Méthodes bayésiennes semi-paramétriques d'extraction et de sélection de variables dans le cadre de la dendroclimatologie

Ophélie Guin

— Université Paris-Sud 11 —

Méthodes bayésiennes semi-paramétriques d'extraction et de sélection de variables dans le cadre de la dendroclimatologie

Thèse présentée pour l'obtention du grade de
DOCTUER DE LUNIVERSITÉ DE PARIS-SUD XI

ÉCOLE DOCTORALE : *Modélisation et Instrumentation en Physique, Energies, Géosciences et Environnement*

par

Ophélie GUIN

soutenue publiquement le 14 avril 2011

devant le jury

M^{me}. BEL Liliane	AgroParisTech	Rapporteure
M. GUIOT Joël	CEREGE	Examinateur
M. LEADLEY Paul	Université Paris-sud XI	Président
M. NAVEAU Philippe	CNRS	Directeur de thèse
M. PERREAULT Luc	Hydro-Québec	Rapporteur

Thèse préparée au Laboratoire des Sciences du Climat et de l'Environnement,
91191 Gif-sur-Yvette

Résumé

Selon le Groupe Intergouvernemental d'experts sur l'Evolution du Climat (GIEC), il est important de connaitre le climat passé afin de replacer le changement climatique actuel dans son contexte. Ainsi, de nombreux chercheurs ont travaillé à l'établissement de procédures permettant de reconstituer les tempratures ou les précipitations passées à l'aide d'indicateurs climatiques indirects. Ces procédures sont généralement basées sur des méthodes statistiques mais l'estimation des incertitudes associées à ces reconstructions reste une difficulté majeure. L'objectif principal de cette thèse est donc de proposer de nouvelles méthodes statistiques permettant une estimation précise des erreurs commises, en particulier dans le cadre de reconstructions à partir de données sur les cernes d'arbres.

De manière générale, les reconstructions climatiques à partir de mesures de cernes d'arbres se déroulent en deux étapes : l'estimation d'une variable cachée, commune à un ensemble de séries de mesures de cernes, et supposée climatique puis l'estimation de la relation existante entre cette variable cachée et certaines variables climatiques. Dans les deux cas, nous avons développé une nouvelle procédure basée sur des modèles bayésiens semi-paramétriques. Tout d'abord, concernant l'extraction du signal commun, nous proposons un modèle hiérarchique semi-paramétrique qui offre la possibilité de capturer les hautes et les basses fréquences contenues dans les cernes d'arbres, ce qui était difficile dans les études dendroclimatologiques passées. Ensuite, nous avons développé un modèle additif généralisé afin de modéliser le lien entre le signal extrait et certaines variables climatiques, permettant ainsi l'existence de relations non-linéaires contrairement aux méthodes classiques de la dendrochronologie. Ces nouvelles méthodes sont à chaque fois comparées aux méthodes utilisées traditionnellement par les dendrochronologues afin de comprendre ce qu'elles peuvent apporter à ces derniers.

Mots clés : estimation bayésienne, modèles hiérarchiques, splines, sélection de variables, dendrochronologie, reconstructions climatiques

Abstract

As stated by the Intergovernmental Panel on Climate Change (IPCC), it is important to reconstruct past climate to accurately assess the actual climatic change. A large number of researchers have worked to develop procedures to reconstruct past temperatures or precipitation with indirect climatic indicators. These methods are generally based on statistical arguments but the estimation of uncertainties associated to these reconstructions remains an active research field in statistics and in climate studies. The main goal of this thesis is to propose and study novel statistical methods that allow a precise estimation of uncertainties when reconstructing from tree-ring measurements data. Generally, climatic reconstructions from tree-ring observations are based on two steps. Firstly, a hidden environmental hidden variable, common to a collection of tree-ring measurements series, has to be adequately inferred. Secondly, this extracted signal has to be explained with the relevant climatic variables. For these two steps, we have opted to work within a semi-parametric bayesian framework that reduces the number of assumptions and allows to include prior information from the practitioner. Concerning the extraction of the common signal, we propose a model which can catch high and low frequencies contained in tree-rings. This was not possible with previous dendroclimatological methods. For the second step, we have developed a bayesian Generalized Additive Model (GAM) to explore potential links between the extracted signal and some climatic variables. This allows the modeling of non-linear relationships among variables and strongly differs from past dendrochronological methods. From a statistical perspective, a new selection scheme for bayesien GAM was also proposed and studied.

Key words : bayesian estimation hierarchical models, spline, variables selection, dendrochronology, climatic reconstructions

Remerciements

Mes premiers remerciements vont à Philippe pour le rôle qu'il a joué tout au long de ma thèse. Merci de ses précieux conseils, de sa confiance et de son soutient qui ont permis à ce travail de recherche d'être ce qu'il est aujourd'hui. Mais au delà du cadre scientifique, merci pour son écoute, sa compréhension et son optimisme qui m'ont fait vivre trois années enthousiasmantes et épanouissantes.

Je souhaiterais ensuite exprimer toute ma gratitude Luc Perreault et Lilane Bel pour avoir accepté de rapporter ma thèse, avec une pensée particulière pour Liliane que je ne remercierais jamais assez des judicieux conseils qu'elle m'a donné au cours de ces trois années. Je remercie également Paul Leadley et Joël Guiot pour mavoir fait l'honneur de participer à mon jury de soutenance. Merci Joël pour les analyses critiques que tu as pu faire de mon travail et qui m'ont permis d'avancer. De manière plus générale, je les remercie de l'intérêt qu'ils ont manifesté à ce travail et de la pertinence de leurs commentaires.

Merci à Jean-Jacques pour nos collaborations scientifiques mais également pour les moments agréables que nous avons partagés. Merci à lui et Marie-Antoinette pour m'avoir si bien reçu chez eux.

Merci à James qui lors de son passage Paris m'a apporté beaucoup et m'a permis de faire un bond dans la compréhension des splines en bayésien.

Cette thèse a été effectuée, au LSCE, au sein de l'quipe Estimr dont je teins à remercier tous les membres. En particulier Pascal pour m'avoir permis d'effectuer cette thèse et pour son aide inconditionnelle; Mathieu pour ses bons conseils et les agréables discutions que nous avons eu ; Cédric, Jérôme et Julien avec qui j'ai pu partager mes angoisses de doctorante ; Malaak et Julie pour nos discutions scientifiques mais également pour les moments de détente que nous avons pu partager.

Ayant commencé ma thèse au sein de l'équipe Clim, je suis restée très proche de ses membres et je les remercie pour les extraordinaires moments passés ensemble. Merci Jean-Yves, Christophe, Gilles R., Florence, Didier P., Sylvie, Massa, Didier R., Gilles C., Marie-Noëlle, Jean-Claude. Avec un merci tout particulier à Nathaëlle avec qui j'ai pu partager mes joies et mes peines au jour le jour. Merci également à Emilie, Katy, Anne.

Merci à la fondation MAIF d'avoir rendu possible ce travail de recherche et de m'avoir fait confiance.

Pour finir un grand merci à mes proches et à ma famille qui m'ont accompagné et qui m'ont soutenu dans les moments difficiles ou de doutes.

Table des matières

Table des figures

Chapitre 1

Introduction générale : La dendroclimatologie

Rings in the branches of sawed trees show the number of years and, according to their thickness, the years which were more or less dry. - Léonard de Vinci, XVème siècle

Ce chapitre d'introduction générale pose quelques bases de la dendroclimatologie. Nous ferons un rapide rappel historique puis nous verrons de quelle manière les statistiques ont très largement participé au développement de cette science et ce qu'elles lui ont apporté. Nous en profiterons également pour positionner cette thèse dans le contexte actuel de la dendroclimatologie et pour décrire la manière dont va s'articuler ce document.

1.1 Préambule

Dans le contexte des changements climatiques actuels (IPCC, 2007), il est important de connaitre le climat passé. En effet, cela nous permettrait de nous rendre compte de ce qu'est l'évolution climatique naturelle et ainsi de replacer le changement climatique d'aujourd'hui dans son contexte. Cependant, la connaissance du climat passé sur plusieurs siècles n'est pas aisée lorsqu'on sait que la période pour laquelle on dispose de mesures directes sur des variables climatiques, telles que les précipitations ou les températures, est bien plus courte que la période qui nous intéresse. Si on prend par exemple la France, nous ne disposons de données météorologiques qu'à partir du début du XXème siècle. Afin de surmonter cette difficulté, de nombreux chercheurs ont travaillé à mettre en place des procédures permettant de reconstituer les températures ou les précipitations passées. On peut notamment citer les travaux de Mann et al. (2008) ou de Esper et al. (2002) qui ont permis la reconstruction des températures annuelles de l'hémisphère nord au cour du dernier millénaire. Pour cela, ils ont dû utiliser des indicateurs climatiques indirects nommés proxies. L'un des proxy les plus connus, qui va nous intéresser tout au long de cette thèse, est la croissance des arbres, et en particulier l'évolution de leurs cernes au cours du temps. En effet, on sait que les arbres créent un cerne par an et que la largeur, l'aire ou la densité de ce cerne est reliée à des variables environnementales. L'étude des relations existant entre ces cernes d'arbres et le climat est connue sous le nom de dendroclimatologie.

1.2 Quelques repères historiques

On fait souvent débuter l'histoire de la dendroclimatologie, et plus largement celle de la dendrochronologie (dendron = arbre et chronos = temps), au début du XXème siècle avec l'astronome Andrew Ellicot Douglass. Cependant, on peut considérer que cette Science a vu le jour dès l'aube de l'humanité, même si les choses n'étaient pas clairement formalisées. En effet, l'être humain s'est intéressé très tôt aux arbres, ces derniers ayant occupé une place majeure dans son développement : ils fournissaient de la nourriture mais également le matériau de construction assurant la sédentarisation (Dumas, 2002). Il est probable que le questionnement sur la croissance radiale des arbres soit apparu en même temps que le développement de leur coupe. L'apparition de la scie, en faisant apparaître une image claire de la succession des cernes, a sans doute rapidement permis de préciser leur caractère annuel. Il faudra pourtant attendre le XVème siècle et Léonard de Vinci pour trouver la première référence écrite portant sur le rôle de l'arbre dans l'enregistrement des fluctuations climatiques (Stalling et al., 1937). Par la suite, grâce au développement des techniques d'observation microscopique, toute une génération de savants s'intéressèrent à l'anatomie de la croissance des arbres. Ils se posèrent des questions portant sur la physique, les sciences de la vie, la botanique ou la foresterie. On peut entre autres citer Malpighi, Grew, Hales, Buffon, Duhamel du Montceau. Dès le milieu du XIXème siècle T. Hartig puis R. Hartig proposent une conception claire du fonctionnement cambial des arbres. A partir de la seconde moitié du XIXème siècle, en Europe comme aux Etats-Unis, émerge une véritable analyse des séquences chronologiques des cernes ainsi qu'une diversification des thématiques liées à ces dernières. On voit apparaître les premiers frémissements de la dendroclimatologie avec notamment Pokorny (1892) qui compare des séquences moyennes de cernes et des données météorologiques ou Kuechler (Campbell, 1949) qui traite du rapport entre la variabilité des cernes et la sécheresse.

A l'issue de tous ces travaux, il est acquis que le fonctionnement cambial de l'arbre conduit à la formation d'un cerne annuel dans lequel on peut même reconnaître la

marque des saisons (Hill, 1770). Mais, c'est avec l'astronome Andrew Ellicot Douglass que le large potentiel que représente les séquences chronologiques de croissance des cernes, tant d'un point de vue archéologique que climatologique, fut réellement exploité. Ses travaux font de lui le père incontestable de la dendrochronologie. Fritts (1976), dans son ouvrage de base, retrace de façon détaillée son cheminement, depuis ses premières interrogations en 1901 sur la mise en évidence d'un lien entre les cycles de l'activité solaire et le climat jusqu'à la création à Tucson du « laboratory of Tree Ring Research » en 1937.

Par la suite ce premier laboratoire entièrement consacré à la dendrochronologie a joué un rôle de modèle pour les multiples unités de recherche qui ont vu le jour après la seconde guerre mondiale et a suscité une large expansion de la discipline. Ainsi, la fin du XXème siècle fut marquée par de très nombreux travaux scientifiques sur ce thème et par l'apparition de deux grandes figures de la dendrochronologie. Tout d'abord, Harold C. Fritts, successeur de Douglass à Tucson. Il est surtout connu pour avoir introduit les techniques de calculs intensifs sur ordinateur en dendrochronologie, nous permettant ainsi d'aller plus loin dans l'analyse des relations cernes-climat dans leur dimension spatiale et temporelle. Son ouvrage majeur « Tree Rings and Climat » (1976), dans lequel toutes les bases conceptuelles de la dendrochronologie sont abordées, reste une référence toujours d'actualité qu'il est impossible d'ignorer. Le second, Hubert Polge, est le fondateur en 1964 de la « Station de recherche sur la qualité du bois » à Champenoux (France) et l'inventeur de l'analyse densitométrique des bois (1966) permettant le développement de chronologies basées non plus sur la largeur des cernes mais sur les fluctuations de la densité du bois.

Ainsi enrichis de l'analyse densitométrique, nous disposons aujourd'hui d'une information environnementale quantitative sur de vastes espaces et sur des centaines d'années, voir des millénaires. Il appartient donc aux dendrochronologues de la déchiffrer et de l'interpréter.

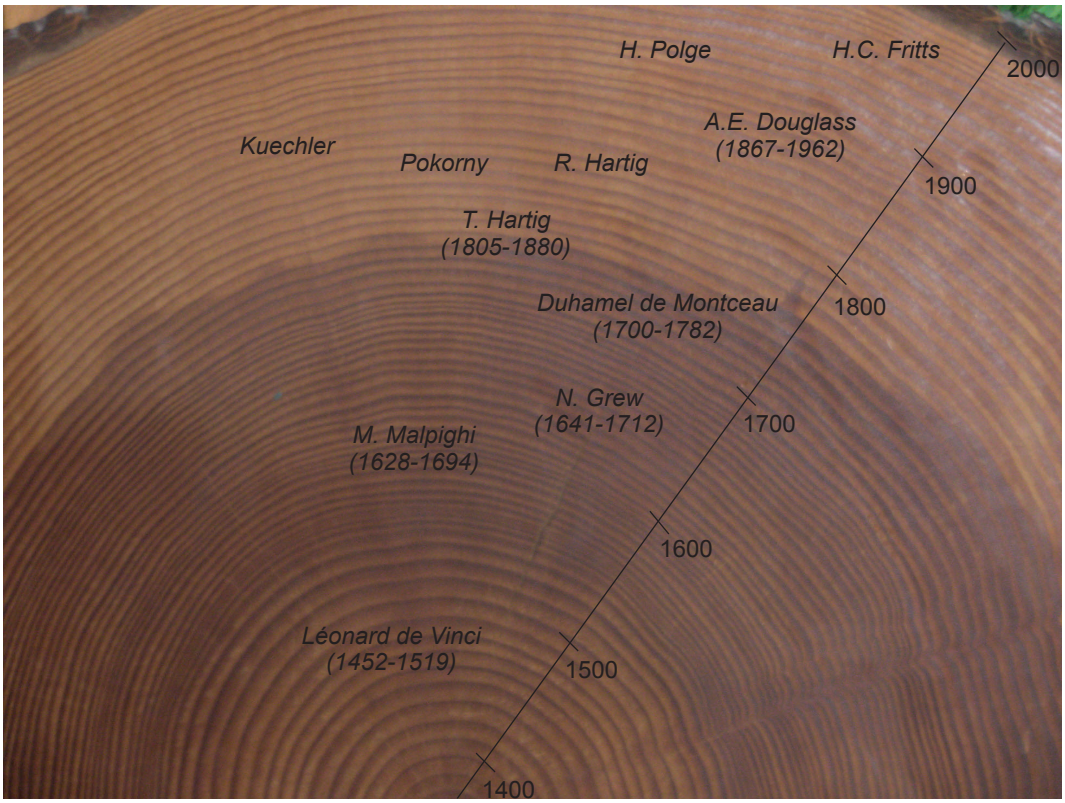

Figure 1.1 Les principaux protagonistes de la dendrochronologie

1.3 Dendroclimatologie et statistiques

Malgré les nombreux travaux qui ont été effectués, l'une des grandes difficultés de la dendroclimatologie aujourd'hui reste l'interprétation environnementale de la croissance des cernes d'arbres et en particulier le choix des variables climatiques qui permettent d'expliquer cette croissance. Est-ce que les différentes mesures de cernes annuels sont liés à la moyenne journalière des précipitations durant les mois d'été, au plus grand nombre de jours consécutifs sans pluie au cours d'une année, aux températures saisonnières, etc... ? Le nombre de possibilités est sans fin et dépend de l'espèce des arbres et de leur région géographique. L'expertise des dendrochronologues sur ce sujet est extrêmement précieuse mais ne permet pas toujours de répondre complètement à cette question. Un solution est donc de faire appel à la statistiques. De nombreuses techniques ont été développées et leur amélioration fait toujours l'objet

d'une recherche très active, ce qui fait aujourd'hui de la statistiques un élément majeur et incontournable de la dendroclimatologie.

Tout d'abord, intéressons nous à l'analyse des séquences chronologiques de cernes d'arbres. De manière classique, les dendrochronologues font l'hypothèse qu'un signal commun est partagé par tous les arbres d'un même site et que ce dernier est sûrement la conséquence de plusieurs facteurs environnementaux, voir climatiques. Ainsi, si on a des séries de mesures de cernes d'arbres pour un site donné, la question est de savoir comment extraire ce signal à partir de notre jeu de données. Pour cela les dendrochronologues modélisent la croissance annuelle d'un arbre à l'aide d'un modèle additif, souvent appelé modèle d'agrégation linéaire (Cook and Kairikukstis, 1990; Buckley, 2009) :

$$R_t = A_t + C_t + D_t + \epsilon_t, \tag{1.1}$$

où, pour l'année t, R_t correspond à la croissance de l'arbre, A_t à une tendance reliée à l'âge de l'arbre (créé par certains éléments physiologiques), C_t à un signal environnemental lié au climat et D_t à des facteurs de perturbations internes ou externes à la forêt (éruption d'insectes, feux, etc...). Quant à ϵ_t, il est supposé représenter du bruit, c'est-à-dire la variabilité qui reste inexpliquée. Dans de nombreuses études, le site d'intérêt est sélectionné afin de minimiser la possibilité de processus écologiques inter-annuels ou extrêmes affectant la croissance des arbres. Si on ajoute à ceci le fait que les dendrochronologues ont tendance à travailler sur des moyennes de séries, le poids de D_t devient très faible. Comme on peut aisément l'imaginer, notre intérêt principal réside dans le fait d'estimer la composante C_t de l'équation (1.1).

Le processus d'estimation de cette composante est généralement appelé standardisation et peut se résumer en trois phases (Melvin and Briffa, 2008). Tout d'abord, une tendance reliée à l'âge des arbres est estimée puis enlevée pour chacune des séquences chronologiques de notre jeu de données afin d'annuler l'effet de l'âge A_t. Ensuite, chaque mesure de cerne est divisée par la valeur ajustée correspondante, obtenue à partir de la régression précédente. On obtient alors une chronologie par

arbre échantillonné. Enfin, pour finir, la composante commune cachée C_t est calculée comme étant la moyenne arithmétique, année par année, de toutes les chronologies obtenues. On peut noter que ces différentes étapes sont équivalentes à un modèle multiplicatif et non à un modèle additif comme celui proposé par l'équation (1.1). Le choix d'un modèle additif est fait pour faciliter la compréhension, mais il est bien connu que certaines propriétés des cernes sont multiplicatives. Les données de croissance de cernes d'arbres connaissent donc une transformation logarithmique dans le modèle (1.1).

Depuis les premiers travaux de Douglass (1920, 1936) de nombreuses variantes de méthodes de standardisation ont été développées, les discussions portant surtout sur la manière d'estimer la tendance A_t reliée à l'âge. Cook and Kairikukstis (1990) les divise en deux classes : les méthodes déterministes et les méthodes stochastiques. Les méthodes déterministes ajustent un modèle mathématique sur les séries de croissance de cernes tel que par exemple un modèle de tendance linéaire, un modèle exponentiel négatif (Fritts et al., 1969) ou un modèle polynomial (Graybill, 1979). Les méthodes stochastiques, qui comme leur nom l'indique font appel à la statistiques, sont plus flexibles et peuvent facilement s'adapter à différentes formes de séries de données. On peut notamment citer la modélisation par processus autorégressifs (Guiot, 1981), le lissage par splines cubiques (Cook and Peters, 1981) ou le lissage exponentiel (Barefoot et al., 1974). Aujourd'hui, ces méthodes bien que toujours très utilisées connaissent quelques critiques. En effet, lorsqu'on enlève la tendance A_t liée à l'effet de l'âge, on supprime en même temps toute l'information basse fréquence contenue dans les cernes d'arbres. Les méthodes de standardisation classiques semblent donc seulement adaptées pour capturer la variabilité annuelle mais pas les tendances décennales ou centennales qui pourraient pourtant contenir une information climatique. Les dendrochronologues ont tenté de répondre à cela en développant de nouvelles méthodes (Regional Curve Standardization, Adaptative Regional Growth Curve (Nicault et al., 2010)) qui s'appuient principalement sur la statistique, renforçant encore leur rôle en dendroclimatologie.

La seconde étape importante pour la reconstruction de températures ou de précipitations est, comme nous l'avons indiqué précédemment, le choix des variables climatiques expliquant la croissance des arbres, et en particulier le signal commun extrait que nous venons de décrire. D'un point de vue statistique, ce problème peut être vu comme un problème de sélection de variables. Les méthodes les plus couramment employées par les dendrochronologues sont les « fonctions de corrélation »et les « fonctions de réponse »(Fritts et al., 1971; Blasing et al., 1984), la relation entre la croissance des cernes d'arbres et le climat étant supposée linéaire. Pour chacune de ces méthodes, le terme de « fonction »indique une séquence de coefficients calculés entre un signal commun extrait ou une chronologie de cernes d'arbres et des variables climatiques. Dans le cas des fonctions de corrélation, ces coefficients sont estimés de façon univariée à l'aide de la corrélation de Pearson (Morrison, 1983), alors que pour les fonctions de réponse les coefficients sont estimés de façon multivariée à l'aide d'un modèle de régression en composantes principales (Briffa and Cook, 1990; Morzukh and Ruark, 1991).

