma^2, \lambda \sim \mathcal{N}_n\left(\mathbf{0}_n, \frac{\sigma^2}{\lambda}\mathbf{K}^-\right)$$

where this last distribution is partially improper (see section 3.6 of Hastie and Tibshirani (1990) for a full discussion). The solution $\hat{\mathbf{f}}$ is then given by the expectation of the posterior distribution of $\boldsymbol{f}$.

The notation $\mathbf{K}^-$ refers to a generalized inverse of a matrix $\mathbf{K}$, with the understanding that an eigenvalue of zero for $\mathbf{K}$ gives an eigenvalue of $+\infty$ for $\mathbf{K}^-$. In the case of smoothing splines $\mathbf{K}$ computes the penalty in (1), $\int(f''(\boldsymbol{x}))^2 d\boldsymbol{x} = \boldsymbol{f}^T\mathbf{K}\boldsymbol{f}$.

Hastie and Tibshirani (1990, 2000) show that this prior covariance matrix $\mathbf{K}^-$ is equal to $\mathbf{B}\boldsymbol{\Omega}\mathbf{B}^T$ evaluated at the data. Let n_u the number of unique value of $\boldsymbol{x}$, the basis matrix $\mathbf{B}$ consist of the vector of $M = n_u + 2$ cubic B-splines basis functions $b(\boldsymbol{x})$ (de Boor, 1978) evaluated at the n_u sample values x_i and the penalty matrix $\boldsymbol{\Omega}$ has elements

$$\Omega_{ij} = \int b_i''(t)b_j''(t)dt.$$

In a Bayesian statistical model, prior distributions also need to be specified for the variance parameter σ^2 and the smoothing parameter λ. This aspect of the problem is discussed in section ??.

Now, we consider a Generalized Additive Model (GAM). Our data consists of n observations $(y_1, ..., y_n)$ and p explanatory variables contained in a matrix $\boldsymbol{X} = \{x_{ij}\}$, with $i = 1, \ldots, n$ and $j = 1, \ldots, p$. We note

$$\boldsymbol{x}_{i\cdot} = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})',$$
$$\boldsymbol{x}_{\cdot j} = (x_{1j}, \ldots, x_{ij}, \ldots, x_{nj})',$$

where $\boldsymbol{x}_{i\cdot}$ is a column vector $p \times 1$ for the case number i and $\boldsymbol{x}_{\cdot j}$ is a column vector $n \times 1$ for explanatory variable j. We then have the smooth function, in vector notation, for the jth explanatory variable

$$\boldsymbol{f}_j = f_j(\boldsymbol{x}_{\cdot j}) = (f_j(x_{1j}), \ldots, f_j(x_{ij}), \ldots, f_j(x_{nj}))',$$

a column vector of dimension $n \times 1$. We get the following model

$$\boldsymbol{y} = \sum_{j=1}^{p} \boldsymbol{f}_j + \boldsymbol{\epsilon}, \text{ with } \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n) \qquad (2)$$

This model, which can also be expressed as a Bayesian model and is the one which will be studied in this paper :

$$\boldsymbol{y}|\mathbf{A},\boldsymbol{\theta},\sigma^2 \sim \mathcal{N}_n\left(\mathbf{A}\boldsymbol{\theta},\sigma^2\mathbf{I}_n\right), \tag{3}$$
$$\boldsymbol{\theta}|\boldsymbol{\Sigma} \sim \mathcal{N}_{np}\left(\mathbf{0}_{np},\boldsymbol{\Sigma}\right), \tag{4}$$

where

$$\begin{aligned}\mathbf{A} &= \left(\mathbf{I}_n,\ldots,\mathbf{I}_n,\ldots,\mathbf{I}_n\right),\\ \boldsymbol{\Sigma} &= \text{Diag}\left(\boldsymbol{\Sigma}_j\right),\\ \boldsymbol{\theta} &= \left(\boldsymbol{f}_1',\ldots,\boldsymbol{f}_j',\ldots,\boldsymbol{f}_p'\right)'.\end{aligned}$$