Afin d'interpréter au mieux les résultats obtenus, il est important de pouvoir estimer de la manière la plus exacte possible la significativité statistique des coefficients calculés. En effet, seuls les variables climatiques correspondant à des coefficients significatifs peuvent être considérées comme ayant un impact sur la croissance des arbres. Cependant, il a été démontré, en particulier pour les fonctions de réponse, que certains coefficients passaient par erreur le test de significativité (Cropper, 1985; Morzukh and Ruark, 1991). Ainsi, le nombre de coefficients significatifs est généralement plus important lorsqu'on utilise les fonctions de réponse que lorsqu'on utilise les fonctions de corrélation (Villalba et al., 1994). La solution retenue pour résoudre cette difficulté a été d'estimer l'erreur faite sur l'estimation des coefficients à l'aide du bootstrap (Efron, 1979; Efron and Tibshirani, 1986; Guiot, 1990, 1991).

1.4 Positionnement de cette thèse

Dans le rapport du Committee on Surface Temperature Reconstructions for the Last 2000 Years (2006) on peut lire que les procédures standards de reconstructions de températures à partir de proxies sont généralement basées sur des méthodes statistiques raisonnables mais qu'elles doivent être associées à l'estimation d'incertitudes. Le comité insiste clairement sur le fait qu'une caractérisation plus rigoureuse des erreurs statistiques commises au cours de la reconstruction est indispensable.

Si cette remarque est valable pour les reconstructions climatiques en général, elle l'est aussi pour les reconstructions à partir de cernes d'arbres. En effet, de nombreuses sources d'incertitudes existent, liées à notre compréhension incomplète de la croissance radiale des arbres et à l'impact de certains effets environnementaux non observables sur cette dernière (Hughes et al., 2011). L'importance de ces incertitudes peut considérablement varier d'une étude à l'autre, mais on ne peut jamais considérer qu'il n'en existe pas. Ainsi par exemple, il est extrêmement rare que l'on connaisse avec exactitude le détail des signaux environnementaux cachés dans les séries de cernes d'arbres et décrit dans la section précédente.

Généralement les discussions portent sur ce qu'on pourrait qualifier d'incertitudes biologiques (l'impact de l'espèce des arbres étudiés, du type de mesure de cernes utilisé ou de l'environnement dans lequel l'arbre pousse) ce qui s'explique simplement par le fait que l'on cherche à identifier des signaux environnementaux à partir de séries temporelles biologiques. Cependant, il existe une autre forme d'incertitudes liée au développement des méthodes d'extraction de chronologies ou de reconstruction. Dans ce cas, on peut parler d'incertitudes statistiques. Par exemple, nous avons vu que les chronologies extraites sont des fonctions moyennes obtenues à partir de plusieurs séries correspondant à différents arbres, le but de cette moyenne étant de faire disparaître le bruit contenu dans les cernes. Pourtant, la plupart du temps, ceci n'est pas suffisant et le bruit n'est jamais complètement éliminé.

L'un des pionniers de la dendroclimatologie, Edmund Schulman, a eu très tôt

conscience de cette difficulté. Il a alors proposé de faire le ratio de la sensibilité moyenne de la chronologie moyenne par la moyenne de la sensibilité moyenne de chaque série individuelle (Schulman, 1956). Ce ratio permettait d'évaluer la « puissance »de la chronologie extraite, autrement dit suivant sa valeur, il indique le fait que la chronologie soit plus ou moins bruitée. Par la suite, avec le développement des ordinateurs dans les années soixante, Fritts introduisit l'utilisation de l'analyse de la variance (ANOVA) afin de décrire de manière quantitative les sources d'incertitudes dans les chronologies des cernes d'arbres (Fritts, 1963). Une autre façon de quantifier les incertitudes en dendroclimatologie est le développement d'intervalles de confiance annuels pour les chronologies de cernes d'arbres. Ces derniers peuvent être facilement estimés en utilisant des méthodes de bootstrap (Till and Guiot, 1990; Cook and Kairikukstis, 1990).

La quantification des incertitudes est donc aujourd'hui l'un des enjeux majeurs de la dendroclimatologie et des reconstructions climatiques en général. La problématique de cette thèse se situe dans la continuité des travaux qui ont été menés pour répondre à ce besoin, l'objectif principal de ce travail étant de proposer de nouvelles méthodes statistiques à la dendroclimatologie permettant une estimation précise des erreurs commises. Pour cela nous avons fait le choix d'avoir recours une à branche particulière de la statistique : la statistique bayésienne non-paramétrique.

1.5 Articulation de ce document

Les chapitres suivants s'attacheront à montrer en quoi la statistique bayésienne peut apporter quelque chose à la dendroclimatologie.

En quelques mots, dans le chapitre 2 nous commencerons par faire un petit rappel sur ce que sont les modèles bayésiens hiérarchiques et plus généralement la statistique bayésienne. Afin d'illustrer notre propos et de donner un exemple concret d'application des modèles bayésiens hiérarchiques à la dendroclimatologie nous présenterons un travail effectué en collaboration avec Jean-Jacques Boreux et visant à extraire un

signal commun haute fréquence à partir de séries de cernes d'arbres.

Le chapitres 3 présentera une nouvelle méthode, basée sur la statistique bayésienne, pour l'extraction d'un signal commun, supposé climatique, contenu dans les cernes d'arbres d'une même région géographique. Nous comparerons cette méthode à l'une des méthodes la plus utilisée aujourd'hui en dendroclimatologie et nous discuterons des avantages et des inconvénients de chacune.

Dans le chapitre 4 nous nous polariserons sur l'un des buts principaux de la dendroclimatologie, à savoir les reconstructions climatiques. Nous verrons les résultats que nous pouvons obtenir dans le cadre de reconstructions de séries de précipitations en Provence calcaire, à l'aide des méthodes bayésiennes qui auront été détaillées précédemment. Nous tenterons également de montrer ce que ces résultats peuvent apporter comme informations supplémentaires sur la connaissance du climat passé.

Le chapitre 5 se présentera de la même manière que le chapitre 3. Nous y présenterons une méthode de sélection de variables dans le cadre de modèles bayésiens additifs généralisés, permettant d'expliquer au mieux le signal extrait des cernes d'arbres de manière climatique. Nous comparerons également cette méthode à une méthode très classique de la dendroclimatologie.

Le chapitre 6, enfin, esquissera une synthèse avant de présenter quelques perspectives et problèmes non résolus.

Signalons enfin que cette introduction s'est attachée à présenter les problématiques de la dendroclimatologie car ce sont principalement ces questions qui ont guidé notre travail vers les problèmes statistiques qui seront présentés tout au long du manuscrit. Ainsi, différents domaines des statistiques ont été visités, allant de l'analyse de séries temporelles à l'estimation de variables latentes (chapitre 3) ou à la sélection de variables (chapitre 5). Certains des résultats de nature statistique qui seront présentés n'en restent pas moins utilisables pour d'autres applications que l'étude des relations cernes-climat.

Bibliographie

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

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

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

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

T.N. Campbell. The pioneer tree-ring work of jacob keuchler. *Tree-Ring Bull.*, 15 (3) :16–20, 1949.

E. Cook and L. Kairikukstis. *Methods of Dendrochronology : Applications in the Environmental Sciences*. Kluwer Academic Publishers, Dordrecht, Netherlands, Boston, 1990.

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

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

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.

R. Dumas. *Traité de l'arbre, essai d'une philosophie occidentale*. Actes Sud, 2002.

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, 2002.

H.C. Fritts. Computer programs for tree-ring research. *Tree-Ring Bull.*, 25(3-4) :2–7, 1963.

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

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

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.A. Graybill. Revised computer programs for tree-ring research. *Tree-Ring Bulletin*, 39 :77–82, 1979.

J. Guiot. *Analyse Mathématiques de Données Géophysiques, Applications à la Dendroclimatologie*. PhD thesis, Louvain-la-Neuve, 1981.

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.

J. Hill. *The construction of timber from it's early growth explained by the microscope and prooved from experiment.* 1770.

M. K. Hughes, T.W. Swetnam, and H.F. Diaz. *Dendroclimatology.* Springer, 2011.

IPCC. Climate change 2007 : The physical science basis. Technical report, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 2007.

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.

T. Melvin and K.R. Briffa. A "signal-free" approach to dendroclimatic standardisation. *Dendrochronologia*, 27 :71–86, 2008.

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.

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

Committee on Surface Temperature Reconstructions for the Last 2000 Years. Surface temperature reconstructions for the last 2000 years. Technical report, National Research Council, 2006.

A. Pokorny. Methode um denmeteorologischen coeffizienten des jährlichen holzzuwashes der dicotyledonenstämme zu ermitteln. *Tharendter forstl. Jb*, 22 :81, 1892.

E. Schulman. *Dendroclimatic changes in semiarid America.* University of Arizona Press, Tucson, 1956.

W.S. Stalling, E. Schulman, and Douglass A.E. Some early paperson tree rings. *Tree-Ring Bull.*, 3(4) :27–28, 1937.

C. Till and J. Guiot. Reconstruction of precipitation in morocco since a d 1100 based on cedrus atlantica tree-ring widths. *Quaternary Research*, 33 :337–351, 1990.

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.

Chapitre 2

Modèles bayésiens hiérarchiques et dendroclimatologie

When we make a scientific generalization we do not assert the generalization and its consequences with certainty ; we assert that they have a high degree of probability on the knowledge available to us at the time, but that this probability may be modified by additional knowledge. - Jeffreys 1931

Dans ce chapitre nous rappellerons rapidement ce que sont les modèles bayésiens hiérarchiques et plus généralement la statistique bayésienne. Nous donnerons quelques pistes de réflexion sur l'intérêt que ces derniers peuvent avoir dans le contexte des sciences du climat et de l'environnement, et en particulier en matière de reconstructions climatiques. Afin d'illustrer notre propos et de donner un exemple concret d'application des modèles bayésiens hiérarchiques à la dendroclimatologie, nous présenterons un travail effectué en collaboration avec Jean-Jacques Boreux visant à extraire un signal commun haute fréquence à partir de séries de cernes d'arbres.

2.1 Modèles bayésiens hiérarchiques, pourquoi ?

Les modèles bayésiens hiérarchiques ont été de plus en plus utilisés au cours de ces dernières décennies et aujourd'hui nous en trouvons de nombreuses applications dans les sciences du climat et de l'environnement. On peut par exemple citer Berliner et al. (2000) qui ont étudié des prédictions long terme de températures pour le Pacifique via un modèle bayésien hiérarchique dynamique, Cooley et al. (2005) qui ont mis en place un modèle bayésien hiérarchique pour estimer les retraits glacières en Bolivie en utilisant la croissance des lichens comme proxy, ou Cooley et al. (2007) qui ont estimé le niveau de retour des précipitations extrêmes en combinant un modèle bayésien hiérarchique avec la théorie des valeurs extrêmes. Concernant les reconstructions climatiques à partir de différents proxies, et en particulier des cernes d'arbres, Robertson et al. (1999) furent les premiers à introduire l'approche bayésienne. Il s'en suivit une série de papiers (Vasko et al., 2000; Toivonen et al., 2001; Korhola et al., 2002) formulant de manière détaillée l'utilisation des techniques bayésiennes aux reconstructions climatiques. Récemment, Hooten and Wikle (2007) utilisèrent un modèle bayésien afin de modéliser la dynamique de croissance spatio-temporelle des feuilles de pins.

Mais qu'est-ce qu'un modèle bayésien hiérarchique ? Que peut-il apporter à la dendroclimatologie et plus largement à tout ce qui touche aux questions climatiques ou environnementales ?

Tout d'abord, concernant la procédure de mise en place d'un modèle bayésien hiérarchique, elle peut se résumer en deux étapes schématisées par la figure 2.1 (Boreux et al., 2009b). Dans la première étape il s'agit de définir un modèle de probabilités total. Plus précisément, on définit les équations de notre modèle et on associe à chacun des paramètres de ce dernier une distribution de probabilités, appelée distribution *a priori*. Comme son nom l'indique, un modèle bayésien hiérarchique doit être défini selon une hiérarchie de couches. La première couche, habituellement appelée couche des données, caractérise les observations, la seconde couche, appelée couche

du processus, modélise le processus latent et la troisième couche, appelée couche des paramètres, décrit l'information concernant les paramètres qui contrôlent le processus latent. L'avantage d'une telle structure est que, au travers de ces différentes couches, nous pouvons modéliser des processus assez complexes d'une manière assez simple. Ceci est très important pour les sciences du climat et de l'environnement car souvent nous cherchons à modéliser des processus complexes qui sont eux-mêmes influencés de manière plus ou moins directe par d'autres processus. D'autre part, la couche du processus latent présente plusieurs avantages. Tout d'abord, il est fréquent que nous connaissions des relations déterministes provenant de la Physique entre certains paramètres. Cette seconde couche nous permet d'intégrer facilement des équations physiques et ainsi de prendre en compte ce type d'information. De plus, même dans le cas de modèles à couche latente simple, la linéarité ne nous est pas imposée contrairement aux modèles classiques tels que les modèles à effets aléatoires ou les modèles mixtes. On peut citer par exemple Eckert et al. (2010) qui, à l'aide d'un modèle bayésien hiérarchique, modélisèrent des ruptures très franches d'altitudes d'arrêt d'avalanches.

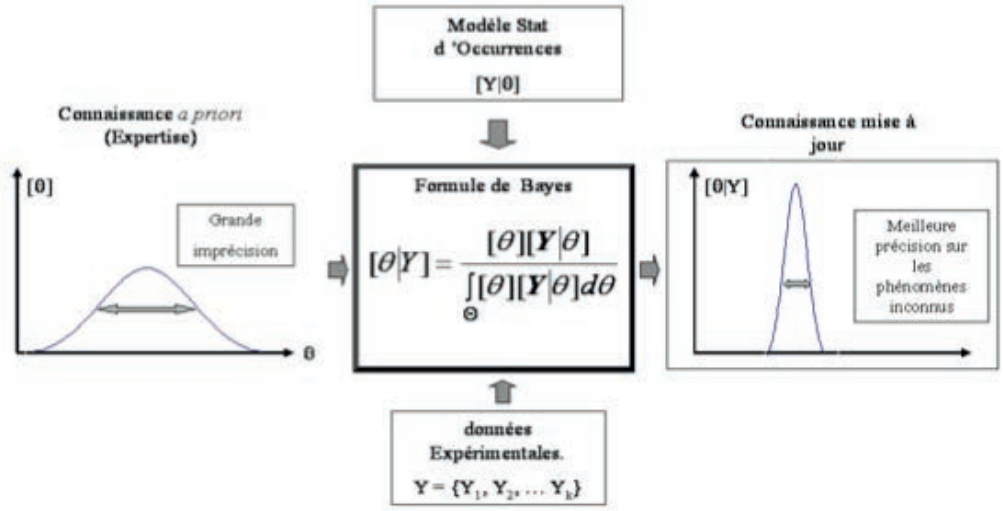

Figure 2.1 L'approche bayésienne (source : Boreux et al. 2009b)

Comme nous l'avons dit précédemment, une fois les équations du modèle établies,

il nous faut associer une distribution *a priori* à chacun des paramètres, ces derniers étant considérés comme des variables aléatoires. Ces distributions *a priori* doivent être établies, indépendamment de notre jeu de données, à partir des connaissances passées et de l'expertise des spécialistes du domaine d'application. Cet aspect de l'approche bayésienne a largement été critiqué car on inclut dans l'estimation une croyance qui peut être extrêmement subjective. D'autre part, il est courant que les experts n'aient pas suffisamment de recul ou d'informations pour orienter notre choix de lois *a priori*. Dans ce cas, nous sommes obligés d'utiliser des lois dites non-informatives (par exemple une loi uniforme ou une loi normale avec un écart-type très important). Cependant, l'introduction d'une connaissance passée peut également être perçue comme un atout car il s'agit d'une source d'information supplémentaire. Robertson et al. (1999) en donnent un exemple : si une série de caractéristiques portant sur les cernes d'un chêne est utilisée comme proxy climatique pour reconstruire les températures moyennes de juillet-août, on peut supposer *a priori* que ces températures ne peuvent pas dépasser l'intervalle 0-50°C car les chênes ne peuvent pas pousser au delà de cette limite.

L'étape suivante est le conditionnement des distributions définies *a priori* par les observations. On cherche à mettre à jour nos connaissances *a priori* à l'aide de nos données afin d'obtenir de nouvelles distributions, appelées distributions *a posteriori*. Ainsi, ces distributions *a posteriori* sont simplement les distributions de nos paramètres sachant nos données et on peut les calculer à l'aide de la règle de Bayes. On pose $\boldsymbol{\theta}$ l'ensemble des paramètres de notre modèle et $\boldsymbol{y}$ nos données. La distribution des paramètres sachant les données, notée $[\boldsymbol{\theta}|\boldsymbol{y}]$, est alors :

$$[\boldsymbol{\theta}|\boldsymbol{y}] = \frac{[\boldsymbol{y}|\boldsymbol{\theta}][\boldsymbol{\theta}]}{\int [\boldsymbol{y}|\boldsymbol{\theta}][\boldsymbol{\theta}] d\boldsymbol{\theta}}. \tag{2.1}$$

La distribution ainsi obtenue est la distribution à partir de laquelle on fait l'inférence, et notamment l'estimation ponctuelle des paramètres du modèle. Ici, nous voyons clairement que l'approche bayésienne permet la quantification de l'incertitude sur les paramètres du modèle, notamment au travers de l'expertise (c'est-à-dire de la

loi *a priori*) et des données. Ceci peut sembler très intéressant dans le cadre de reconstructions climatiques car on sait qu'il y a de nombreuses sources d'erreurs tout au long du traitement des données (erreurs de mesures, erreurs liées à la modélisation, etc...) et l'une des difficultés principales est donc de modéliser de manière fiable et réaliste les incertitudes avec lesquelles les variables climatiques sont reconstruites.

Cependant, l'obtention des distributions *a posteriori* des paramètres peut être compliquée. En effet, si l'on revient à l'équation (2.1), une fois que l'on a multiplié la vraisemblance $[\boldsymbol{y}|\boldsymbol{\theta}]$ par la distribution *a priori* $[\boldsymbol{\theta}]$, il arrive assez régulièrement que l'on ne puisse pas intégrer ce produit ni utiliser d'approximation asymptotique. On peut alors utiliser des méthodes statistiques que l'on regroupe sous le nom de méthodes de Monte Carlo par Chaîne de Markov (MCMC). Ces dernières n'ont pu être développées qu'avec les progrès de l'informatique, ce qui explique l'accroissement récent de l'utilisation des méthodes bayésiennes. Parmi les techniques MCMC les plus utilisées, on peut citer l'échantillonnage de Gibbs ou l'algorithme de Metropolis-Hastings (Gelman et al., 2003).

Une fois nos distributions *a posteriori* obtenues, reste bien sûr l'étape classique de validation du modèle. On cherche a voir si notre modèle ajuste bien les données, si les conclusions que l'on obtient sont raisonnables et comment les résultats sont sensibles aux hypothèses de modélisation faites dans la première étape. Si les réponses à ces questions ne sont pas celles que l'on attend, il faut recommencer et proposer un nouveau modèle.

2.2 Etude introductive : extraction d'un signal commun haute fréquence à partir de cernes d'arbres

Afin de mieux comprendre ce qu'est un modèle bayésien hiérarchique, nous proposons un exemple concret. Il s'agit d'un travail effectué en collaboration avec Bo-

reux et al. (2009a) qui proposent une méthode d'extraction d'un signal commun haute fréquence pour des séries de cernes d'arbres, basée sur un modèle bayésien hiérarchique. Cette étude a permis la publication d'un article (Annexe A) dans Climat of the Past. Si on se replace dans le contexte de la dendroclimatologie, nous avons dit en introduction que la première étape à l'interprétation climatique de la croissance des arbres était d'extraire un signal caché commun à tous les arbres d'une même zone géographique, ce dernier étant supposé représenter le climat de la région. De nombreuses méthodes pour résoudre ce problème existent et nous allons détailler l'une d'entre elle qui s'appuie sur le choix de l'approche bayésienne.

Le jeu de données utilisé pour cet exemple est composé de quinze séries, correspondant à quinze arbres, d'aires de cernes sur une période de 158 ans. Il s'agit de cernes d'Epinettes noires qui est une espèce très répandue dans le nord du Québec. Nous avons fait le choix d'utiliser des aires de cernes plutôt que des largeurs car ceci diminue l'impact de l'effet géométrique : le diamètre de l'arbre augmentant, plus les arbres sont vieux plus ils ont des cernes fins. Les arbres ont été échantillonnés sur un site proche du lac Hurault (54°15'N, 70°47'W) dans le nord du Québec, localisé sur la Figure 2.2 à l'aide de l'étoile rouge et nommé HM-1. Ce site a été choisi car il a l'avantage d'appartenir à une région climatique homogène relativement préservée de l'activité humaine. D'autre part, il s'agit d'un site intéressant tout particulièrement pour Hydro-Québec qui a grandement participé à ce projet du fait de ses capacités hydro-électriques.

Afin de nous faire une idée de notre jeu de données, les graphiques de droite de la figure 2.3 représentent le comportement temporel de trois séries d'aires de cernes, choisies au hasard à partir des 15 arbres. A la vue de ces courbes, il semble clair que chacun de ces arbres a une tendance différente et donc qu'il est difficile de trouver un signal commun dans le domaine de la basse fréquence. Afin de contourner cette difficulté, on propose d'appliquer une transformation simple, ce qui nous permettra de faire disparaître les tendances de chaque série et ainsi de travailler à partir de séries

Figure 2.2 Le graphique du haut représente le nord du Québec. Le graphique du bas est un zoom sur la région de Caniapiscau et la croix rouge, appelée HM-1, représente le site sur lequel les quinze arbres ont été échantillonnés.

stationnaires. Cette transformation est définie de la manière suivante :

$$Y_{ts} = \log X_{ts} - \log X_{t-1s}, \tag{2.2}$$

où $t = 2, ..., T$, $s = 1, ...S$ et où X_{ts} représente l'aire de cerne annuelle produite durant l'année t par l'arbre s. Cette transformation est assez classique et très couramment utilisée en finances (Gencay et al., 2002). Les séries ainsi transformées sont représentées par les graphiques de gauche de la figure 2.3. On note que Y_{ts} étant définie comme la log-différence entre deux valeurs d'aires de cernes consécutives, t ne correspond plus à une année mais à l'augmentation d'une année. D'autre part, à chaque fois que cette différence est proche de zéro cela signifie que les aires de cernes de deux années consécutives sont proches et, si cette différence est importante, les aires de cernes sont très différentes. Il faut garder cela en mémoire lors de l'interprétation des résultats.

Comme nous l'avons indiqué dans la section précédente, concernant la mise en place du modèle bayésien hiérarchique permettant l'extraction d'un signal commun caché supposé climatique, il nous faut tout d'abord mettre en place les équations du modèle. Pour cela on suppose que la variable aléatoire Y_{ts}, définie par l'équation (2.2)

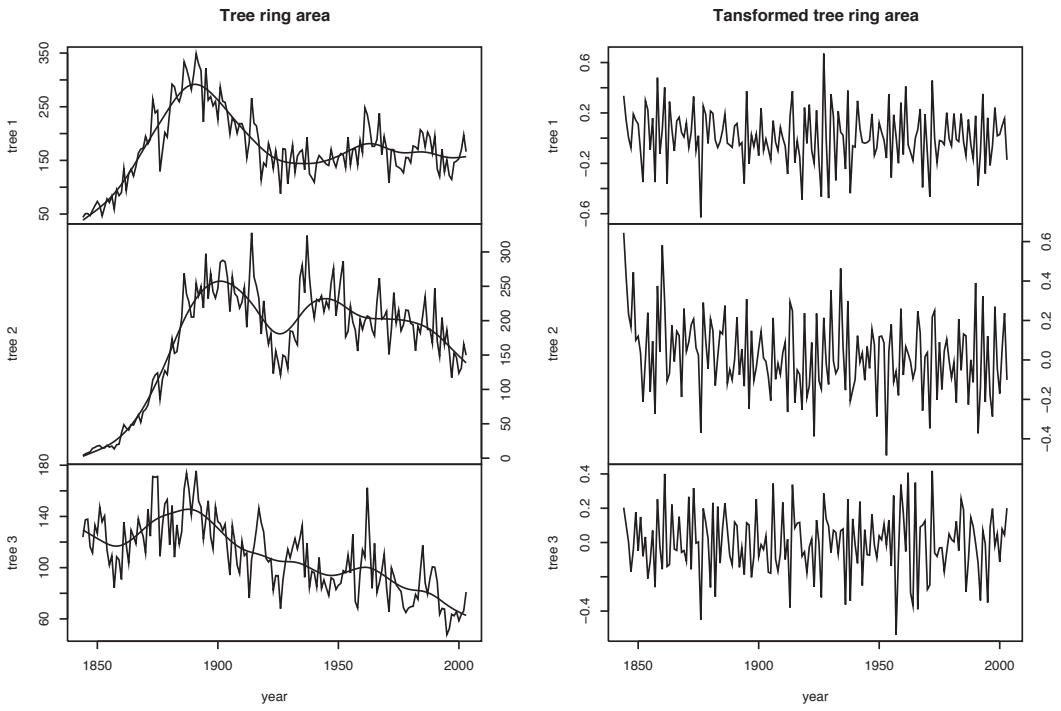

Figure 2.3 Comportement temporel de trois séries d'aires de cernes (choisies aléatoirement parmi les quinze arbres) sur la période 1846-2003. Les graphiques de gauche correspondent aux aires de cernes mesurées avec l'ajustement d'un spline cubic. Les graphiques de droite représentent la log-différence de ces mêmes aires de cernes, voir éq. (2.2).

suit un modèle additif comportant une variable latente Z_t

$$Y_{ts} = \mu_s + \lambda_s Z_t + \epsilon_{ts}, \tag{2.3}$$

avec $t = 2, ..., T$ et $s = 1, ..., S$, et ou μ_s correspond au niveau moyen de l'arbre s, Z_t représente le signal régional caché commun à tous les arbres et ϵ_{ts} décrit des fluctuations locales de l'arbre s durant l'année t. Les variations prises en compte par ϵ_{ts} peuvent être dues à des réserves accumulées par l'arbre s ou à des facteurs liés à la localisation de l'arbre (par exemple le fait qu'un cours d'eau passe ou non à son pied). Ces facteurs, bien qu'environnementaux et étant propres à chaque arbre, ne peuvent pas être pris en compte dans Z_t. Certains d'entre eux peuvent avoir une mémoire temporelle (par exemple le stress hydrique) ce qui nous a fait opter pour

une modélisation simple à l'aide d'un processus autorégressif gaussien d'ordre un et de moyenne nulle. Autrement dit, $\epsilon_{ts} = \phi_s \epsilon_{t-1s} + V_{ts}$ avec V_{ts} qui suit une distribution gaussienne de moyenne nulle et de précision η_s. Pour chaque année t, le produit $\lambda_s Z_t$ mesure la manière dont le facteur caché Z_t contribue à la croissance de l'arbre s. On suppose que Z_t et ϵ_{ts} sont des processus indépendants et comme pour ϵ_{ts} on autorise Z_t à avoir une mémoire d'une année sur l'autre. De même que précédemment, on suppose donc que le processus latent peut être modélisé par un processus autorégressif gaussien d'ordre un et de moyenne nulle, i.e. $Z_t = \rho Z_{t-1} + U_t$ avec U_t qui suit une distribution gaussienne de moyenne nulle et de précision τ. Notre modèle total compte $2 + 4S$ paramètres que l'on note (ρ, τ) et $\boldsymbol{\theta}_s = (\lambda_s, \mu_s, \phi_s, \eta_s)$ avec $s = 1, ..., S$.

Si l'on en revient à ce que l'on disait précédemment, à savoir qu'un modèle bayésien hiérarchique se décline en plusieurs couches, les variables aléatoires Y_{ts} correspondent à la couche des données et Z_t représente la couche du processus.

Les équations du modèle étant établies, il nous reste à définir les probabilités *a priori* que l'on met sur nos paramètres. On suppose que les distributions *a priori* $[\rho, \tau]$, $[\boldsymbol{\theta}_1]$,... et $[\boldsymbol{\theta}_S]$ sont indépendantes les unes des autres. Tout d'abord, intéressons-nous à la distribution *a priori* de $[\rho, \tau]$. En écrivant cette distribution jointe comme le produit d'une distribution conditionnelle et d'une distribution marginale, elle peut prendre la forme $[\rho, \tau] = [\rho|\tau][\tau]$. De manière générale, on suppose que le paramètre de précision τ suit une distribution Gamma avec deux hyperparamètres qui doivent être fixés afin de refléter au mieux la croissance *a priori*. En revanche, le choix de la loi *a priori* pour le coefficient auto-régressif $[\rho|\tau]$ est plus délicat. En général, on suppose que les processus auto-régressifs sont stationnaires, ce qui implique que le coefficient auto-régressif doit appartenir à l'intervalle $[-1, 1]$. Cependant, la philosophie bayésienne veut que les caractéristiques sous-jacentes du processus caché Z_t ne soient pas imposées mais résultent des données via la règle de Bayes ou via la connaissance a priori. On suppose donc que $[\rho|\tau]$ suit une distribution gaussienne de moyenne nulle et de précision proportionnelle à τ.

Concernant la loi *a priori* du vecteur aléatoire $\boldsymbol{\theta}_s = (\lambda_s, \mu_s, \phi_s, \eta_s)$, on sup-

pose l'indépendance conditionnelle, c'est-à-dire que $[\boldsymbol{\theta}_s] = [\lambda_s|\eta_s][\mu_s|\eta_s][\phi_s|\eta_s][\eta_s]$. De même que précédemment, on fait l'hypothèse que la précision η_s suit une distribution Gamma avec deux hyperparamètres fixés. Les distributions *a priori* $[\lambda_s|\eta_s]$, $[\mu_s|\eta_s]$ et $[\phi_s|\eta_s]$ sont supposées être gaussiennes. Ainsi, comme pour le coefficient d'auto-régression de Z_t cela signifie que le coefficient ϕ_s d'auto-régression de ϵ_{ts} n'est pas supposé être *a priori* dans l'intervalle $[-1, 1]$.

Notre modèle de probabilité total étant défini, il nous faut l'actualiser à l'aide de nos données d'aires de cernes d'arbres. Afin d'obtenir les lois *a posteriori* du vecteur latent Z_t et des autres paramètres de notre modèle, on utilise l'échantillonnage de Gibbs décrit dans l'annexe A. L'inférence bayésienne est faite à l'aide du logiciel libre R.

Ainsi, nous pouvons obtenir la distribution a posteriori de notre variable latente Z_t. Dans la figure 2.4, la courbe noire représente la valeur de la médiane *a posteriori* du facteur commun Z_t sur la période 1846-2003. L'aire grisée correspond à l'intervalle de crédibilité de 90%. On note que les valeurs de Z_t et de λ_s sont estimées à une constante près car il est toujours possible de multiplier Z_t par une constante et de diviser λ_s par la même constante sans être capable d'identifier ce facteur multiplicatif dans (2.3). La figure 2.4 compare également les résultats obtenus avec notre modèle bayésien hiérarchique aux résultats que l'on peut obtenir avec les techniques classiques employées par les dendrochronologues. En effet, la courbe en pointillé représente ce qu'on appelle l'indice de croissance des arbres. Il s'agit d'une moyenne arithmétique de ratios calculés sur tous les arbres. Chacun des ratios est obtenu en divisant l'air du cerne par une courbe lissée de la série temporelle des aires de cernes d'un arbre donné (Cook and Kairikukstis, 1990). On peut voir, qu'à une constante près (ce qui explique les deux échelles différentes de l'axe des ordonnées), l'indice de croissance des arbres se comporte de la même manière que Z_t en restant dans l'intervalle de crédibilité de ce dernier sur une longue période de temps. Une divergence, mais qui reste très localisée dans le temps, existe entre les deux courbes de 1875 à 1900, où l'indice de croissance prend des valeurs supérieures à l'intervalle de crédibilité à 90%

que nous avons obtenu. On peut donc en conclure que notre approche à l'aide d'un modèle bayésien hiérarchique nous permet d'obtenir des sorties significatives pour les dendrochronologues car elles ne contredisent pas leurs résultats passés. Elle permet donc d'apporter de l'information supplémentaire, avec notamment la modélisation des incertitudes, et d'offrir une nouvelle approche statistique à la communauté des dendrochronologues.

Afin de tester la qualité de nos estimations, la figure 2.5 nous montre, pour les arbres sélectionnés dans la figure 2.3, les Y_{ts} observés (c'est-à-dire les aires de cernes mesurées) en fonction d'une estimation naïve $\hat{Y}_{ts}$ obtenue en additionnant les médianes *a posteriori* des paramètres de (2.3) sans le bruit. Comme on pouvait s'y attendre, la relation apparait comme linéaire. On trouve le même type de résultats pour les autres arbres.

Cependant, si ce travail nous a paru extrêmement intéressant puisqu'il a permis de mettre en évidence l'intérêt des modèles bayésiens hiérarchiques en dendrochronologie et notamment dans le cadre d'extraction de signaux cachés, nous lui avons trouvé quelques limites. Tout d'abord, dans le modèle (2.3) la croissance propre de l'arbre est supposée être constante alors que, dans le modèle d'agrégation linéaire (1.1) utilisé par les dendrochronologues, elle évolue dans le temps. On peut donc se demander s'il ne serait pas intéressant de remplacer le paramètre μ_s par une fonction évoluant avec le temps. D'autre part, comme nous l'avons dit les données de mesures de cernes d'arbres sont standardisées avant d'être traitées. On peut donc faire à cette nouvelle méthode la même critique qu'aux méthodes classiques : on perd toute l'information climatique basse fréquence qui peut être contenue dans les cernes d'arbres. Forts de ces considérations, nous avons repris le modèle (2.3) et cherché à l'améliorer afin qu'il réponde au mieux aux attentes de la communauté des dendrochronologues.

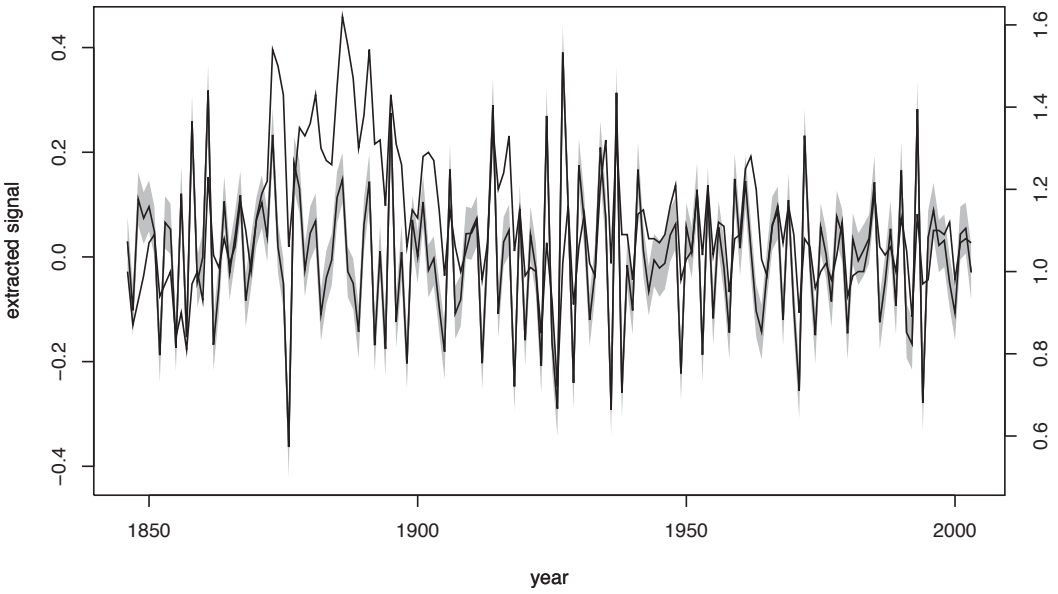

Figure 2.4 Médiane *a posteriori* du signal Z_t sur la période 1846-2003 et son intervalle de crédibilité à 90% en grisé. La courbe en pointillé représente ce que l'on appelle l'indice de croissance des arbres.

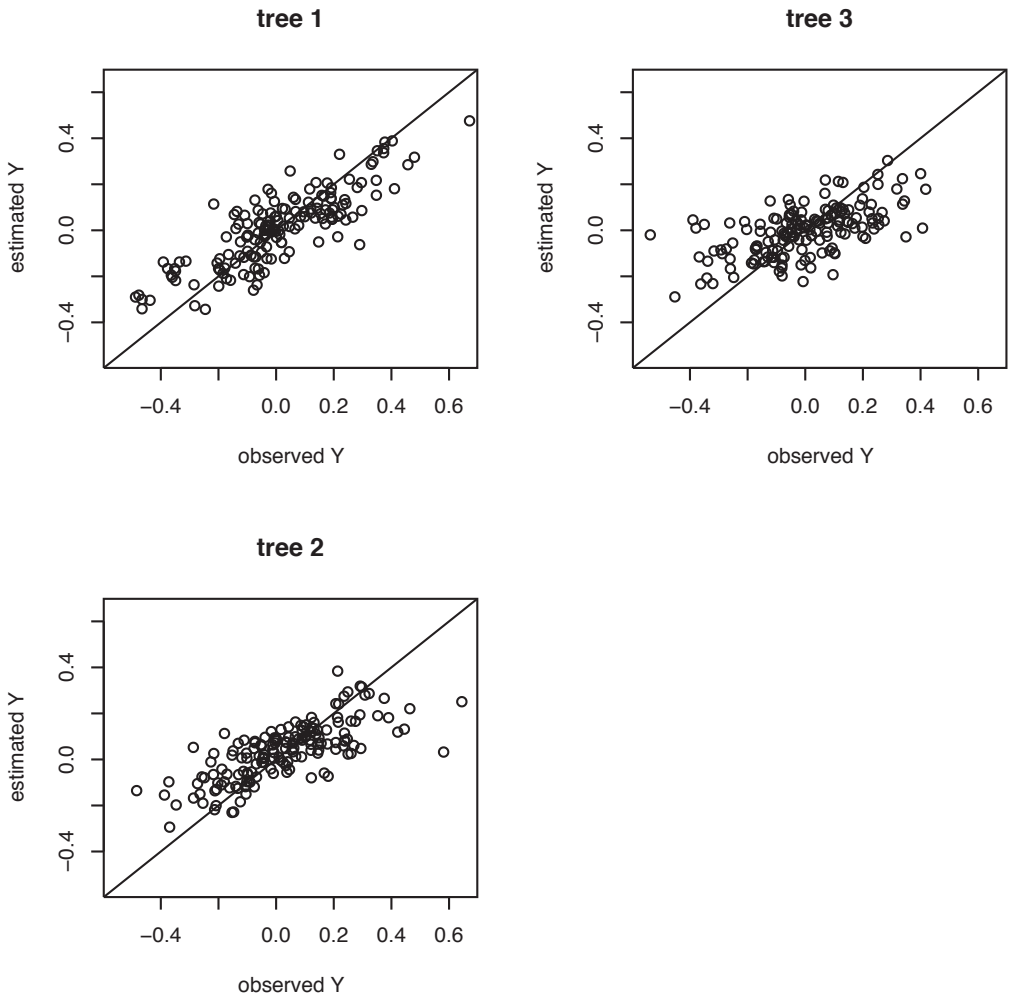

Figure 2.5 Les Y_{ts} observés en fonction de leur estimation pour chacun des trois arbres choisis aléatoirement dans la Figure 2.3

Bibliographie

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

J.-J. Boreux, P. Naveau, O. Guin, L. Perreault, and J. Bernier. 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, 2009a. doi : 10.5194/cp-5-607-2009.

J.-J. Boreux, E. Parent, and J. Bernier. *Pratique du calcul bayésien*. Springer, 2009b.

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.

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

D. Cooley, D. Nychka, and P. Naveau. Bayesian spatial modeling of extreme precipitation return levels. *J. Am. Stat. Assoc.*, 102(479) :824–840, 2007.

N. Eckert, H. Baya, and M. Deschâtres. Assessing the response of snow avalanche runout altitudes to climate fluctuations using hierarchical modelling : application to 61 winters of data in France. *Journal of Climate*, 2010.

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

R. Gencay, F. Selcuk, and B. Whitcher. *An Introduction to Wavelets and Other Filtering Methods in Finance and Eco- nomics*. Academic Press, San Diego, 2002.

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

A. Korhola, K. Vasko, H. Toivonen, and H. Olandor. Holocene temperature changes in northern fenno-scandia reconstructed from chironomids using bayesian modelling. *Quatern. Sci. Rev.*, 21 :1841–1860, 2002.

I. Robertson, D. Lucy, L. Baxter, A.M. Pollard, R.G. Aykroyd, A.C. Barker, A.H.C. Carter, V.R. Switsur, and J.S. Waterhouse. A kernel-based bayesian approach to climatic reconstruction. *The Holocene*, 9(4) :495–500, 1999.

H.T.T. Toivonen, H. Mannila, A. Korhola, and H. Olander. Applying bayesian statistics to organism-based environmental reconstruction. *Ecol. Appl.*, 11 :618–630, 2001.

K. Vasko, H. T. Toivonen, and A. Korhola. A bayesian multinomial gaussian response model for organism-based environmental reconstruction. *J. Paleolimn.*, 24 :243–250, 2000.

Chapitre 3

Extraction de tendances cachées dans les cernes d'arbres à l'aide d'un modèle bayésien hiérarchique

Dans ce chapitre nous proposons une nouvelle méthode permettant l'extraction d'un signal commun à partir de plusieurs séries chronologiques de mesures de cernes d'arbres. Cette méthode est basée sur un modèle bayésien hiérarchique semi-paramétrique permettant de capturer les basses fréquences contenues dans les cernes d'arbres, fréquences difficiles à capturer avec les méthodes classiques de la dendrochronologie. Dans une première partie, nous décrirons le modèle utilisé et dans une seconde partie nous comparerons nos résultats à ceux obtenus avec une méthode plus classique.

3.1 Extraction de tendances cachées dans les cernes d'arbres à l'aide d'un modèle bayésien hiérarchique semi-paramétrique

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

Résumé : La statistique est devenue une composante essentielle des reconstructions climatiques, qui sont elles même 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, mise en place dans le but d'extraire un signal climatique pertinent pour chaque proxies. Ce papier s'intéresse tout particulièrement à cette étape pour l'un des proxies 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 vue statistique, ce problème d'extraction peut être vu comme la recherche d'une variable cachée qui représente un signal commun à un ensemble 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 activ-

ity 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 ad-

dition, 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

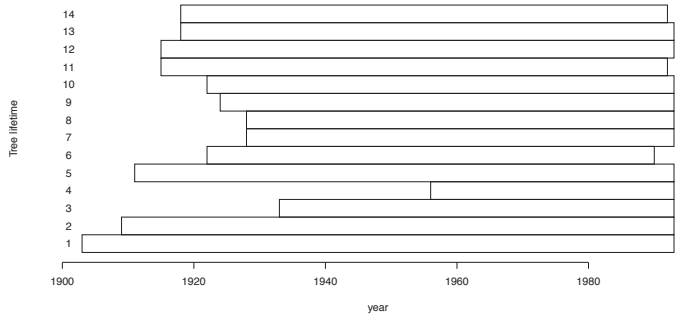

Figure 1: The lifetime of the fourteen trees that has been used in our application. The x-axis corresponds to the years and the y-axis to the tree label.

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.

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 re-

gion 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.