$\mathbf{A}$ is a matrix of dimension $n \times np$ made up of a row of $n \times n$ identity matrices and $\boldsymbol{\Sigma}$, the prior covariance matrix, is a block diagonal matrix of dimension $np \times np$ with diagonal matrix elements $\boldsymbol{\Sigma}_j = (\sigma^2/\lambda_j)\mathbf{K}_j^-$, of dimension $n \times n$ for $j = 1,\ldots,p$. The matrices $\mathbf{K}_j^-$ are defined in the same way as in the previous univariate case. As in the case with one explanatory variable, λ_j represents the smoothing parameter for the jth spline function, *i.e.* the spline function for the jth explanatory variable. The parameter λ_j can take values over $\mathbb{R}_+$ and therefore in practice, its prior illicitation and its interpretation from the posterior distribution are difficult. We thus suggest to use an alternate parameterization

$$\phi_j = \frac{1}{1+\lambda_j}, \forall j.$$

It is directly seen that each ϕ_j varies between 0 and 1. It is worth noting that when λ_j goes to 0 (interpolation of the data), ϕ_j goes to 1, and when λ_j goes to ∞ (linear relation), ϕ_j goes to 0. Therefore, it is possible to specify a prior distribution for each ϕ_j which reflects our knowledge on the type of relation which is anticipated for an explanatory variable j. We can now write the covariance matrix for the vector of functional elements, $\boldsymbol{\theta}$, as

$$\boldsymbol{\Sigma} = \sigma^2\text{Diag}\left(\frac{\phi_j}{1-\phi_j}\mathbf{K}_j^-\right).$$

3 Bayesian variables selection for Generalized Additive Models

First, we considering the model with 1 explanatory variable, we have

$$\begin{aligned}\left[\boldsymbol{y}|\boldsymbol{\theta}_1,\sigma^2\right] &\equiv \mathcal{N}_n\left(\boldsymbol{\theta}_1,\sigma^2\mathbf{I}_n\right),\\ \left[\boldsymbol{\theta}_1|\boldsymbol{\Sigma}_1\right] &\equiv \mathcal{N}_n\left(\mathbf{0},\boldsymbol{\Sigma}_1\right),\end{aligned}$$

where $\boldsymbol{\Sigma}_1 = (\sigma^2/\lambda_j)\mathbf{K}_1^-$. The first distribution is proper while the second distribution is partially improper. From Lindley and Smith (1972; LS72), we get the posterior distribution of $\boldsymbol{\theta}_1$ which is proper (see for example Green and Silverman, 1994)

$$\left[\boldsymbol{\theta}_1|\boldsymbol{\Sigma}_1,\sigma^2,\boldsymbol{y}\right] \equiv \mathcal{N}_n\left(\boldsymbol{\theta}_1^*,\boldsymbol{\Sigma}_1^*\right),$$

where

$$\begin{aligned}\boldsymbol{\theta}_1^* &= \left(\mathbf{I}_n + \lambda_1 \mathbf{K}_1\right)^{-1} \boldsymbol{y},\\ \boldsymbol{\Sigma}_1^* &= \sigma^2 \left(\mathbf{I}_n + \lambda_1 \mathbf{K}_1\right)^{-1}.\end{aligned}$$

The marginal distribution is partially improper as can be found by calculating

$$\left[\boldsymbol{y}|\boldsymbol{\Sigma}_1, \sigma^2\right] = \frac{\left[\boldsymbol{y}|\boldsymbol{\theta}_1, \sigma^2\right]\left[\boldsymbol{\theta}_1|\boldsymbol{\Sigma}_1\right]}{\left[\boldsymbol{\theta}_1|\boldsymbol{\Sigma}_1, \sigma^2, \boldsymbol{y}\right]}.$$

More explicitly, it is given by

$$\left[\boldsymbol{y}|\boldsymbol{\Sigma}_1, \sigma^2\right] = \left(2\pi\sigma^2\right)^{\frac{-(n-2)}{2}} |\mathbf{I}_n + \lambda_1^{-1}\mathbf{K}_1^-|_+^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{y}'\left(\mathbf{I}_n + \lambda_1^{-1}\mathbf{K}_1^-\right)^- \boldsymbol{y}\right\}.$$