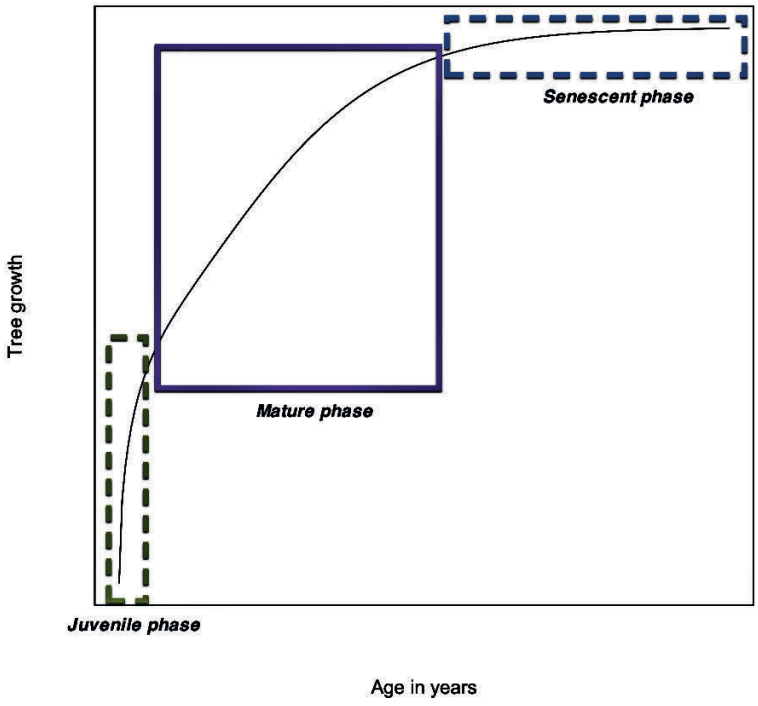

Figure 2: Idealized tree age effect behavior over time. After a juvenile phase (youth) with an accelerating rate of growth, the tree enters a mature phase with a roughly constant rate of growth, follows by a senescent phase with a decelerating rate of growth. In practice it is difficult to statistically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

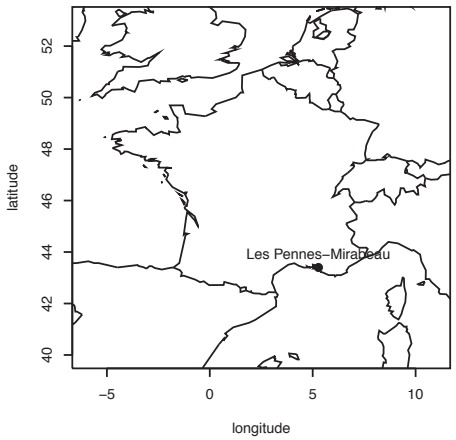

Figure 3: The "Les Pennes-Mirabeau" site located in the South of France where *Pinus halepensis Mill* tree ring densities series shown in Figure 4 were recorded.

(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

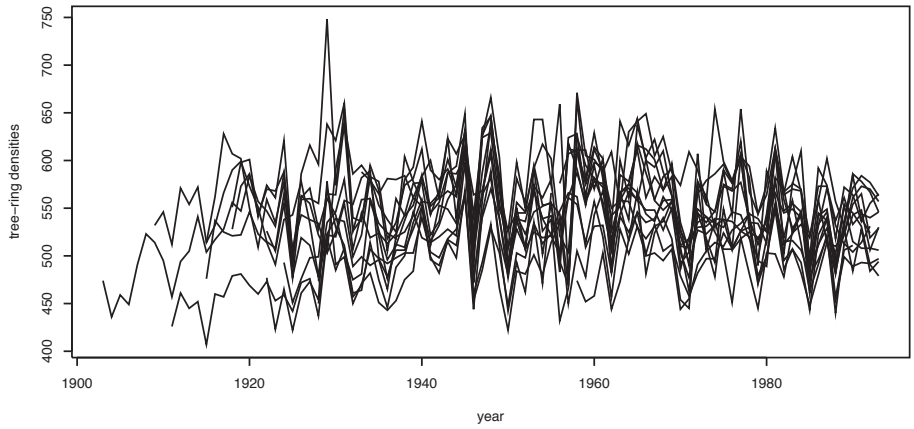

Figure 4: Fourteen *Pinus halepensis Mill* tree ring density time series (in mg/cm3) from the "Les Pennes-Mirabeau" site located in Figure 3. The x-axis (years) covers the period $1903 - 1993$ and each time series has a different length, see Figure 1.

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 $\int (f''(\mathbf{x}))^{\mathbf{2}}\mathbf{dx} = \mathbf{f^T K f}$. 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 $\mathbf{f}$

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

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

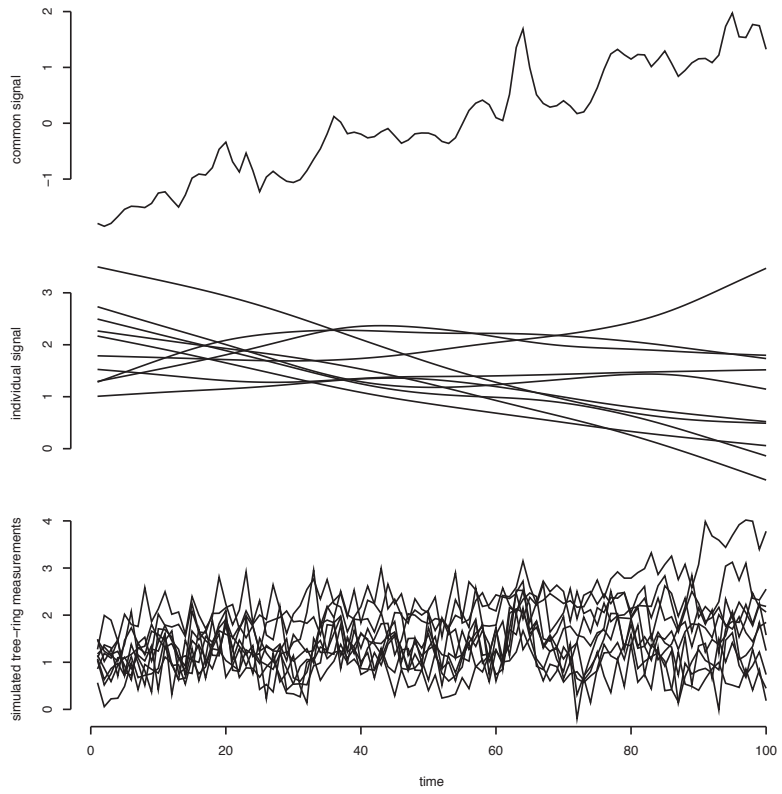

Figure 5: Simulations of tree ring measurements from the additive model (2). The top panel corresponds to the common signal $\mathbf{f}$, the second panel to individual growth tree effect signals $\mathbf{g}_j$ and the bottom panel to simulated tree ring series $\mathbf{y}_j$, respectively. Our objective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

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 $[0, 1]$

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

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

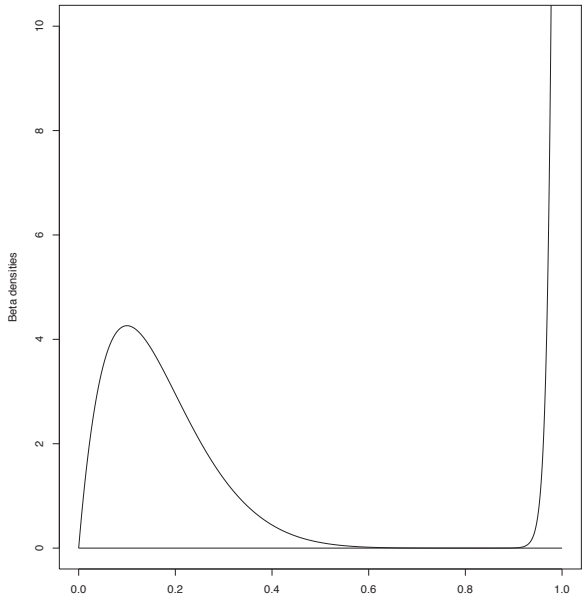

Figure 6: Informative Beta prior densities for the smoothness parameter $\varphi_0 \sim \text{Beta}(2, 10)$ (solid line) of the function $\mathbf{f}$ and for $\varphi_j \sim \text{Beta}(100, 1)$ (dotted line) of the function $\mathbf{g}_j$ for $j = 1, \ldots, p$. A value near one (near zero) corresponds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

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.

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 surpris-

ing 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).

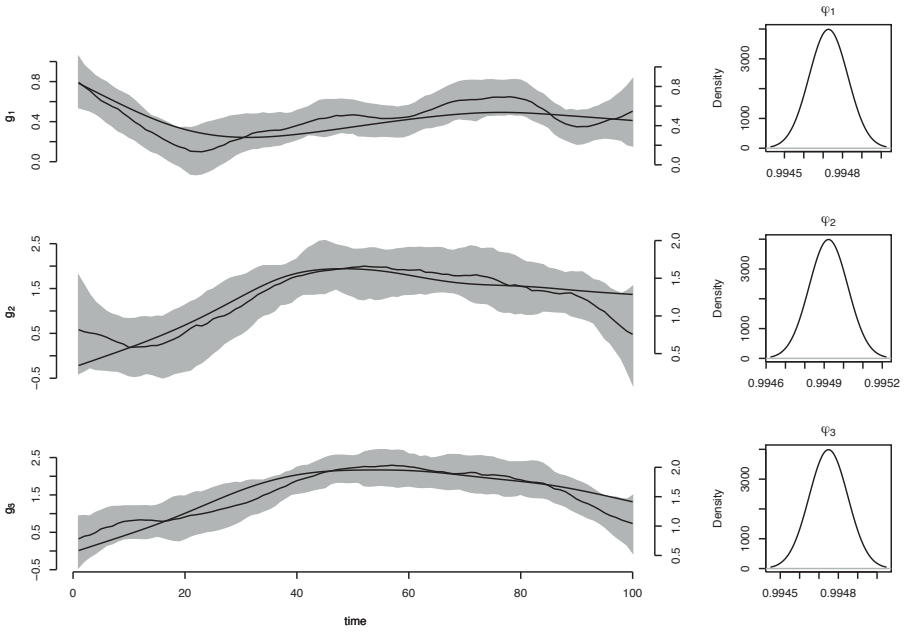

Figure 7: Posterior information about the tree age effect $\mathbf{g}_j$ for $j = 1, 2, 3$ obtained from the simulated tree series shown in the bottom panel of Figure 5. The solid and dotted lines in the left panels correspond to the true $\mathbf{g}_j$ and the estimated posterior median, respectively. 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.

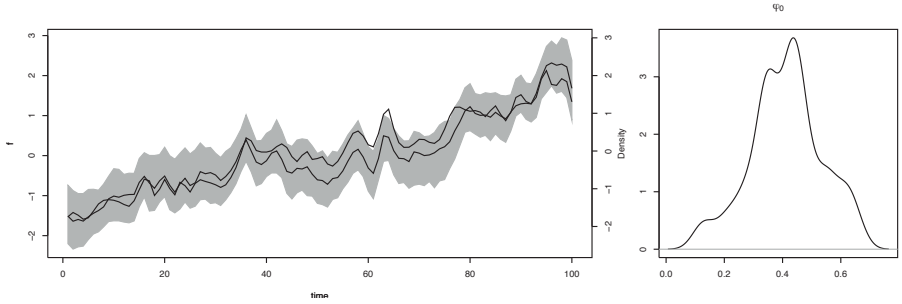

Figure 8: Left panel: posterior median and 95% credibility intervals of the common signal shared by all trees $\mathbf{f}$ obtained from the simulated tree series shown in the bottom panel of Figure 5. The solid corresponds to the true $\mathbf{f}$. Right panel: posterior pdf of the smoothness parameter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

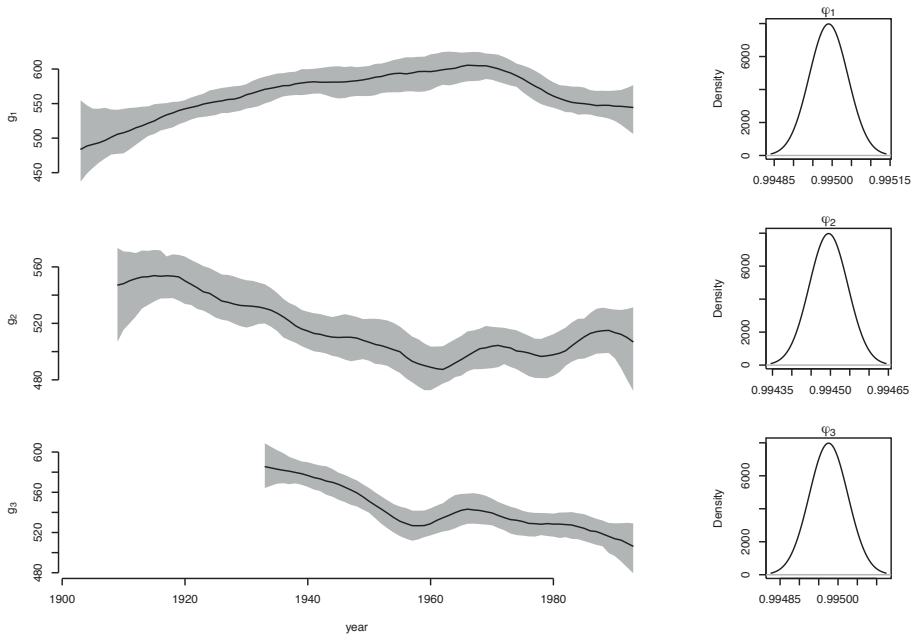

Figure 9: Left panels: posteriors of the three individual age effect trends $\mathbf{g}_1$, $\mathbf{g}_2$ and $\mathbf{g}_3$ obtained from our analysis of the fourteen tree density series shown in Figure 4. Black lines correspond to posterior medians and gray areas to 95% credibility intervals. Right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

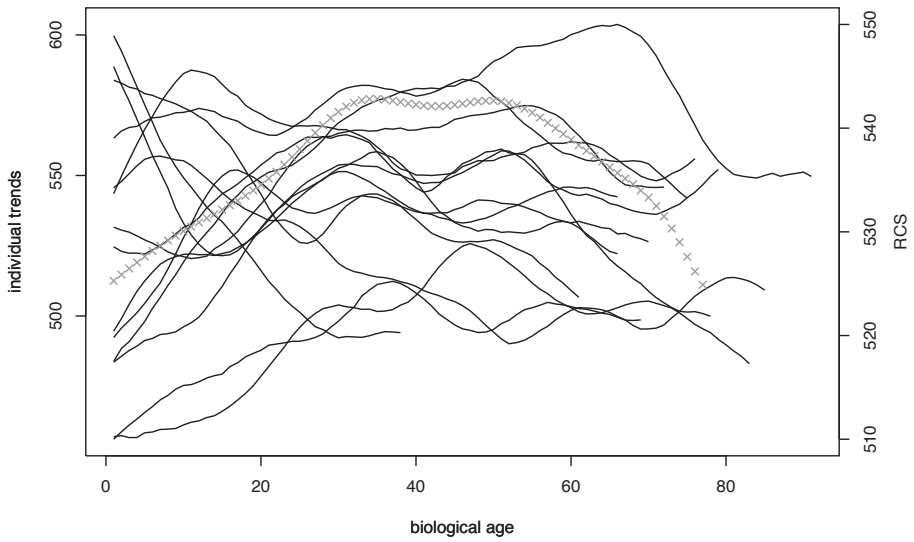

Figure 10: Posterior median of individual age effect profiles $\mathbf{g}_j$ that have been aligned according to their biological age (not their chronological age). The line thickness is proportional to the posterior median coefficient β_j. The gray line represents the classical global biological trend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

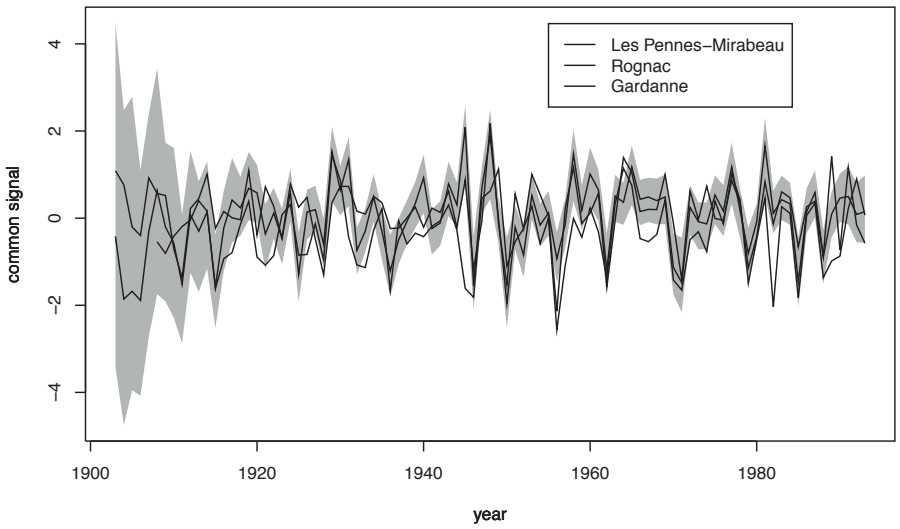

Figure 11: Posterior median of the common signal $\mathbf{f}$ obtained from trees measured at the site of “Les Pennes Mirabeau" (solid line), the site of “Rognac" (dashed line) and the site of “Gardanne" (dotted line). The three sites belong to the same climatological region and have the same tree species. 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-

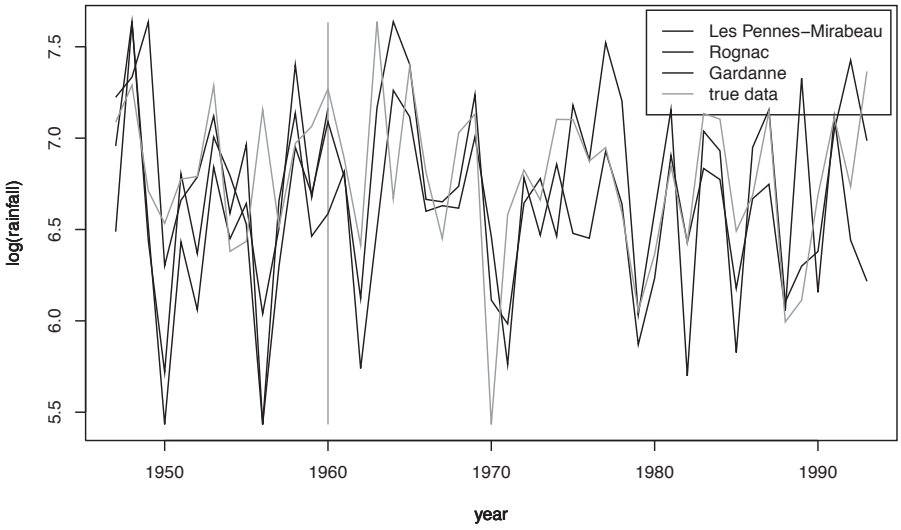

Figure 12: Rainfall reconstruction. The grey line represents the logarithm of observed total Summer precipitation recorded during 1947-1993 in Marseille, source ECA&D (`http://eca.knmi.nl/`). During the period 1961-1993, a linear estimation between log(rainfall) and the signal $\mathbf{f}$ was implemented 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

910

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.

3.2 Comparaison avec une méthode classique de la dendrochronologie : la méthode RCS

Comme cela a été dit dans le chapitre 1, les méthodes classiques d'extraction d'un signal climatique caché dans les cernes d'arbres sont regroupées sous le nom de standardisation et elles se déroulent en trois étapes : estimation d'une tendance reliée à l'âge des arbres, retrait de cette tendance pour chaque arbre et moyenne des chronologies obtenues. On a vu que l'une des difficultés majeures de la standardisation est que lorsqu'on enlève la tendance reliée à l'âge des arbres, on supprime en même temps toute l'information basse fréquence contenue dans les cernes. Afin de contourner cette difficulté les dendrochronologues ont mis en place une méthode appelée RCS (Regional Curve Standardization), permettant de préserver une partie de l'information climatique basse fréquence. Elle est basée sur la production d'une tendance de croissance globale biologique des arbres obtenue en faisant la moyenne des largeurs de cernes qui ont été alignées selon leur âge biologique (et non leur âge chronologique). La courbe obtenue est alors lissée à l'aide d'une fonction spline et chaque chronologie de cernes d'arbres est divisée par cette dernière. Comme dans toutes les méthodes classiques de standardisation, le signal commun est estimé en faisant la moyenne des chronologies obtenues pour chaque arbre. La méthode RCS est aujourd'hui l'une des plus utilisée. La procédure d'extraction que nous avons décrite dans la section 3.1 ayant elle aussi pour but de conserver l'information basse fréquence contenue dans les cernes d'arbres, il nous paraît opportun de comparer nos résultats avec ceux que l'on obtiendrait à l'aide de cette dernière.

Pour cela nous utiliserons un jeu de données bien connu et utilisé par Büntgen et al. (2006) afin de reconstruire les variations de températures dans les Alpes entre 755 et 2004. Plus précisément, il s'agit d'un jeu de données comportant 180 séries de densités de cernes de mélèzes [*Larix decidua Mill.*]. On peut diviser ces séries en deux groupes : des séries récentes échantillonnées sur des arbres vivants (86 séries) et

des séries anciennes échantillonnées sur des arbres morts (94 séries). La figure 3.13 représente ces différentes séries de densités de cernes ainsi que la manière dont elles se répartissent dans le temps.

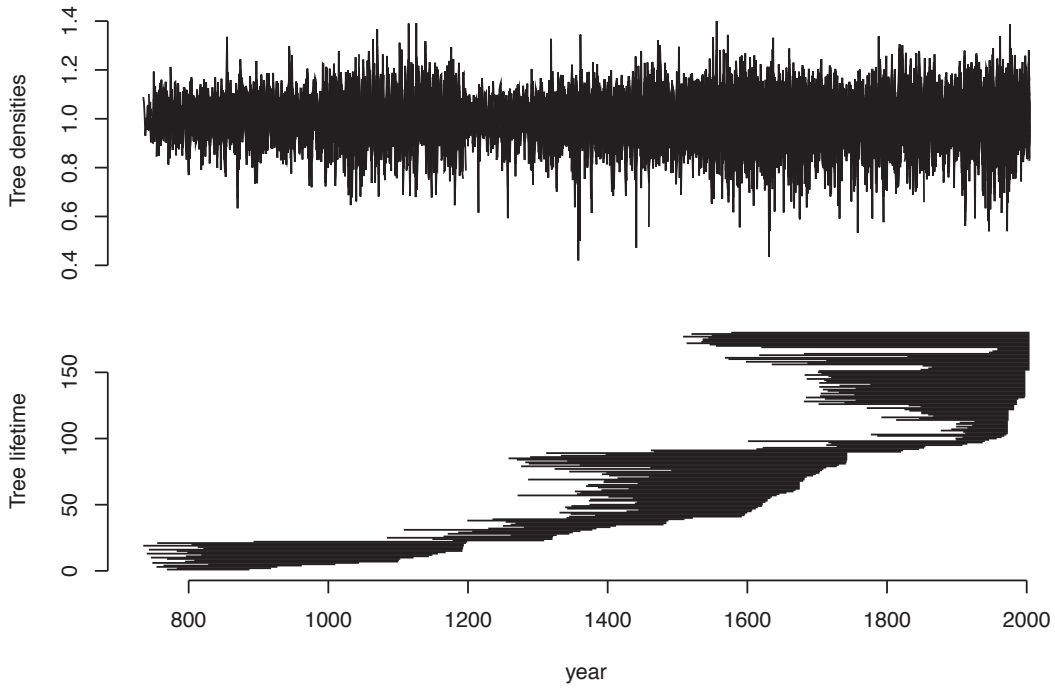

Figure 3.13 Représentation des séries de densités de cernes de Mélèzes dans les Alpes. Le graphique du haut représente les données de mesures de cernes et le graphique du bas la manière dont les différentes séries se répartissent dans le temps.

945 On applique notre nouveau modèle à ce jeu de données et on s'intéresse tout d'abord à l'estimation de la tendance de croissance globale biologique des arbres. Dans la méthode RCS les arbres sont tous supposés avoir la même forme structurelle qui ne varie pas dans le temps alors qu'avec notre méthode nous permettons à chaque arbre d'avoir sa propre tendance de croissance. La figure 3.14 compare la tendance 950 de croissance biologique estimée par la méthode RCS (courbe grise) et les médianes *a posteriori* des effets de l'âge que l'on a obtenues à l'aide de notre modèle bayésien hiérarchique semi-paramétrique et qui ont été alignées suivant l'âge biologique des

arbres (courbes noires). On peut noter la grande diversité des profils propres à chaque arbre obtenus à l'aide de notre modèle. Ceci nous laisse donc penser que la tendance de croissance n'est pas forcement la même pour tous les arbres, conclusion qui rejoint l'analyse de Fritts (1976) disant que tous les individus d'une espèce n'atteignent pas forcement leur optimum de croissance au même âge et que leur taux de croissance peut varier suivant certains facteurs du sol ou de la compétition existant entre les arbres. Avoir une information détaillée sur la croissance propre de chaque arbre peut donc sembler intéressante pour les dendrochronologues et les aider à interpréter le comportement local des arbres.

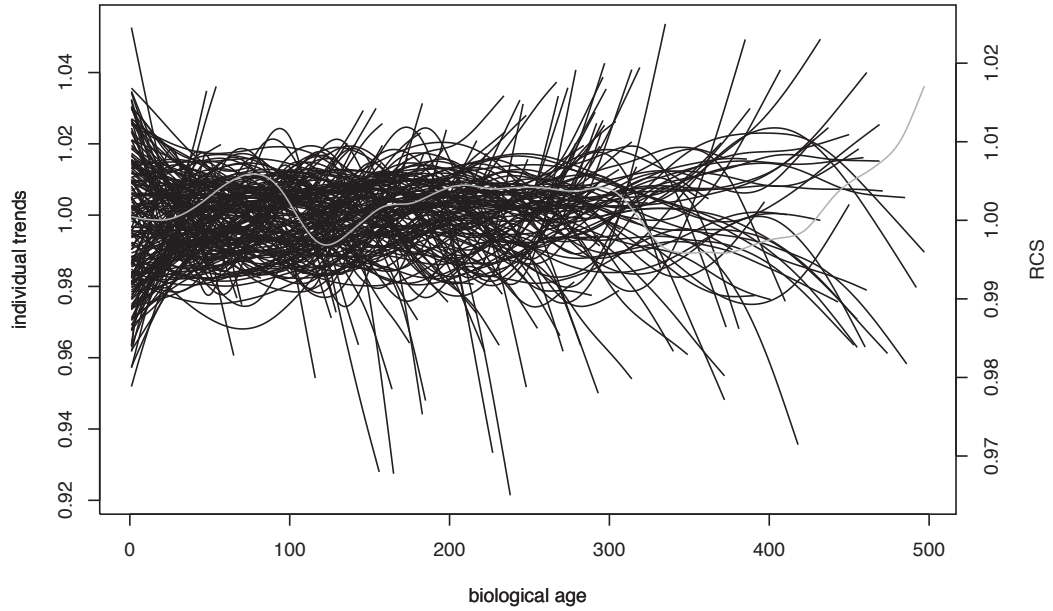

Figure 3.14 Médianes *a posteriori* des tendances de croissance propres à chaque arbre, alignées en fonction de leur âge biologique. La courbe grise représente la tendance biologique globale obtenue par la méthode RCS en fonction de l'âge des arbres.

On peut également remarquer sur la figure 3.14 que les courbes de croissances obtenues pour chaque arbre n'ont pas forcement le même lissage. En effet, le modèle décrit dans la section 3.1 n'impose pas de coefficient de lissage fixe. Certes, compte tenu de

965 l'expérience des dendrochronologues et pour des questions d'identifiabilité, nous avons mis un *a priori* fort sur ces derniers mais nous laissons l'échantillonnage de Gibbs estimer leur valeur pour chaque arbre. Ceci apporte une flexibilité supplémentaire par rapport à la méthode RCS où le lissage de la courbe globale de croissance est choisi de manière « arbitraire »ce qui peut entrainer un biais dans les résultats.

970 La figure 3.15 représente les signaux communs extraits avec chacune des méthodes étudiées. Sur le premier graphique nous avons représenté la médiane *a posteriori* obtenue à l'aide de notre modèle bayésien hiérarchique avec son intervalle de crédibilité à 95% en grisé. Le second graphique, quant à lui, correspond à l'estimation que nous donne la méthode RCS. Les deux chronologies extraites étant longues, il est difficile
975 de les comparer visuellement et de voir ce qu'elles ont ou non en commun. Nous avons donc représenté sur la figure 3.16 le signal extrait par la méthode RCS en fonction de notre médiane *a posteriori*. On voit que le nuage de points obtenu suit une droite, ce qui veut dire que la médiane *a posteriori* que nous avons extraite à l'aide de notre modèle est proche du signal estimé par la méthode RCS. Ainsi notre nou-
980 velle méthode d'extraction, décrite dans la section 3.1, nous permet de retrouver des résultats proches de ceux obtenus de manière classique par les dendrochronologues. L'un des principaux avantages d'extraire notre signal commun à l'aide d'un modèle bayésien réside dans le fait que l'on connaît l'incertitude de notre estimation (figure 3.15). Cette information supplémentaire nous paraît extrêmement importante dans
985 le sens où elle peut avoir un impact sur l'interprétation climatique d'un tel signal ou encore dans le cadre de reconstructions climatiques. Par exemple, on peut noter que l'incertitude portant sur le signal commun extrait est importante avant 800 et vers 1200. On peut expliquer ce phénomène en se reportant au second graphique de la figure 3.13 : il s'agit de deux périodes pour lesquelles on dispose de peu d'arbres ce
990 qui entraine assez naturellement une incertitude plus grande sur notre signal. Cette information n'apparaît pas avec la méthode RCS mais pourtant elle est tout aussi vraie et il faudrait prendre les résultats obtenus sur ces périodes avec prudence.

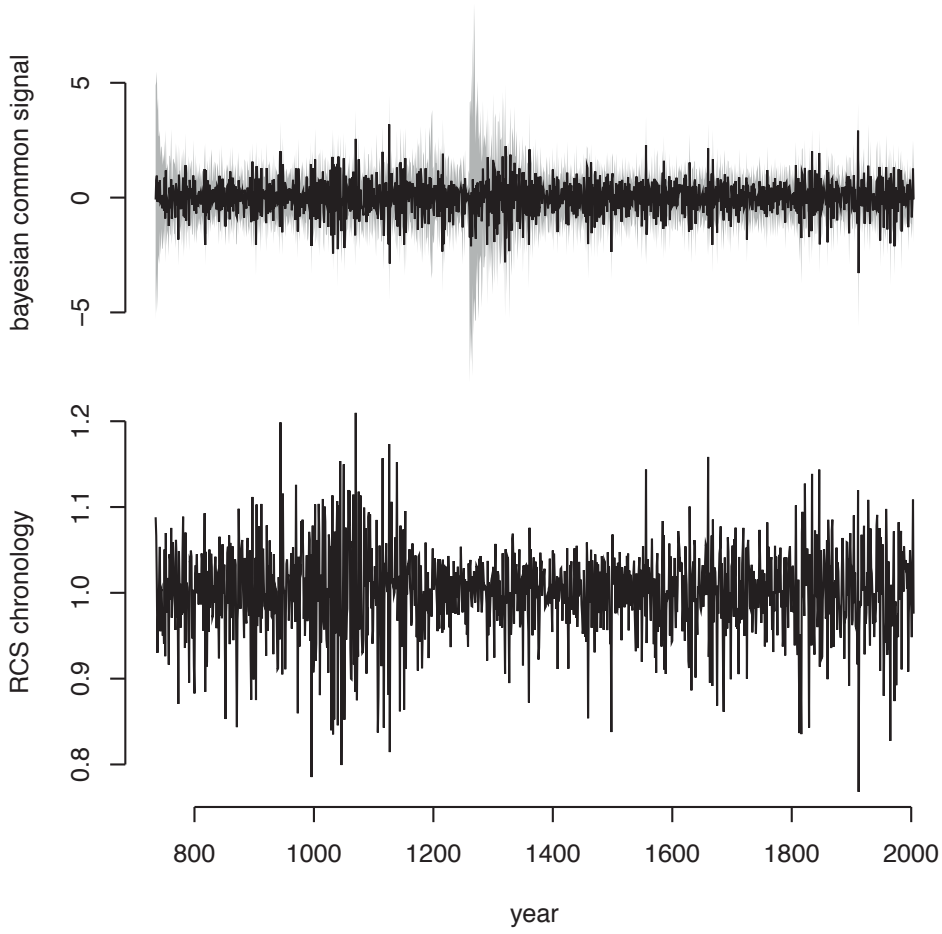

Figure 3.15 Représentation des signaux extraits avec notre méthode et la méthode RCS. Le premier graphique représente la médiane *a posteriori* obtenue à l'aide de notre modèle avec son intervalle de crédibilité à 95% et le second graphique représente l'estimation que nous donne la méthode RCS.

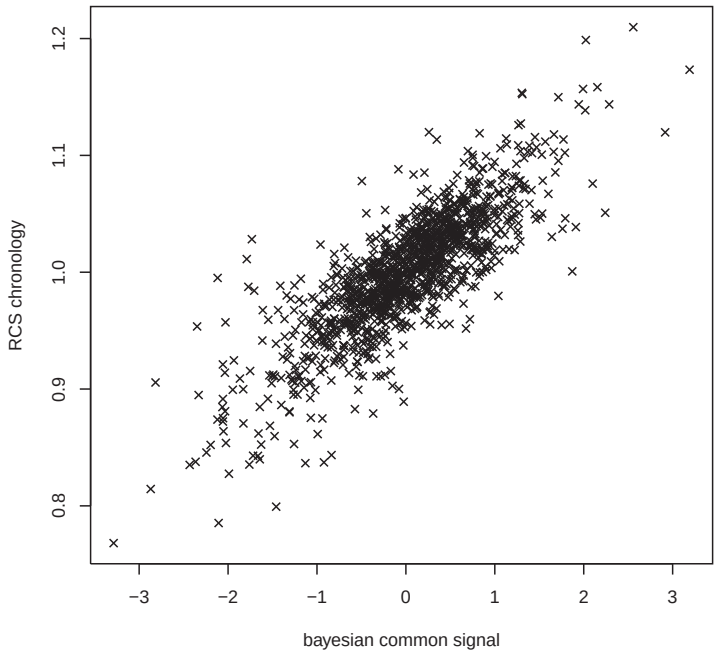

Figure 3.16 Comparaison signaux extraits avec notre méthode et la méthode RCS.

Nous nous intéressons maintenant aux résidus obtenus pour chacune des deux méthodes étudiées. Nous avons tracé un Q-Q plot des résidus de l'arbre 1 (figure 3.17). On note que les résidus obtenus avec notre modèle bayésien hiérarchique semblent gaussiens, ce qui est un peu moins vrai pour ceux obtenus avec la méthode RCS. Cette impression nous est confirmée par la figure 3.18 qui représente l'histogramme des résidus pour l'arbre 1 pour chacune des méthodes. L'histogramme du premier graphique, correspondant à notre nouvelle méthode, est très proche de la densité d'une loi normale. Nous avons le même type de résultats pour les autres arbres de notre jeu de données. Le bruit résiduel semble donc être plus compliqué lorsque l'on utilise la méthode RCS. Ceci peut s'expliquer simplement par le fait que certains arbres ont une tendance propre assez éloignée de la courbe de croissance estimée par la méthode RCS. Ainsi on retrouve cette information, qui n'a pas été prise en compte,

dans le bruit.

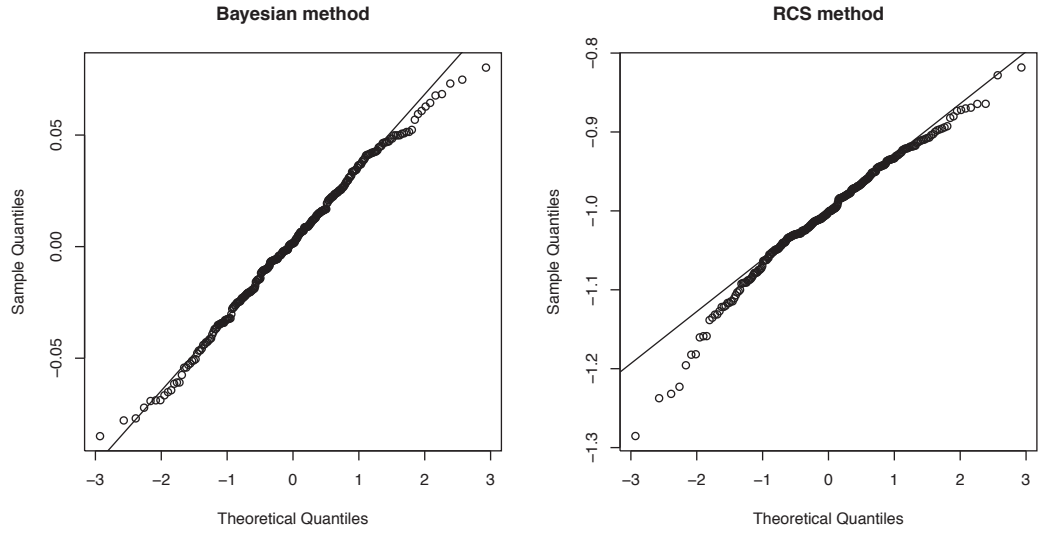

Figure 3.17 Q-Q plot des résidus obtenus avec notre nouvelle méthode et la méthode RCS pour l'arbre 1

Pour terminer, nous souhaiterions discuter de l'une des difficultés de la méthode RCS concernant l'extraction du signal commun. En effet, lorsqu'on se trouve dans une situation où tous les arbres sont quasiment nés en même temps, le fait d'aligner les séries de mesures de cernes en fonction de leur âge cambial ou de leur âge chronologique revient au même. Ainsi, lors de l'estimation de la tendance de croissance globale, on peut prendre en compte une partie de l'information climatique commune. Afin d'illustrer ce phénomène nous avons repris les données que nous avions simulées dans la section 3.1 et nous leur avons appliqué les deux méthodes d'extraction. Dans la figure 3.19, la courbe en traits discontinus correspond au signal commun que nous avions simulé et la courbe en pointillé à la chronologie obtenue à l'aide de la méthode RCS. On voit que la tendance basse fréquence contenue dans le signal commun simulé a disparu et a été prise en compte dans la courbe de croissance globale. Ce défaut semble absent pour notre nouvelle méthode, la courbe en trait plein de la figure 3.19 représentant la médiane *a posteriori* obtenue à l'aide de notre modèle bayésien hiérarchique, avec son intervalle de crédibilité à 95%. Cette estimation est très proche

du signal commun simulé qui est compris dans l'intervalle de crédibilité. Cependant, il faut quand même noter que dans certains cas très particuliers notre méthode a elle aussi ses limites. Si les tendances de croissances propres à chaque arbre sont très proches les unes des autres, notre modèle considérera qu'il s'agit de quelque chose de commun à toutes les séries, et ainsi cette information sera prise en compte dans le signal commun et non dans les tendances propres à chaque arbre (figure 3.20).

Pour conclure, nous avons mis en place une nouvelle méthode permettant l'extraction d'un signal commun, supposé climatique, à partir de séries de mesures de cernes d'arbres. Pour cela nous avons combiné deux types de modèles spline : un pour le signal commun caché et un pour l'effet de l'âge individuel de chaque arbre. Cette nouvelle approche semble présenter de nombreux avantages car en modélisant de manière individuelle les tendance de croissance des arbres, nous faisons moins d'hypothèses que pour la méthode RCS et ainsi nous obtenons une plus grande flexibilité dans notre modèle. D'autre part, le choix de l'estimation bayésienne nous permet d'estimer l'incertitude associée à la variable extraite, ce qui n'est pas le cas des méthodes classiques. Dans le cadre de reconstructions climatiques le fait de prendre en compte cette information supplémentaire peut avoir un impact sur les résultats.

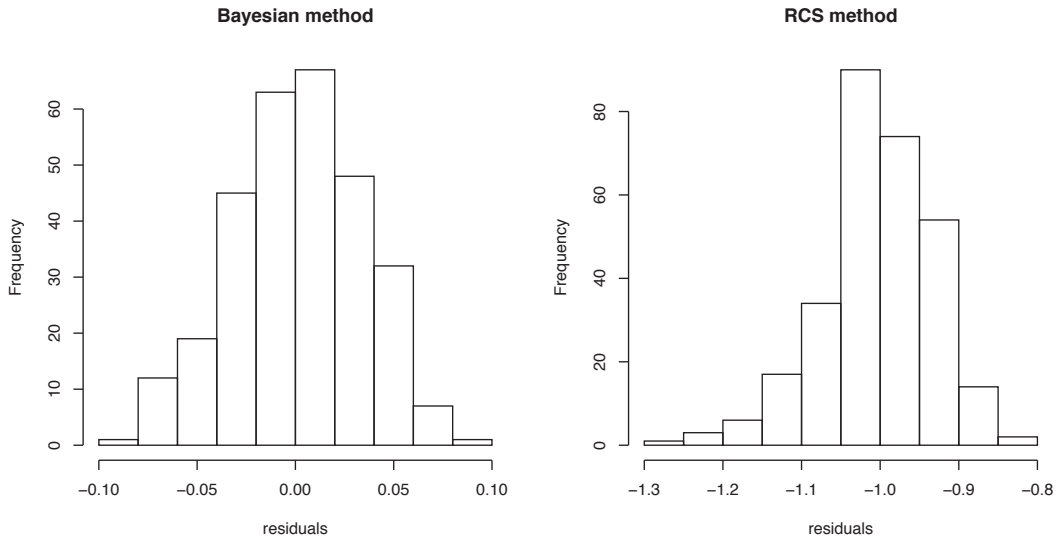

Figure 3.18 Histogramme des résidus obtenus avec notre nouvelle méthode et la méthode RCS pour l'arbre 1

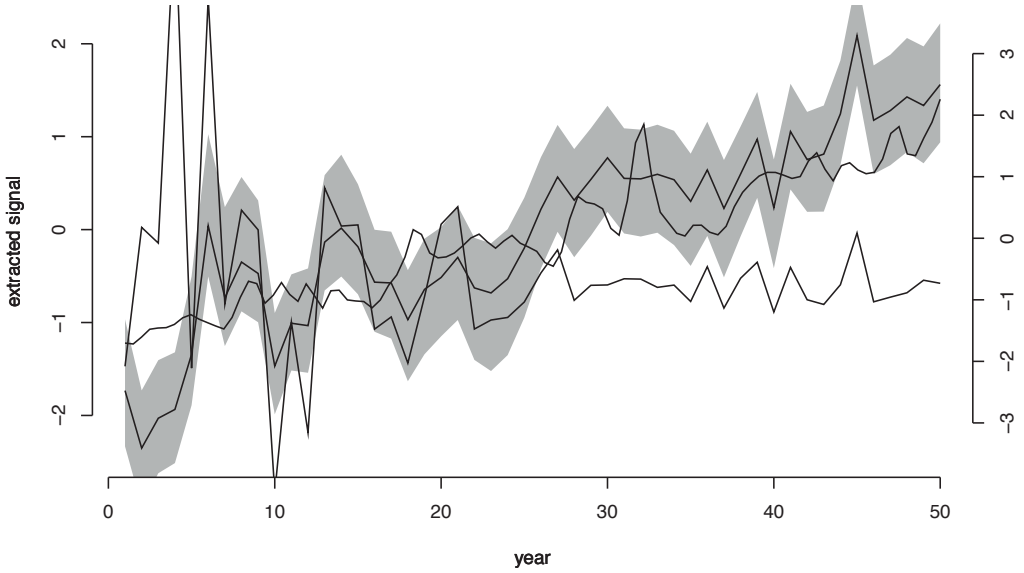

Figure 3.19 Signal commun extrait pour un jeu de données simulées. La courbe en traits interrompus correspond au signal commun simulé et la courbe en pointillé au signal commun extrait avec la méthode RCS. La courbe en trait plein représente la médiane *a posteriori* obtenue avec notre modèle bayésien hiérarchique avec son intervalle de confiance à 95% en grisé.