So far, nothing has been assumed concerning the prior distributions of the variance σ^2 and of the smoothing parameter λ_1 (or equivalently $\phi_1 = 1/(1+\lambda_1)$ (see section 2)). Leaving aside ϕ_1 for the moment and assuming that $\sigma^2 \sim \mathcal{IG}(a_\sigma, b_\sigma)$, we then have the following marginal distribution

$$\left[\boldsymbol{y}|\lambda_1, \mathbf{K}_1\right] = \frac{\left[\boldsymbol{y}|\lambda_1, \mathbf{K}_1, \sigma^2\right]\left[\sigma^2\right]}{\left[\sigma^2|\lambda_1, \mathbf{K}_1, \boldsymbol{y}\right]},$$

which we calculate to be

$$\left[\boldsymbol{y}|\lambda_1, \mathbf{K}_1\right] = \frac{\Gamma\left(\frac{2a_\sigma+(n-2)}{2}\right)}{\left(\pi b_\sigma\right)^{\frac{n-2}{2}}\Gamma\left(\frac{2a_\sigma}{2}\right)} |\mathbf{I}_n + \lambda_1^{-1}\mathbf{K}_1^-|_+^{-1/2} \left\{1 + \frac{\boldsymbol{y}'\left(\mathbf{I}_n + \lambda_1^{-1}\mathbf{K}_1^-\right)^- \boldsymbol{y}}{2b_\sigma}\right\}^{-\left(\frac{2a_\sigma+(n-2)}{2}\right)}$$

This is in the form of a multivariate t distribution; with the notation of Berger (1985) or Robert (2001), if it was proper, it could be written as

$$\mathcal{T}_{n-2}\left(2a_\sigma, \mathbf{0}, (b_\sigma/a_\sigma)\left\{\mathbf{I}_n + \lambda_1^{-1}\mathbf{K}_1^-\right\}\right).$$

It is partially improper though and this is thus a partially improper multivariate t distribution.

Now, if we do the same exercise with p variables, we get

$$\begin{aligned}\left[\boldsymbol{y}|\left\{\lambda_j, \mathbf{K}_j\right\}_{j=1,\ldots,p}\right] &= \frac{\Gamma\left(\frac{2a_\sigma+(n-2p)}{2}\right)}{\left(\pi b_\sigma\right)^{\frac{n-2p}{2}}\Gamma\left(\frac{2a_\sigma}{2}\right)}\\ &\times |\mathbf{I}_n + \sum_{j=1}^{p}\lambda_j^{-1}\mathbf{K}_j^-|_+^{-1/2}\\ &\times \left\{1 + \frac{\boldsymbol{y}'\left(\mathbf{I}_n + \sum_{j=1}^{p}\lambda_j^{-1}\mathbf{K}_j^-\right)^- \boldsymbol{y}}{2b_\sigma}\right\}^{-\left(\frac{2a_\sigma+(n-2p)}{2}\right)}. \qquad (5)\end{aligned}$$

Finally, concerning the prior distributions for the λ_js, it's better to put prior distributions on the ϕ_js given in section 2. Then, we have the following marginal distribution

$$\left[\boldsymbol{y}|\left\{\mathbf{K}_j\right\}_{j=1,\ldots,p}\right] = \frac{\left[\boldsymbol{y}|\left\{\phi_j, \mathbf{K}_j\right\}_{j=1,\ldots,p}\right]\left[\left\{\phi_j\right\}_{j=1,\ldots,p}\right]}{\left[\left\{\phi_j\right\}_{j=1,\ldots,p}|\left\{\mathbf{K}_j\right\}_{j=1,\ldots,p}, \boldsymbol{y}\right]}$$