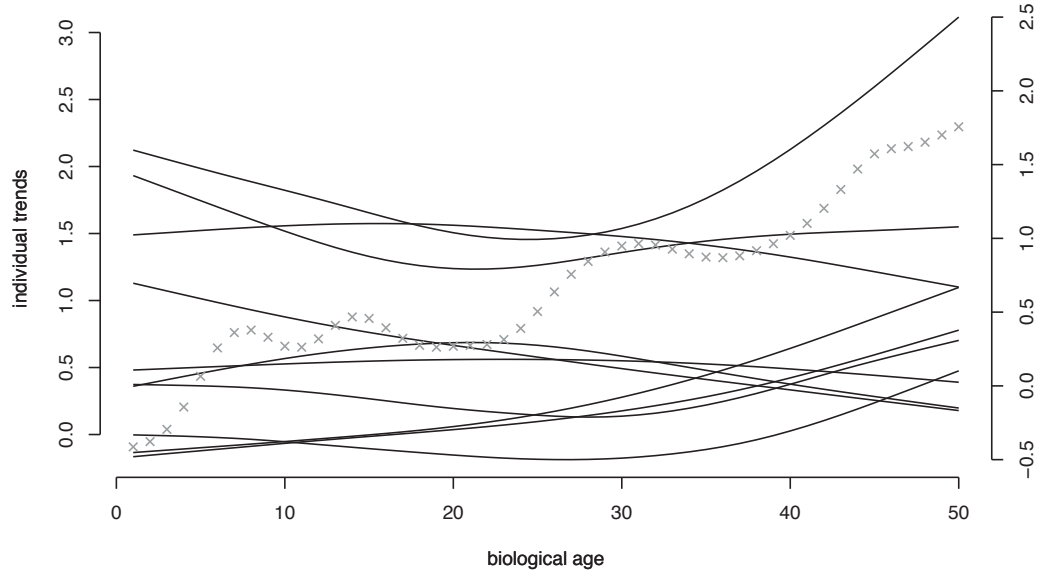

Figure 3.20 Médianes *a posteriori* des tendances de croissance propres à chaque arbre pour un jeu de données simulées, alignées en fonction de leur âge biologique. La courbe grise représente la tendance biologique globale obtenue par la méthode RCS en fonction de l'âge des arbres.

Bibliographie

U. Büntgen, D.C. Frank, D. Nievergelt, and J. Esper. Summer temperature variations in the european alps, a.d. 755-2004. *Journal of Climate*, 19 :5606–5623, 2006.

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

Chapitre 4

Exemples de reconstructions de précipitations en Provence calcaire

Dans ce chapitre nous allons essayer de montrer ce que la méthode décrite dans le chapitre précédent peut apporter à la dendroclimatologie et notamment aux reconstructions climatiques. Tout d'abord, nous verrons de quelle manière nous pouvons prendre en compte les informations supplémentaires apportées par notre nouvelle méthode dans les reconstructions climatiques, puis nous proposerons un exemple de reconstruction de séries de précipitations en Provence calcaire.

4.1 Description de la méthode utilisée

Comme nous venons de le signaler dans la partie précédente, le fait d'avoir une estimation de l'incertitude du signal commun extrait à partir de séries de mesures de cernes d'arbres peut entrainer des répercussions dans le cadre de reconstructions climatiques. Est-ce que si l'on tient compte de cette information supplémentaire la reconstruction de séries de précipitations ou de températures peut être différente de ce que l'on obtient à l'aide des méthodes classiques de la dendrochronologie ? Dans la section 3.1, nous avons pu voir que le signal commun caché dans les séries de cernes d'arbres du site Les Pennes-Mirabeau était relié de manière linéaire au logarithme des précipitations d'été enregistrées à Marseille. On se propose ici de reprendre le même jeu de données et de reconstruire les précipitations d'été passées en prenant ou non en compte l'incertitude estimée sur la variable extraite.

De manière classique, les dendrochronologues supposent que le signal extrait est relié de manière linéaire ou log-linéaire à une ou plusieurs variables climatiques et que cette relation est constante dans le temps. On a alors l'équation suivante :

$$\text{variable climatique} = a + b \text{ signal extrait} + \text{bruit blanc gaussien.} \tag{4.1}$$

Afin de reconstruire la variable climatique considérée, on commence par calibrer cette relation, c'est-à-dire par estimer les paramètres a et b sur la période récente pour laquelle nous disposons de mesures climatiques directes. Ensuite, à l'aide des estimations obtenues pour les paramètres du modèle (4.1) et de la variable extraite sur la période ancienne, on « prédit »les valeurs passées de la variable climatique étudiée. Afin d'estimer l'erreur commise au cours de la reconstruction, les dendrochronologues font généralement appel à des méthodes de bootstrap (Efron, 1979; Efron and Tibshirani, 1986; Guiot, 1990, 1991). Dans notre exemple, on dispose des précipitations d'été à Marseille sur la période 1947-1993. On propose donc d'utiliser les données dont on dispose sur la période 1961-1993 afin de calibrer la relation et de reconstruire le logarithme des précipitations d'été sur la période 1947-1960. On pourra ainsi comparer

nos prédictions aux vraies valeurs et tester la qualité de nos résultats. Le premier graphique de la figure 4.1 représente la reconstruction du logarithme des précipitations d'été obtenu avec son intervalle de confiance à 95%. Cette reconstruction est comparée aux vraies valeurs mesurées sur la période 1947-1960, représentées par la courbe en pointillé.

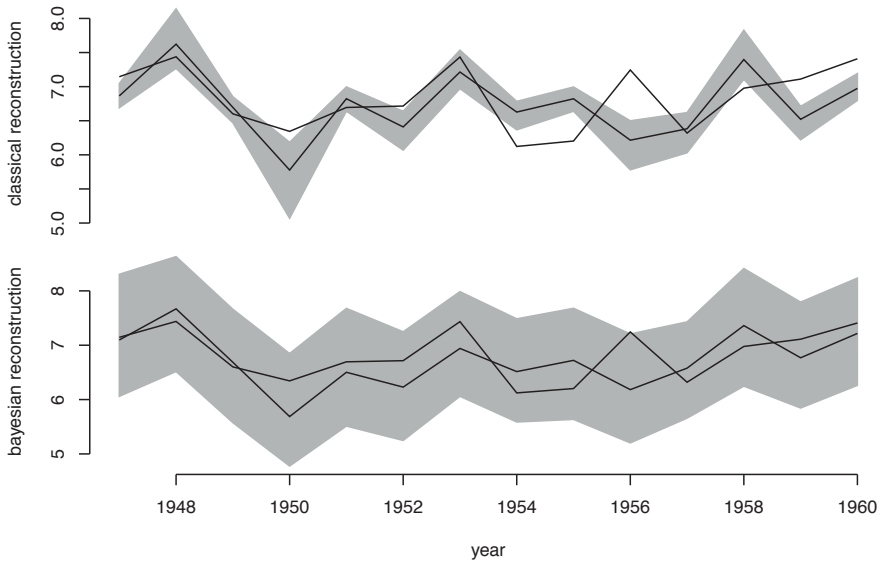

Figure 4.1 Reconstructions du logarithme des précipitations d'été. Le graphique du haut représente la reconstruction obtenue à l'aide du bootstrap avec son intervalle de confiance à 95% et le graphique du bas la médiane *a posteriori* obtenue à l'aide de l'estimation bayésienne avec son intervalle de crédibilité à 95%. Les courbes en pointillé correspondent aux vraies valeurs des précipitations.

On souhaite maintenant prendre en compte l'incertitude estimée sur le signal commun extrait. Pour cela et afin de rester cohérent avec ce qui a été fait précédemment, on se place dans un cadre bayésien. On note $\boldsymbol{P} = (\boldsymbol{P}_O, \boldsymbol{P}_N)^T$ le vecteur du logarithme des précipitations. $\boldsymbol{P}_N$ correspond aux précipitations enregistrées sur la période récente et $\boldsymbol{P}_O$ aux précipitations passées qu'on cherche à reconstruire. De même $\boldsymbol{f} = (\boldsymbol{f}_O, \boldsymbol{f}_N)^T$ correspond au signal

commun extrait pour la période récente et passée. A l'aide de ces nouvelles notations, on peut donc réécrire la relation (4.1)

$$\begin{pmatrix} \boldsymbol{P}_O \\ \boldsymbol{P}_N \end{pmatrix} = a \begin{pmatrix} \mathbf{1}_m \\ \mathbf{1}_{n-m} \end{pmatrix} + b \begin{pmatrix} \boldsymbol{f}_O \\ \boldsymbol{f}_N \end{pmatrix} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n). \tag{4.2}$$

Comme nous nous plaçons dans un contexte bayésien, il nous faut déterminer la vraisemblance de notre modèle et les lois *a priori* de chacun des paramètres afin d'effectuer nos estimations. Concernant la vraisemblance, on déduit de l'équation (4.2) que $\boldsymbol{P}$ suit une loi normale multivariée

$$[\boldsymbol{P}|a, b, \boldsymbol{f}, \sigma^2] = (2\pi)^{-n/2}\sigma^{-n} \exp(-\frac{1}{2\sigma^2}(\boldsymbol{P} - a\mathbf{1}_n - \boldsymbol{f})^T(\boldsymbol{P} - a\mathbf{1}_n - \boldsymbol{f})).$$

On a alors la loi conditionnelle des données $\boldsymbol{P}_N$

$$\begin{aligned}[\boldsymbol{P}_N|a, b, \boldsymbol{f}, \boldsymbol{P}_O, \sigma^2] \quad = \quad & (2\pi)^{-n/2}\sigma^{-n} \exp[-\frac{1}{2\sigma^2}(\boldsymbol{P}_N - a\mathbf{1}_{n-m} - \boldsymbol{f}_N)^T(\boldsymbol{P}_N - a\mathbf{1}_{n-m} - \boldsymbol{f}_N) \\ & -\frac{1}{2\sigma^2}(\boldsymbol{P}_O - a\mathbf{1}_n - \boldsymbol{f}_O)^T(\boldsymbol{P}_O - a\mathbf{1}_n - \boldsymbol{f}_O)].\end{aligned}$$

Pour le choix des lois *a priori*, de manière assez classique, on suppose que les paramètres de régression a et b suivent des lois normales. Afin de tenir compte de l'incertitude estimée sur la variable cachée extraite à l'aide de la méthode décrite dans la section (3.1), il nous paraît opportun de prendre comme loi *a priori* sur $\boldsymbol{f}$ la distribution marginale *a posteriori* obtenue en sortie de notre modèle bayésien hiérarchique semi-paramétrique pour le signal commun. Toutefois, cette distribution n'est pas explicite ce qui rend l'estimation des paramètres du modèle (4.2) assez compliquée. On a démontré que la distribution conditionnelle *a posteriori* du signal extrait est une loi normale multivariée donc on propose, dans un premier temps, d'approcher la distribution marginale de $\boldsymbol{f}$ par une loi normale multivariée en ajustant au mieux les histogrammes de sortie du modèle bayésien hiérarchique semi-paramétrique. Par ailleurs, on sait que la distribution du logarithme des précipitations est très proche d'une loi normale, ce qui justifie le choix d'une telle loi *a priori* pour $\boldsymbol{P}_O$

$$\boldsymbol{P}_O \sim \mathcal{N}_m(\boldsymbol{\mu}_O, \boldsymbol{\Sigma}_O).$$

Une fois notre modèle de probabilité totale établi, on peut calculer les distributions *a posteriori* de chacun des paramètres du modèle. On a en particulier

$$[\boldsymbol{P}_O|a, b, \boldsymbol{f}, \sigma^2, \boldsymbol{P}_N] \propto [\boldsymbol{P}_N|a, b, \boldsymbol{f}, \boldsymbol{P}_O, \sigma^2][\boldsymbol{P}_O].$$

D'où

$$\boldsymbol{P}_O|a, b, \boldsymbol{f}_O, \sigma^2 \sim \mathcal{N}_m(\boldsymbol{\mu}_O^*, \boldsymbol{\Sigma}_O^*)$$

avec

$$\begin{aligned}\boldsymbol{\Sigma}_O^* &= \sigma^2(\mathbf{I}_m + \sigma^2\boldsymbol{\Sigma}_o^{-1})^{-1} \\ \boldsymbol{\mu}_O^* &= \boldsymbol{\Sigma}_O^*\boldsymbol{\Sigma}_O^{-1}\boldsymbol{\mu}_O + (1/\sigma^2)\boldsymbol{\Sigma}_O^*(a\mathbf{1} + b\boldsymbol{f}_O).\end{aligned}$$

Le détail des calculs et les lois *a posteriori* des autres paramètres du modèle sont donnés en annexe B.

Les lois *a posteriori* étant connues de manière explicite, on peut estimer les paramètres de notre modèle à l'aide de l'échantillonnage de Gibbs. Comme pour l'analyse basée sur les méthodes classiques, on suppose qu'on ne dispose que de données de précipitations $\boldsymbol{P}_N$ sur la période 1961-1993 et on cherche à estimer $\boldsymbol{P}_O$ sur la période 1947-1960. Le second graphique de la figure 4.1 représente la médiane *a posteriori* du logarithme des précipitations d'été reconstruites et son intervalle de crédibilité à 95%. Là encore, la courbe en pointillé correspond aux vraies valeurs mesurées sur la période 1947-1960. Si on compare les deux reconstructions obtenues, on peut noter que les fluctuations sont les mêmes mais l'incertitude sur la seconde reconstruction est plus importante que sur la première. Un tel résultat ne nous surprend pas puisque dans le second cas nous avions en entrée de notre régression linéaire une source d'incertitude supplémentaire. Cependant, même si l'incertitude sur les précipitations reconstruites est plus importante, le résultat semble meilleur dans le sens où les vraies valeurs de précipitations sont comprises dans l'intervalle de crédibilité.

4.2 Reconstructions de précipitations en Provence calcaire

En appliquant la méthode que nous venons de décrire, nous allons tenter dans cette section de reconstruire des séries de précipitations en Provence calcaire. Pour cela, nous allons nous appuyer sur les travaux de thèse de Nicault (1999) et en particulier sur les données de cernes d'arbres qu'il a échantillonnées. Il s'agit de séries de cernes de pins d'Alep, portant sur 21 sites situés en Provence calcaire et qui se répartissent selon un gradient nord-sud, la limite nord étant définie par la limite septentrionale du pin d'Alep et la limite sud par la mer. Le gradient est-ouest moins marqué, s'étend, au nord, de la vallée de la Durance à la vallée du Rhône, et au sud, de Toulon à l'étang de Berre. Chaque site a été choisi pour être relativement homogène quant à sa topographie (exposition, altitude, pente) et à la structure de son peuplement. Une quinzaine d'arbres ont été échantillonnés par site, sur une surface maximum d'un hectare. Les arbres ont été choisis dans la mesure du possible, au vu de leur position dominante au sein du peuplement et de leur bon état sanitaire. Les carottes prélevées ont été préparées et mesurées en laboratoire afin d'obtenir l'ensemble des mesures classiques et radiodensitométriques. Ici nous avons retenu la densité moyenne des cernes pour notre étude.

Dans un premier temps, nous allons nous limiter à quatre sites, situés dans la région appelée "zone littorale" (Guiot, 1983; Tessier, 1984). Ces quatre sites sont localisés en noir sur la carte de la figure 4.2. A partir de ces derniers nous allons tenter de reconstruire les précipitations d'été à Marignanne et à Aix-en-Provence, stations pour lesquelles nous disposons de données météorologiques sur la période récente. Plus précisément, à chaque site météorologique on associe deux sites dendrochronologiques. Ainsi les sites de Rognac et Les-Pennes-Mirabeau nous servirons à reconstruire les précipitations de Marignane et les sites de Gardanne et Simiane les précipitations d'Aix-en-Provence.

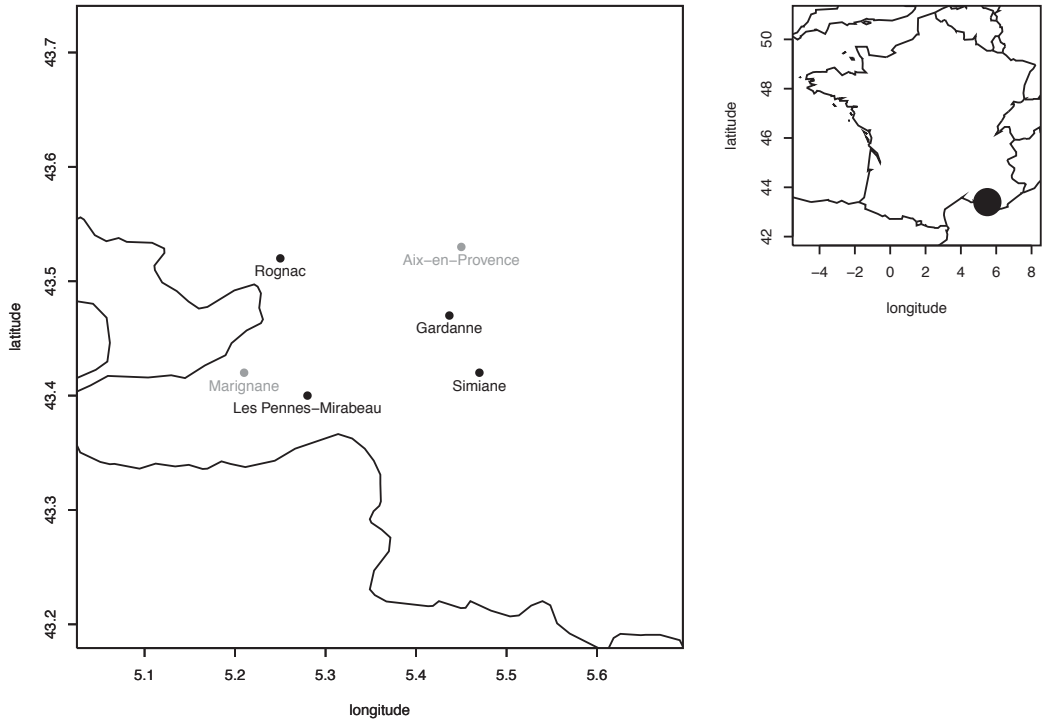

Figure 4.2 Localisation des quatre sites dendrochronologiques en noir et des deux sites météorologiques en gris. La seconde carte situe la Provence calcaire en France.

Reconstruction des précipitations à Marignane Comme nous venons de le voir dans le chapitre précédent, de manière générale, la première étape lors de reconstructions climatiques est d'extraire un signal commun caché dans les cernes d'arbres pour un site dendrochronologique donné. Ainsi, pour chacun des deux sites utilisés pour la reconstruction des précipitations de Marignane, nous appliquons notre modèle bayésien hiérarchique semi-paramétrique décrit dans la section 3.1. Nous obtenons une chronologie par site. La figure 4.3 représente les chronologies obtenues pour les sites de Rognac et Les-Pennes-Mirabeau. La courbe noire correspond à la médiane *a posteriori* du signal extrait et la zone grisée à son intervalle de crédibilité à 95%.

Afin d'effectuer notre reconstruction, nous disposons des précipitations à Marignane entre 1947 et 1993. Comme dans la section précédente, nous proposons de

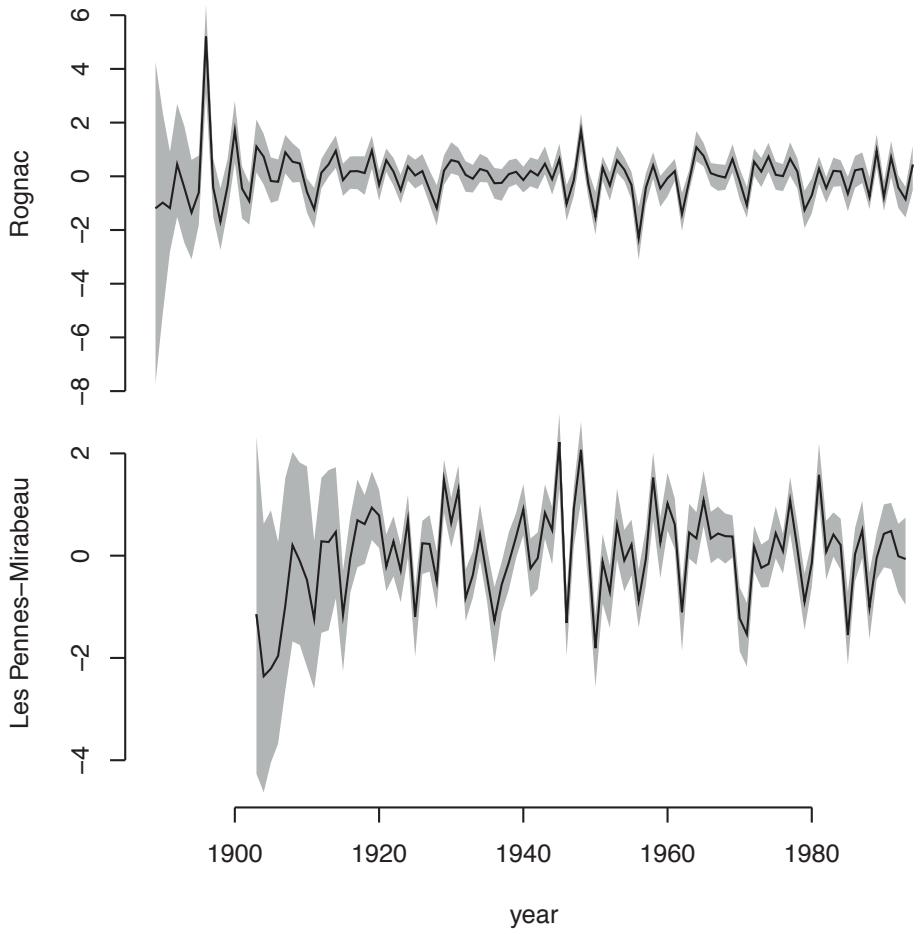

Figure 4.3 Médiane *a posteriori* et intervalle de crédibilité à 95% pour le signal commun extrait des sites de Rognac et Les-Pennes-Mirabeau.

diviser cette période en deux afin d'obtenir une période pour calibrer le modèle (1961-1993) et une période pour comparer nos prédictions aux vraies valeurs afin de tester la qualité de nos résultats (1947-1960). En appliquant le modèle (4.2) à chacun des signaux cachés extraits et aux données de précipitations sur la période 1961-1993, on obtient une reconstruction du logarithme des précipitations d'été à Marignane avant 1961. Pour chacun des sites dendrochronologiques étudiés, la figure 4.4 représente la médiane *a posteriori* des précipitations reconstruites avec leur intervalle de crédibilité à 95%. La courbe en pointillé correspond aux vraies valeurs de précipitations sur la période 1947-1960. On peut noter que dans les deux cas les vraies valeurs sont bien

comprises dans l'intervalle de crédibilité à 95% de nos reconstructions.