But, here we can not explicitly calculate this distribution. To circumvent this issue we propose to use a method described by Chib and Jeliazkov (2001). The main idea is that for appropriate $\left\{\phi_j^*\right\}_{j=1,\ldots,p}$ we have

$$\left[\boldsymbol{y}|\left\{\mathbf{K}_j\right\}_{j=1,\ldots,p}\right] = \frac{\left[\boldsymbol{y}|\left\{\phi_j^*, \mathbf{K}_j\right\}_{j=1,\ldots,p}\right]\left[\left\{\phi_j^*\right\}_{j=1,\ldots,p}\right]}{\left[\left\{\phi_j^*\right\}_{j=1,\ldots,p}|\boldsymbol{y}, \left\{\mathbf{K}_j\right\}_{j=1,\ldots,p}\right]}$$

from which the marginal likelihood can be estimated by finding an estimate of the posterior $\left[\left\{\phi_j^*\right\}_{j=1,\ldots,p}|\boldsymbol{y}, \left\{\mathbf{K}_j\right\}_{j=1,\ldots,p}\right]$. Chib and Jeliazkov show a simulation-consistent estimate of the posterior ordinate is available as

$$\frac{G^{-1}\sum_{g=1}^{G}\alpha(\left\{\phi_j^g, \phi_j^*\right\}_{j=1,\ldots,p}|\boldsymbol{y})q(\left\{\phi_j^g, \phi_j^*\right\}_{j=1,\ldots,p}|\boldsymbol{y})}{L^{-1}\sum_{l=1}^{L}\alpha(\left\{\phi_j^*, \phi_j^l\right\}_{j=1,\ldots,p}|\boldsymbol{y})}$$

where $\left\{\phi_j^g\right\}_{j=1,\ldots,p}$ are the sample draws from the posterior distribution with the Metropolis-Hasting algorithm, $\left\{\phi_j^l\right\}_{j=1,\ldots,p}$ are draws from the proposal density $q(\left\{\phi_j^*, \phi_j\right\}_{j=1,\ldots,p}|\boldsymbol{y})$ and α is the probability to move.

4 Data analysis

4.1 Simulations results

We consider the following full model

$$\boldsymbol{y}|\sigma^2 \sim \mathcal{N}_3(\sum_{j=1}^{3} f_j(\boldsymbol{x}_{.j}), \sigma^2\mathbf{I}_3) \tag{6}$$

where the predictors $\boldsymbol{x}_j$, $j = 1, 2, 3$ were generated with an uniform distribution $[0, 10]$. The true model was established with $\boldsymbol{y}$ expected value equal to $f_1(\boldsymbol{x}_{.1}) + f_2(\boldsymbol{x}_{.2})$. We suppose $f_1(\boldsymbol{x}) = \sin(2\boldsymbol{x} + 2)$, $f_2(\boldsymbol{x}) = \cos(\boldsymbol{x})$, $f_3(\boldsymbol{x}) = 0.5\boldsymbol{x}$ and $\sigma^2 = 0.1$. Figure 2 represents simulated $\boldsymbol{y}$ with such a model.

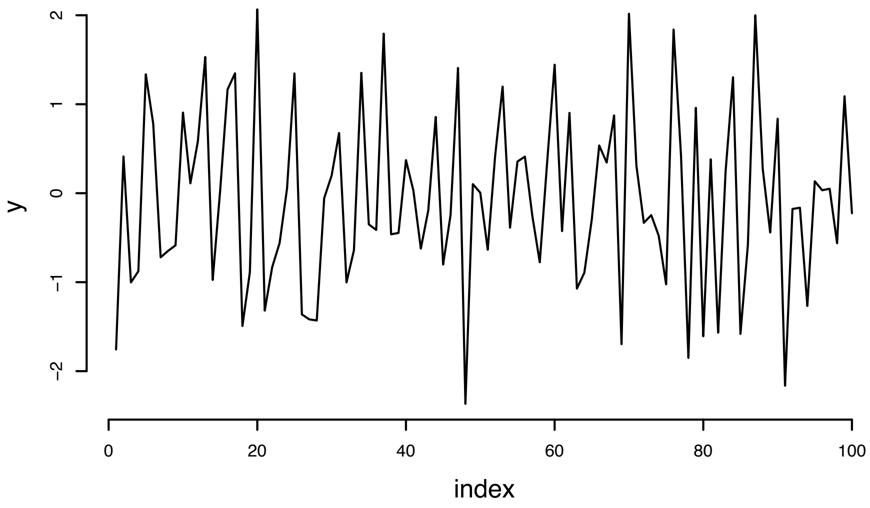

Figure 2: Simulated data with model (6)

For each possible model, i.e. for all explanatory variables combinations we can estimate the marginal distribution $\left[\boldsymbol{y}|\left\{\mathbf{K}_j\right\}_{j=1,\ldots,p}\right]$. These results are summarized in the Table 1. We know the best model have the more important probability, i.e. the model with the two first variablse in our case. This result seems coherent because this is the true model.

Selected variables	$\log P(\boldsymbol{y}\|\left\{\mathbf{K}_j\right\}_{j=1,\ldots,p})$
123	-179.55
12	-173.68
13	-194.93
23	-177.47
1	-176.15
2	-192.45
3	-196.28

Table 1: Estimated marginal distribution for the different models. The first column corresponds to the selected explanatory variables and the second column to the logarithm of associated marginal distribution

Of course, for the arrested model, we can estimate the smooth functions $f_1(.)$ and $f_2(.)$. Figures 3 and 4 represent the posterior median for these two estimated functions there true value used for the simulation. The true functions are represented by a solid line. The posterior median (dotted line) and the 95% credibility intervals (gray area) adequately follow the behavior of the true $f_1(.)$ and $f_2(.)$.

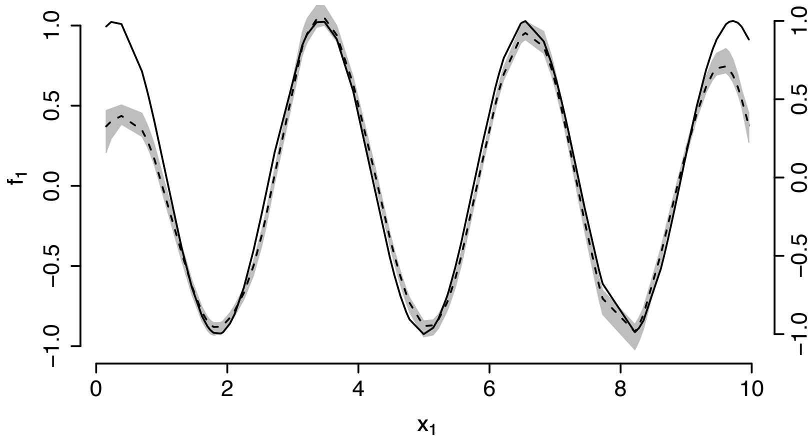

Figure 3: Posterior information about the function $f_1(.)$. The solid and dotted lines correspond to the true $f_1(.)$ and the posterior estimated median, respectively. The gray area represents the 95 % credibility intervals.

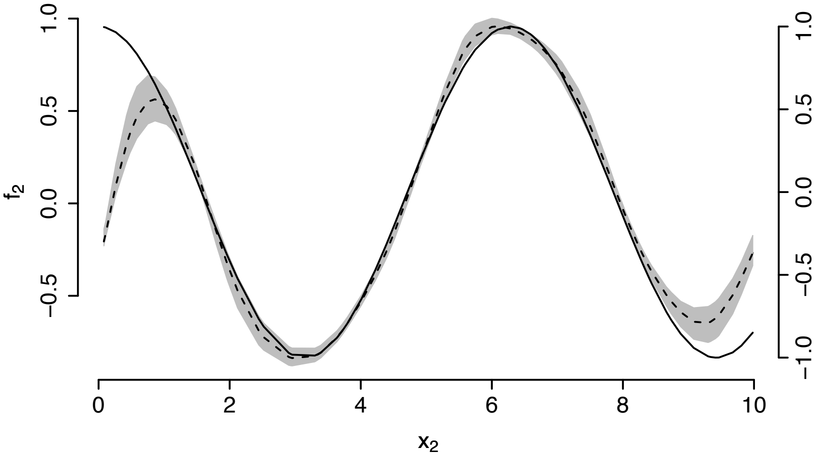

Figure 4: Posterior information about the function $f_2(.)$. The solid and dotted lines correspond to the true $f_2(.)$ and the posterior estimated median, respectively. The gray area represents the 95 % credibility intervals.