Les résultats obtenus semblent assez encourageant car nous retrouvons dans les séries de précipitations reconstruites des années caractéristiques. Par exemple, on peut noter les faibles précipitations des années 1911 et 1946 (croix bleus sur la figure 4.4) qui sont bien connues pour être des années de forte sécheresse.

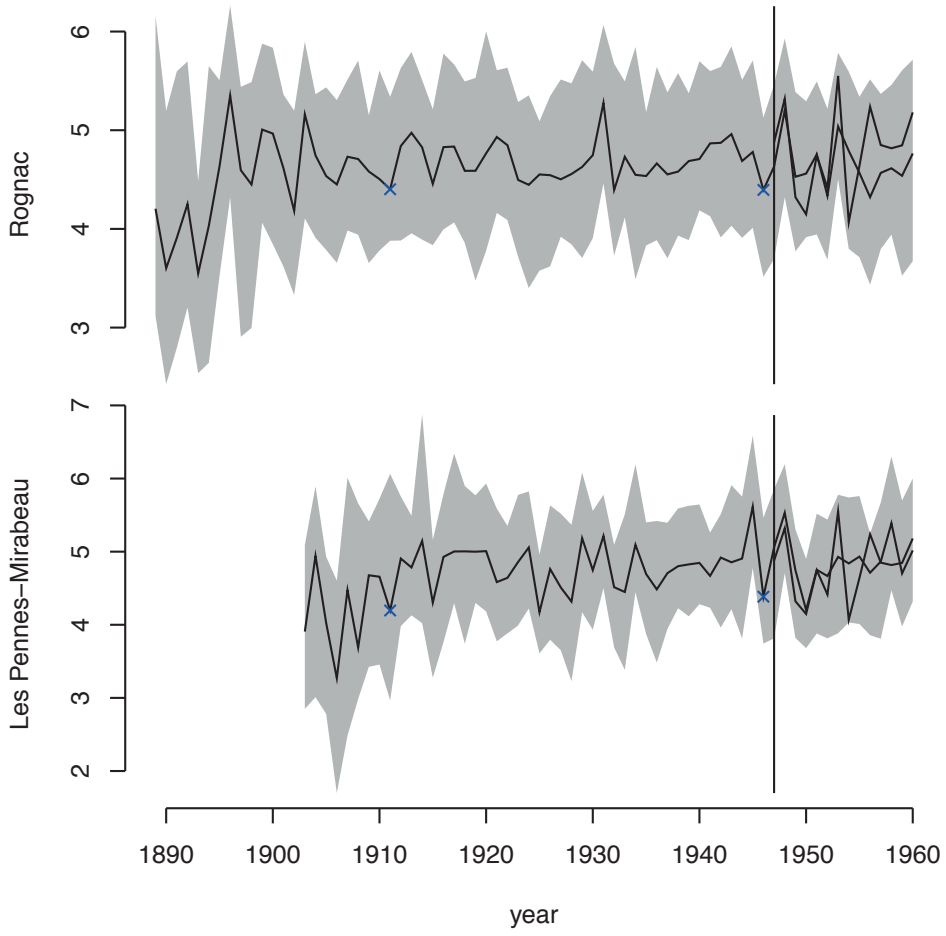

Figure 4.4 Médiane *a posteriori* et intervalle de crédibilité à 95% du logarithme des précipitations reconstruites pour les sites de Rognac et Les-Pennes-Mirabeau. La courbe en pointillé correspond aux vraies valeurs de précipitations sur la période 1947-1960.

Reconstruction des précipitations à Aix-en-Provence De la même manière que précédemment, on cherche à reconstruire les précipitations à Aix-en-Provence. Pour cela, comme nous l'avons signalé plus haut, nous allons utiliser des séries de cernes d'arbres provenant des sites de Gardanne et Simiane. Pour chacun de ces deux sites nous commençons par extraire un signal commun. La figure 4.5 représente les médianes *a posteriori* de chacune des chronologies extraites avec leur intervalle de crédibilité à 95%.

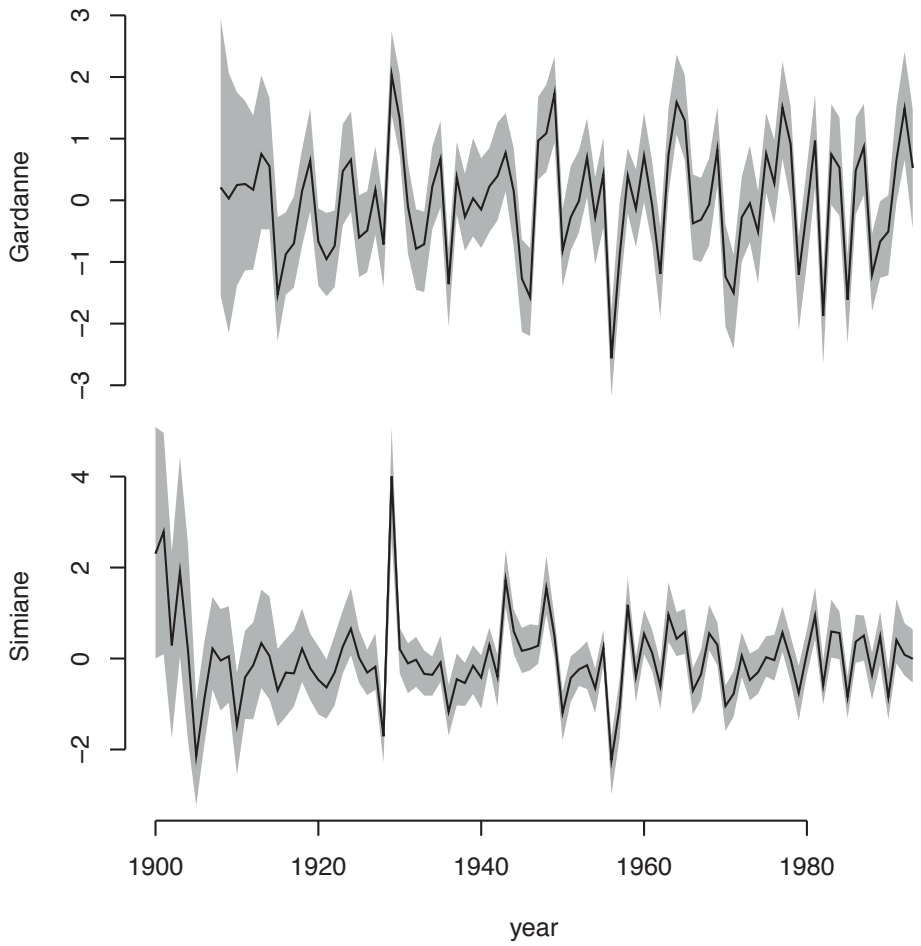

Figure 4.5 Médiane *a posteriori* et intervalle de crédibilité à 95% pour le signal commun extrait des sites de Gardanne et Simiane.

Ici, nous nous trouvons dans un contexte un peu différent que lorsque nous avons reconstruit les températures à Marignane puisque nous disposons des précipitations entre 1900 et 1993 à Aix-en-Provence. Or, les signaux communs que nous avons extraits ne vont pas au delà de 1900, ce qui ne nous permet donc pas de reconstruire des précipitations antérieures à cette date. Par contre, notre série de précipitations n'a pas de données entre 1927 et 1960. Nous proposons donc de reconstruire les précipitations durant cette période. On aura alors la période 1961-1993 comme période de calibration et la période 1900-1926 comme période de contrôle. En appliquant le modèle (4.2) aux signaux cachés extraits pour les sites de Gardane et Simiane, on obtient une recons-

truction du logarithme des précipitations d'été à Aix-en-Provence avant 1961. Pour chacun des sites dendrochronologiques étudiés, la figure 4.6 représente les médianes *a posteriori* des précipitations reconstruites avec leur intervalle de crédibilité à 95%. La courbe en pointillé correspond aux vraies valeurs de précipitations sur la période 1900-1926. On peut noter que dans les deux cas les vraies valeurs sont bien comprises dans l'intervalle de crédibilité à 95% de nos reconstructions, avec en particulier de très bons résultats pour le site de Simiane.

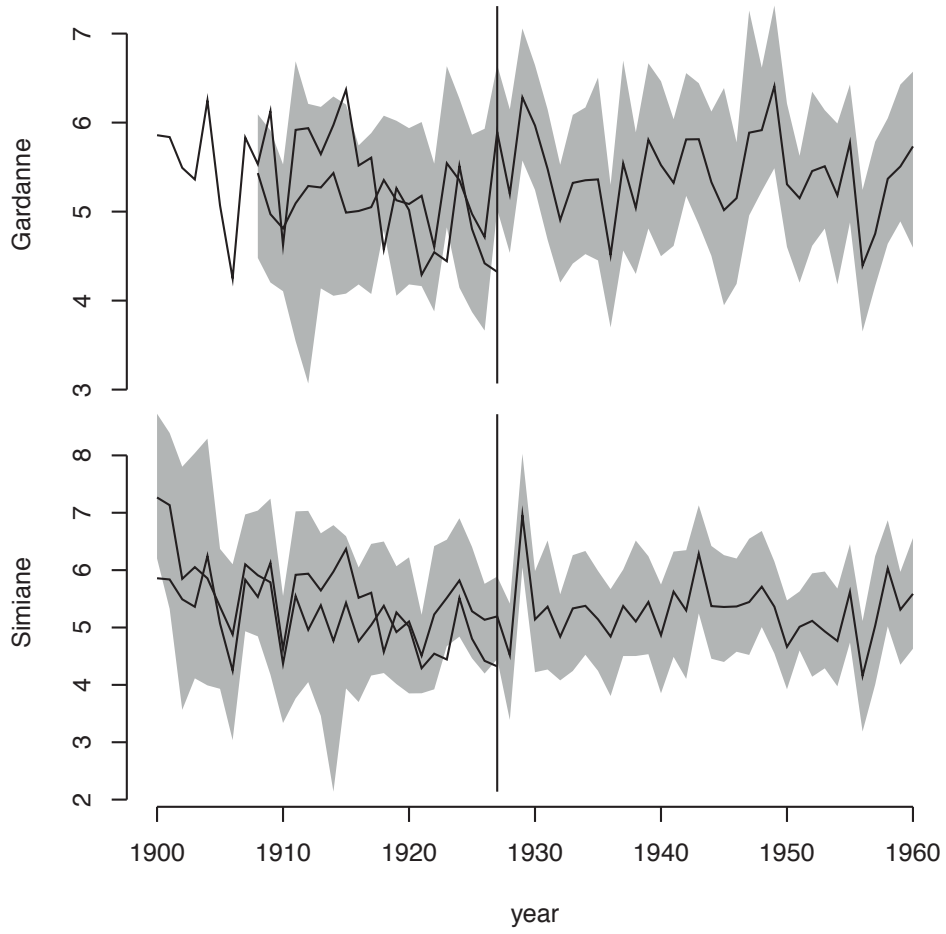

Figure 4.6 Médiane *a posteriori* et intervalle de crédibilité à 95% du logarithme des précipitations reconstruites pour les sites de Rognac et Gardane et Simiane. La courbe en pointillés correspond aux vraies valeurs de précipitations sur la période 1900-1926.

Ainsi, en reprenant les hypothèses de base de la dendroclimatologie, à savoir que les relations cernes-climat sont linéaires et constantes dans le temps, nous avons proposé une méthode nous permettant de prendre en compte l'incertitude associée au signal commun caché dans les mesures de cernes et extrait à l'aide de la méthode décrite dans la section 3.1. Le fait de prendre en compte cette information augmente l'incertitude associée à la reconstruction de variables climatiques mais semble plus cohérente lorsque l'on teste les résultats sur des périodes pour lesquelles on possède des données de précipitations. Dans les reconstructions effectuées, on a pu retrouver certaines années caractéristiques de sécheresse, ce qui semble encourageant.

Bibliographie

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. Guiot. Analyse statistique multidimentionnelle pour la mise en oeuvre d'une climatographie spatio-temporelle, méthodes + exemples. In *Contribution n°38*, page 43. Univ. Catholique de Louvain-la-Neuve, 1983.

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.

A. Nicault. *Analyse de l'influence du climat sur les variations inter et intraannuelle de la croissance radiale du pin d'Alep (Pinus halpensis Mill.) en Provence calcaire.* PhD thesis, Université d'Aix-Marseille III, 1999.

L. Tessier. *Dendroclimatologie et écologie de Pinus sylvestris L. et Quercus pubescens Willd dans le sud-est de la France.* PhD thesis, Université d'Aix-Marseille III, 1984.

Chapitre 5

Sélection de variables pour les modèles additifs généralisés dans le cadre de reconstructions climatiques

Dans ce chapitre, nous proposons une méthode de sélection de variables dans le cadre de modèles bayésiens additifs généralisés. Cette méthode permet de sélectionner les variables climatiques expliquant au mieux le signal commun extrait à partir de séries de cernes d'arbres et de modéliser leur relation. Le fait d'utiliser un modèle additif généralisé apporte une plus grande flexibilité dans la modélisation des relations cernes-climats par rapport aux méthodes classiques de la dendrochronologie qui imposent des relations linéaires.

Contexte statistique : sélection bayésienne de variables Concernant le choix de modèle ou la sélection de variables dans un contexte bayésien, il existe une large littérature pour ce qui est du cas de la régression linéaire (voir par exemple Mitchell and Beauchamp (1988) ; George and McCulloch (1993) ; Chipman (1996) ; Smith and Kohn (1996) ; George and McCulloch (1997) ; Brown et al. (1998) ; Philips and Guttman (1998) ; George (2000) ; Kohn et al. (2001) ; Dupuis and Robert (2003) ; Nott and Green (2004) ; Schneider and Corcoran (2004) ; Casella and Moreno (2004) ; Celeux et al. (2006) ; Cui and George (2008) ; Liang et al. (2008)). Il s'agit de sélectionner un sous espace d'un ensemble total de q variables explicatives. Généralement, chacun des modèles possibles est indexé par un vecteur $\boldsymbol{\gamma}$ de dimension q tel que $\gamma_j = 1$ signifie que la variable explicative j est sélectionnée dans le modèle et $\gamma_j = 0$ signifie qu'elle ne l'est pas. Soit $\mathcal{M}_\gamma$ un modèle indexé par un $\boldsymbol{\gamma}$ donné, l'approche bayésienne de la sélection de modèle implique de spécifier des distributions *a priori* sur les paramètres inconnus de chaque modèle et d'actualiser la vraisemblance $p(\mathcal{M}_\gamma)$ de chacun de ces modèles afin d'obtenir leur probabilité *a posteriori*

$$p(\mathcal{M}_\gamma|\boldsymbol{y}) = \frac{p(\mathcal{M}_\gamma)p(\boldsymbol{y}|\mathcal{M}_\gamma)}{\sum_\gamma p(\mathcal{M}_\gamma)p(\boldsymbol{y}|\mathcal{M}_\gamma)}$$

Le modèle retenu est celui dont la probabilité *a posteriori* est la plus grande. Dans ce chapitre, nous allons donc transposer cette approche au cas de modèles additifs généralisés.

Résumé : La statistique est devenue l'une des composantes essentielles des reconstructions climatiques, qui sont elles-même extrêmement importante pour quantifier le réchauffement climatique actuel. Ainsi, ces dernières années, il y a eu un effort de recherche scientifique important pour combiner de manière spatiale et temporelle différents proxies (c'est-à-dire des mesures indirectes du climat). Cet article s'intéresse à l'un des proxy les plus connu : les mesures de cernes d'arbres. Dans le contexte des reconstructions climatiques, l'une des difficultés statistiques porte sur le choix délicat de variables climatiques explicatives permettant de décrire la croissance des arbres. Le nombre de possibilités est sans fin et dépend de l'espèce des arbres et de la région d'étude. L'expertise des dendrochronologues est extrêmement précieuse,afin d'effectuer une pré-sélection de variables explicatives possibles, mais ne permet pas toujours de répondre complètement à cette question. Le statisticien peut voir le problème des reconstructions climatiques à partir de cernes d'arbres comme un problème de sélection de variables avec une procédure de régression inverse. Dans cet article, on propose un modèle bayésien additif généralisé afin de sélectionner ces variables climatiques et d'estimer la relation entre le climat et la croissance des arbres. Comparé aux méthodes dendrochronologiques passées, ce type de modèle permet des relations non-linéaires. Notre nouveau 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.

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

May 7, 2011

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). This paper focuses on one of the most used climate proxy, tree ring measurements. In a reconstruction context, one of the statistical difficulties concerns the delicate choice of the explanatory climatic variables and their time scales to explain tree growth. The number of possibilities is endless and depends on the tree specie and the region of interest. Hence the dendrochronologist's expertise is invaluable to pre-select possible meaningful explanatory variables which 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 propose a Bayesian Generalized Additive Model to select climatic variables and to estimate relation between climate and tree growth. Compared to past dendroclimatology studies, this type of model allows non-linear relations. Our new model is tested on simulated data and applied to tree rings density measurements (Pinus halepensis Mill.) recorded in French Mediterranean.

1 Climatic reconstruction and tree-rings

In order to understand past and recent climate changes it is necessary to construct long temperature and precipitation series. However, 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 is to derive, study and apply efficient statistical procedures to identify links between information from proxies and temperature or precipitation. One of the best known and

1280

common climate proxy is tree-ring measurements. Since the work of Douglass (1920, 1936), there has been active and extensive research activities 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 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 important 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.

In a reconstruction context, one of the statistical difficulties concerns the delicate choice of the explanatory climatic variables and their time scales to explain tree growth. 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. Hence the dendrochronologist's expertise is invaluable to pre-select possible meaningful explanatory variables which 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. 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 greater 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.

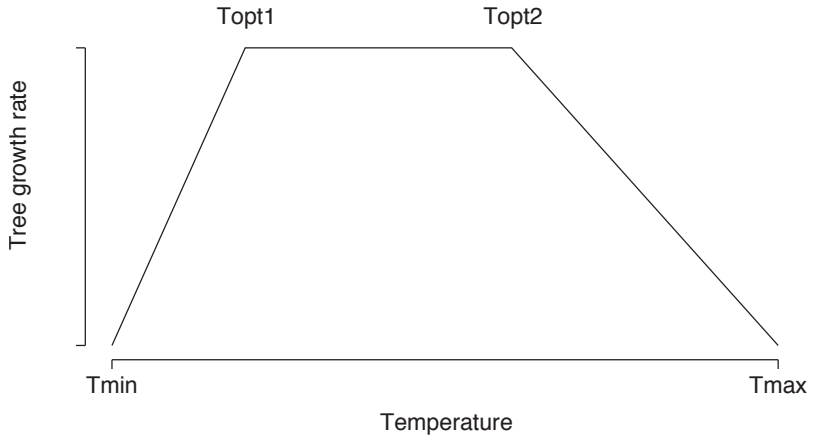

Figure 1: Tree growth rate as a function of temperature. The first segment corresponds to the period of rising growth rate with increasing temperatures below growth-optimal temperature range (Topt1), the second segment to a relatively constant rates within an optimal range of temperatures $[\text{Topt1},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

In light of these reflections, we suppose tree-ring growths are not a simple linear regression of climatic variables but linear regression of climatic data functions. More precisely, we propose to model the relation between tree-rings and climatic factors with a Generalized Additive Model (GAM). As with classical models used by dendrochronologists, the question of the significant climatic variables is crucial. In this paper, we describe a Bayesian variable selection method for GAM. The Bayesian choice is adopted in order to construct a coherent probabilistic framework in which it is possible to formerly assess the variable selection procedure (for other Bayesian dendrochronological papers, see for example (Boreux et al., 2009; Guin et al., 2010)).

1285

2 Bayesian Generalized Additive Models

2.1 Bayesian formulation of smoothing splines

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

$$y_i = f(x_i) + \epsilon_i,\ \epsilon_i \sim \mathcal{N}_1(0, \sigma^2).$$

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

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

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

Another characterization of $\boldsymbol{f}^*$ can be obtained through a Bayesian formulation of the problem (Wahba, 1978; Hastie and Tibshirani, 1990). The Bayesian model can be written as

$$\begin{aligned}\boldsymbol{y}|\boldsymbol{f}, \sigma^2 &\sim \mathcal{N}_n\left(\boldsymbol{f}, \sigma^2\mathbf{I}_n\right),\\ \boldsymbol{f}|\sigma^2, \boldsymbol{\Sigma} &\sim \mathcal{N}_n^{\dagger}\left(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}\right),\end{aligned}$$

where $\boldsymbol{\Sigma} = \lambda^{-1}\mathbf{K}^{-}$. The prior distribution is partially improper, hence the presence of the symbol † (see section 3.6 of Hastie and Tibshirani (1990) for a full discussion). The solution $\boldsymbol{f}^*$ is then given by the expectation of the proper posterior distribution of $\boldsymbol{f}$ which can be obtained from the results of Lindley and Smith (1972)

$$\boldsymbol{f}|\sigma^2, \boldsymbol{\Sigma}, \boldsymbol{y} \sim \mathcal{N}_n\left(\boldsymbol{f}^*, \sigma^2\boldsymbol{\Sigma}^*\right),$$

where

$$\boldsymbol{f}^* = \boldsymbol{\Sigma}^*\boldsymbol{y} = \mathbf{S}_\lambda\boldsymbol{y}, \tag{2}$$

$$\boldsymbol{\Sigma}^* = \left(\mathbf{I}_n + \lambda\mathbf{K}\right)^{-1}, \tag{3}$$

where $\mathbf{S}_\lambda$ is known as the smoothing matrix, a symmetric positive semidefinite matrix.