4.2 True data

5 Conclusion

References

F. Biondi. Evolutionary and moving response functions in dendrochronology. *Dendrochronologia*, 15:139–150, 1997.

T.J. Blasing, A.M. Solomon, and D.N. Duvick. Response functions revisited. *Tree-Ring Bulletin*, 44:1–15, 1984.

K.R. Briffa and Cook E.R. *Methods of dendrochronology*, pages 165–178. Kluwer Academic Publishers, Dordrecht, Netherlands, Boston, 1990.

S. Chib and I. Jeliazkov. Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association*, 96:270–281, 2001.

E. Cook and Leonardas. Kairikukstis. *Methods of dendrochronology : applications in the environmental sciences / edited by E.R. Cook and L.A. Kairiukstis*. Kluwer Academic Publishers, Dordrecht, Netherlands, Boston, 1990. ISBN 0792305868.

J.P. Cropper. *Tree-ring response functions : An elevation by means of simulations*. PhD thesis, The University of Arizona, 1985.

C. de Boor. *A Practical Guide to Splines*. Applied Mathematical Sciences, Springer-Verlag, New-York, 1978.

A.E. Douglass. Evidence of climatic effects in the annual rings of trees. *Ecology*, 1: 24–32, 1920.

A.E. Douglass. Climatic cycles and tree-growth. *Carnergie Institution of Washington publication*, 289(3), 1936.

B. Efron. Bootstrap methods : another look at the jackknife. *Annals of Statistics*, 7(1): 1–26, 1979.

B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

J. Esper, E. R. Cook, and F. H. Schweingruber. Low-frequency signals in long tree-ring chronologies for reconstructing past temperature variability. *Science*, 295:2250–2253, 2002.

M.N. Evans, B.K. Reichert, A. Kaplan, K.J. Anchukaitis, E.A. Vaganov, M.K. Hughes, and M.A. Cane. A forward modeling approach to paleoclimatic interpretation of tree-ring data. *J. Geophys. Res.*, 111, 2006.

H.C. Fritts. *Tree Rings and Climate*. Academic Press London, 1976.

H.C. Fritts, T.J. Blasing, B.P. Hayden, and J.E. Kutzbach. Multivariate techniques for specifying tree-growth and climate relationships and for reconstructing anomalies in paleoclimate. *Journal of Applied Meteorology*, 10(5):845–864, 1971.

D.M. Gates. *Biophysical Ecology*. Springer, New-York, 1980.

L.J. Graumlich. A 1000-year record of temperature and precipitation in the sierra nevada. *Quaternary Research*, 39:249–255, 1993.

J. Guiot. *Methods of dendrochronology*, chapter Methods of calibration, pages 165–178. Kluwer Academic Publishers, Dordrecht, Netherlands, Boston, 1990.

J. Guiot. The bootstrapped response function. *Tree-Ring Bulletin*, 51:39–41, 1991.

T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman and Hall, London, 1990.

T. Hastie and R. Tibshirani. Bayesian backfitting. *Statistical Science*, 15(3):196–223, 2000.

P.J. Kramer and T.T. Kozlowski. *Physiology if Woody Plants*. Elsevier, New-York, 1979.

M. E. Mann, Z. Zhang, M. K. Hughes, R. S. Bradley, S. K. Miller, and S. Rutherford. Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proc. Natl. Acad. Sci.*, 105:13252–13257, 2008.

M.E. Mann, R.S. Bradley, and M.K. Hughes. Northern hemisphere temperatures during the past millennium : inferences, uncertainties and limitations. *Geophysical Research Letters*, 2:759–762, 1999.

D.F. Morrison. *Applied Linear Statistical Methods*, page 562. Prentice-Hall, Englewood Cliffs, 1983.

B.J. Morzukh and G.A. Ruark. Principal components regression to mitigate the effect of multicillinearity. *Forest Science*, 37(1):191–199, 1991.

R. Villalba, T.T. Veblen, and Ogden J. Climatic influences on growth of subalpine trees in the colorado front range. *Ecology*, 75(5):1450–1462, 1994.

G. Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Royal Statist. Soc.*, 40:364–372, 1978.

C.A. Woodhouse. Artificial neural networks and dendroclimatic reconstructions : an example from the front range, colorado, usa. *The Holocene*, 9:521–529, 1999.