The prior covariance matrix of $\boldsymbol{f}$ is proportional to the matrix $\mathbf{K}^{-}$ which depends on the input values and the natural cubic spline basis induced by these input values. It is a generalized inverse of the 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, the integral of the penalty term in (1) is proportional to $\mathbf{K}$; more explicitely, it is given by $\int (f^{(2)}(x))^2 dx = \boldsymbol{f}'\mathbf{K}\boldsymbol{f}$, which justifies the prior distribution given above.

Hastie and Tibshirani (1990, 2000) show that this prior covariance matrix $\mathbf{K}^{-}$ is equal to $\mathbf{B}\boldsymbol{\Omega}^{-}\mathbf{B}'$ 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

$$\Omega_{ij} = \int b_i^{(2)}(x) b_j^{(2)}(x) dx.$$

In a Bayesian statistical model, prior distributions also need to be specified for the variance parameter σ^2 and the smoothing parameter λ since these two quantities are not known in applications. This aspect of the problem will be addressed after the formulation of the Bayesian Generalized Additive Model (BGAM).

2.2 Bayesian formulation of Generalized Additive Models

We now suppose that 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 write

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

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. The smooth function, in vector notation, for the jth explanatory variable is then given by

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

a column vector of dimension $n \times 1$. With this notation, an additive model for the observed data can be written as

$$\boldsymbol{y} = \sum_{j=1}^{p} \boldsymbol{f}_j + \boldsymbol{\epsilon}, \text{ with } \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2 \mathbf{I}_n), \tag{4}$$

With a similar structure as the one given in the previous section, the Bayesian model is as follows

$$\boldsymbol{y}|\mathbf{A}, \boldsymbol{\theta}, \sigma^2 \sim \mathcal{N}_n\left(\mathbf{A}\boldsymbol{\theta}, \sigma^2 \mathbf{I}_n\right), \tag{5}$$

$$\boldsymbol{\theta}|\sigma^2, \boldsymbol{\Sigma} \sim \mathcal{N}_{np}^{\dagger}\left(\mathbf{0}_{np}, \sigma^2 \boldsymbol{\Sigma}\right), \tag{6}$$

where

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

$\mathbf{A}$ is a matrix of dimension $n \times np$ made up of a row of $n \times n$ identity matrices and $\mathbf{\Sigma}$, the prior covariance matrix, is a block diagonal matrix of dimension $np \times np$ with diagonal matrix elements $\mathbf{\Sigma}_j = \lambda_j^{-1}\mathbf{K}_j^-$, of dimension $n \times n$ for $j = 1, \ldots, p$. The matrices $\mathbf{K}_j^-$ are defined in the same manner as in the previous univariate case and each λ_j represents the smoothing parameter for the jth spline function, *i.e.* the spline function for the jth explanatory variable.

For fixed values of the λ_j's and a given global variance σ^2, the posterior distribution of the vector of functions, $\boldsymbol{\theta}$, can be calculated explicitely (see Lindley and Smith, 1972)

$$\boldsymbol{\theta}|\sigma^2, \mathbf{\Sigma}, \boldsymbol{y} \sim \mathcal{N}_{np}\left(\boldsymbol{\theta}^*, \sigma^2\mathbf{\Sigma}^*\right),$$

with $\boldsymbol{\theta}^* = \mathbf{\Sigma}^*\mathbf{A}'\boldsymbol{y}$, and

$$\mathbf{\Sigma}^* = \left\{\mathbf{A}'\mathbf{A} + \text{Diag}\left(\lambda_j\mathbf{K}_j\right)\right\}^{-1} = \mathbf{M}^{-1}.$$

It is interesting to note that the jth block diagonal element of $\mathbf{M}$ is given by $\mathbf{S}_{\lambda_j}^{-1} = \mathbf{I}_n + \lambda_j\mathbf{K}_j$, the inverse of the smoothing matrix for the explanatory variable j in a one dimensional context (see equation (2)). It is also worthwhile to point out that the Bayesian formulation takes into account the fact that the functions are fitted simultaneously through the matrix $\mathbf{\Sigma}^*$, or more precisely through the presence of the matrix $\mathbf{A}'\mathbf{A}$ which is a $np \times np$ matrix made up of identity matrices.

So far, nothing has been assumed concerning the prior distributions of the global variance σ^2 and of each smoothing parameter λ_j. Leaving aside the λ_j's for the moment, we take the prior distribution for the global variance to be an inverse gamma distribution:

$$\sigma^2 \sim \mathcal{IG}(a_\sigma, b_\sigma). \tag{7}$$

With this prior distribution, the posterior distribution can be obtained in closed form and it is given by

$$\begin{aligned}
\sigma^2|\mathbf{\Sigma}, \boldsymbol{y} &\sim \mathcal{IG}(a_\sigma^*, b_\sigma^*), \\
a_\sigma^* &= a_\sigma + \frac{(n-2p)}{2}, \\
b_\sigma^* &= b_\sigma + \frac{\boldsymbol{y}'\left(\mathbf{I}_n + \sum_{j=1}^p \mathbf{\Sigma}_j\right)^- \boldsymbol{y}}{2}.
\end{aligned}$$

The posterior distribution of $\boldsymbol{\theta}$, unconditional on σ^2, can be calculated to be: $\boldsymbol{\theta}|\mathbf{\Sigma}, \boldsymbol{y} \sim \mathcal{T}_{np}\left(2a_\sigma^*, \boldsymbol{\theta}^*, \left(\frac{a_\sigma^*}{b_\sigma^*}\right)\mathbf{\Sigma}^*\right)$.

Each smoothing parameter λ_j can take values over $[0, +\infty)$ and therefore in practice, it is difficult to assess a prior distribution and interpret the posterior distribution. To alleviate these difficulties, we 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

1290

relation), ϕ_j goes to 0. Therefore, for an explanatory variable j, it is possible to specify a prior distribution for each ϕ_j which reflects our knowledge on the type of relation which is anticipated. We choose independent prior Beta distributions for the ϕ_j's which we can write as

$$\phi_j \sim \mathcal{B}\left(c_j, d_j\right), \forall j. \tag{8}$$

As far as we know, it is not possible to obtain the posterior distributions of the ϕ_j's explicitely, which can be appreciated by looking out how these parameters appear in the covariance matrix of the prior distribution. Therefore, we rely on the Metropolis-Hastings procedure to get samples from the posterior distributions.

For what follows, we continue to use the $\boldsymbol{\Sigma}$ notation for the prior covariance matrix, with the understanding that each block diagonal element is now given by: $\boldsymbol{\Sigma}_j = \left(\frac{\phi_j}{1-\phi_j}\right)\mathbf{K}_j^-$. Furthermore, the vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_j, \ldots, \phi_p)'$ will be used to designate all the ϕ_j's.

3 Bayesian variable selection for Generalized Additive Models

There is a large literature on model choice and variable selection for the linear regression context (see for instance Mitchell and Beauchamp (1988); George and McCulloch (1993); Chipman (1996); Smith and Kohn (1996); George and McCulloch (1997); Brown et al. (1998); Philips and Guttman (1998); George (2000); Kohn et al. (2001); Dupuis and Robert (2003); Nott and Green (2004); Schneider and Corcoran (2004); Casella and Moreno (2004); Celeux et al. (2006); Cui and George (2008); Liang et al. (2008)). The model choice problem involves selecting a subset of a total of q predictor variables. Generally, the model space is indexed by $\boldsymbol{\gamma}$, a q-dimensional vector of indicator variables with $\gamma_j = 1$ meaning that $\boldsymbol{x}_{.j}$ is included in the set of predictor variables and with $\gamma_j = 0$ meaning that $\boldsymbol{x}_{.j}$ is excluded. Let $\mathcal{M}_{\boldsymbol{\gamma}}$ be a model indexed by a given $\boldsymbol{\gamma}$, the Bayesian approach to model selection involves specifying priors on the unknown parameters in each model, and in turn updating prior probabilities of models $p(\mathcal{M}_{\boldsymbol{\gamma}})$ to obtain posterior probabilities of each model

$$p(\mathcal{M}_{\boldsymbol{\gamma}}|\boldsymbol{y}) = \frac{p(\mathcal{M}_{\boldsymbol{\gamma}})p(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{\gamma}})}{\sum_{\boldsymbol{\gamma}} p(\mathcal{M}_{\boldsymbol{\gamma}})p(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{\gamma}})}$$

A key component in the posterior model probabilities is the marginal likelihood of the data $p(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{\gamma}})$ under the model $\mathcal{M}_{\boldsymbol{\gamma}}$, obtained by integrating the likelihood with respect to the prior distributions for model specific parameters. If we take the prior probabilities, $p(\mathcal{M}_{\boldsymbol{\gamma}})$, to be the same across models, we see that the posterior probabilities only depend on the marginal likelihoods. In this context, the different models can be compared through their marginal likelihoods. In this section we describe a method to calculate the marginal likelihoods for GAMs.

A model $\mathcal{M}_{\boldsymbol{\gamma}}$ in our context is defined by which p variables (out of q) are included in the model, as defined by the vector $\boldsymbol{\gamma}$, and the model distributions given in equations

(5), (6), (7), and (8). In order to get the marginal distribution for $\mathcal{M}_{\boldsymbol{\gamma}}$, all model parameters must be integrated. More specifically, one needs to integrate over the parameters $\boldsymbol{\theta}$, σ^2, and $\boldsymbol{\phi}$. Using standard Bayesian results, this can be performed for the first two parameters to obtain

$$p\left(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{\gamma}},\boldsymbol{\phi}\right) \equiv \mathcal{T}_n^{\dagger}\left(2a_\sigma, \mathbf{0}, (b_\sigma/a_\sigma)\left\{\mathbf{I}_n + \sum_{j=1}^{p}\boldsymbol{\Sigma}_j\right\}\right),$$

where the conditional distribution takes into account the fact that the matrices $\boldsymbol{\Sigma}_j$ only depend on the vector $\boldsymbol{\phi}$ since the $\mathbf{K}_j$ are fully determined by the vectors $\boldsymbol{x}_{\cdot j}$ in our context.

In order to calculate the full marginal distribution, it is necessary to integrate this last distribution relative to the parameter $\boldsymbol{\phi}$. To our knowledge, it is not possible to perform this operation explicitely but we propose to use a method developed by Chib and Jeliazkov (2001). The idea behind the method is to inverse Bayes' theorem to get the following expression

$$p\left(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{\gamma}}\right) = \frac{p\left(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{\gamma}},\boldsymbol{\phi}\right)\pi\left(\boldsymbol{\phi}|\mathcal{M}_{\boldsymbol{\gamma}}\right)}{\pi\left(\boldsymbol{\phi}|\mathcal{M}_{\boldsymbol{\gamma}},\boldsymbol{y}\right)}.$$

For an appropriate value of $\boldsymbol{\phi}^*$, *i.e.* a value of high density, we then have

$$p\left(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{\gamma}}\right) = \frac{p\left(\boldsymbol{y}|\mathcal{M}_{\boldsymbol{\gamma}},\boldsymbol{\phi}^*\right)\pi\left(\boldsymbol{\phi}^*|\mathcal{M}_{\boldsymbol{\gamma}}\right)}{\pi\left(\boldsymbol{\phi}^*|\mathcal{M}_{\boldsymbol{\gamma}},\boldsymbol{y}\right)},$$

for which only the denominator, the posterior distribution evaluated at $\boldsymbol{\phi}^*$, needs to be approximated.

Chib and Jeliazkov show a simulation-consistent estimate of the posterior ordinate is available as

$$\frac{G^{-1}\sum_{g=1}^{G}\alpha(\boldsymbol{\phi}^{(g)},\boldsymbol{\phi}^*|\boldsymbol{y})q(\boldsymbol{\phi}^{(g)},\boldsymbol{\phi}^*|\boldsymbol{y})}{L^{-1}\sum_{l=1}^{L}\alpha(\boldsymbol{\phi}^*,\boldsymbol{\phi}^{(l)}|\boldsymbol{y})},$$

where $\boldsymbol{\phi}^{(g)}$ are the sample draws from the posterior distribution with the Metropolis-Hasting algorithm, $\boldsymbol{\phi}^{(l)}$ are draws from the proposal density $q(\boldsymbol{\phi}^*,\boldsymbol{\phi}|\boldsymbol{y})$ and α is the probability to move.

4 Data analysis

4.1 Simulations results

We consider the following full model

$$\boldsymbol{y}|\sigma^2 \sim \mathcal{N}_3(\sum_{j=1}^{3} f_j(\boldsymbol{x}_{\cdot j}), \sigma^2\mathbf{I}_3) \tag{9}$$

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.

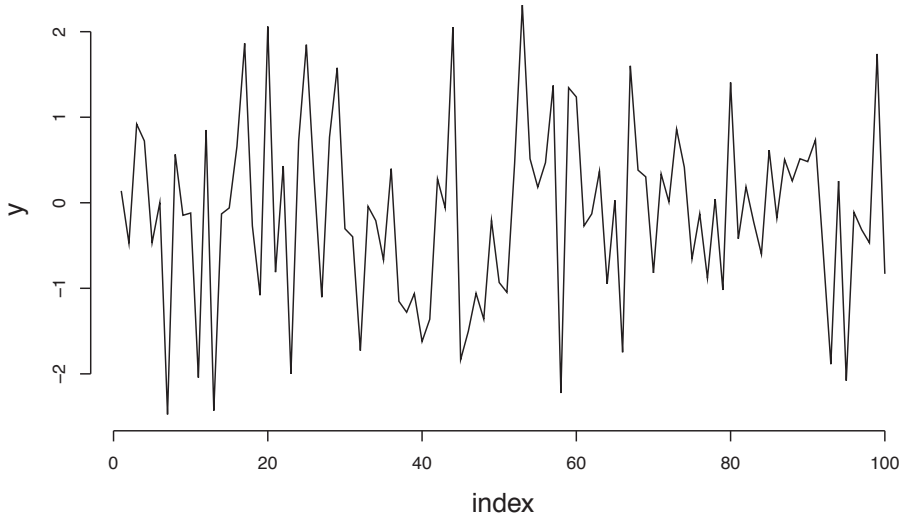

Figure 2: Simulated data with model (9)

For each possible model, i.e. for all explanatory variables combinations we can estimate the marginal likelihood $\left[\boldsymbol{y}|\left\{\mathbf{K}_j\right\}_{j=1,\ldots,p}\right]$. 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 variable in our case. This result seems coherent because this is the true model.

Selected variables	$\log p(\boldsymbol{y}\|\left\{\mathbf{K}_j\right\}_{j=1,\ldots,p})$
1,2,3	-58.41465
1,2	-54.67726
1,3	-113.8362
2,3	-101.1376
1	-116.5670
2	-100.4158
3	-122.7672

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 chosen 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 $f_2(.)$.

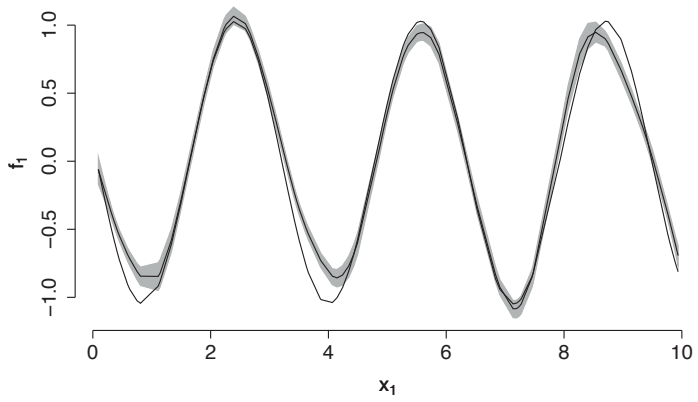

Figure 3: Posterior information about the function $f_1(.)$. The solid and dotted lines correspond to the true $f_1(.)$ and the posterior estimated median, respectively. The gray area represents the 95 % credibility intervals.

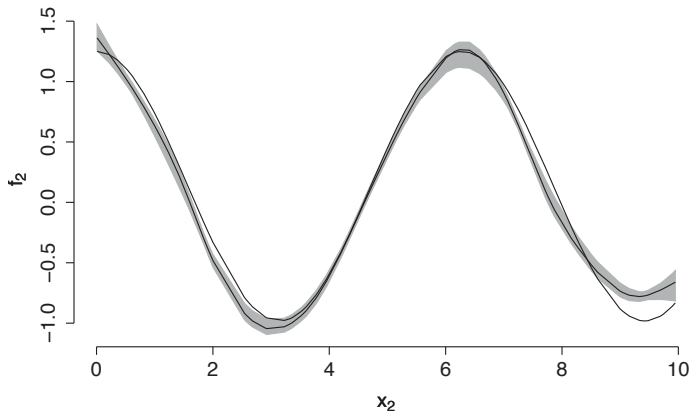

Figure 4: Posterior information about the function $f_2(.)$. The solid and dotted lines correspond to the true $f_2(.)$ and the posterior estimated median, respectively. The gray area represents the 95 % credibility intervals.

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

To exemplify and discuss our approach, we analyzed a set of fourteen Pinus halepensis Mill. Figure 5 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 like Spring and Summer precipitation or Spring temperatures. This past study provides a referential for our climatic variables selection.

Our example data set is composed by fourteen Pinus halepensis Mill tree ring density time series (in mg/cm3), corresponding to the growth of fourteen trees. Basically individual trees at an environmentally homogenous site can have their own physiological aging process but they can also share a common element due to the local environment. One of the main dendroclimatologist interests resides in finding this common element to make climatic interpretation. So, a first set is to extract a climatically-related environmental common signal from this series. We use a recent method develop by Guin et al. (2010) and we obtain the signal represented by Figure 6.

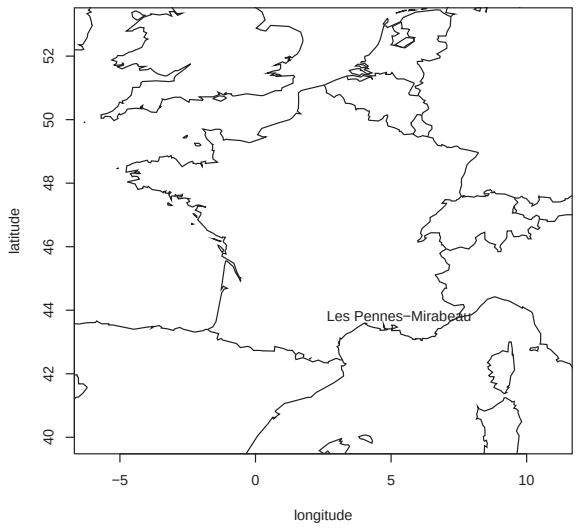

Figure 5: The "Les Pennes-Mirabeau" site located in the South of France where *Pinus halepensis Mill* tree ring densities series were recorded.

We investigate potential links between our extracted signal and climatic variables. We focus on explanatory variables identified by Nicault et al. (2001) : Spring temperatures, Spring and Summer precipitation. We check there are no correlation between this variables and for all explanatory variables combinations we can estimate the marginal

likelihood $\left[\boldsymbol{y}|\left\{\mathbf{K}_j\right\}_{j=1,\ldots,p}\right]$. These results are summarized in the Table 2. The best model which have the more important probability is the model with just Spring temperatures explanatory variable.

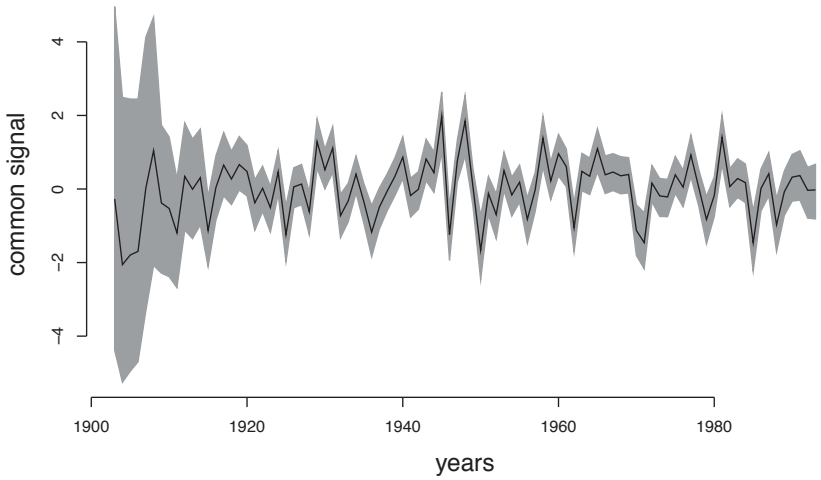

Figure 6: Posterior median of the common signal obtained from trees measured at the site of “Les Pennes Mirabeau" (solid line) with its 95% credibility interval (gray area).

Selected variables	$\log p(\boldsymbol{y}\|\left\{\mathbf{K}_j\right\}_{j=1,\ldots,p})$
Spring precipitation + Summer precipitation + Spring temperatures	-56.69
Spring precipitation + Summer precipitation	-72.66
Spring precipitation + Spring temperatures	-87.19
Summer precipitation + Spring temperatures	-82.23
Spring precipitation	-80.20
Summer precipitation	-76.82
Spring temperatures	-33.47

Table 2: 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

For the chosen model, we can estimate the smooth functions $f(.)$. Figure 7 represents the posterior median for $f(.)$ and the gray area the 95 % credibility interval. We note this function is not linear and it seems to have threshold effect.

If we apply classical response functions (Guiot, 1991) to our dataset two variables are supposed significant : Spring temperatures and Summer precipitation. To compare the two methods we propose to study residuals for each model. Figure 8 represents residuals for response functions and Figure 9 for Bayesian Generalized Additive Model. In this two figures, the first graph corresponds to a residuals normal Q-Q plot

1300

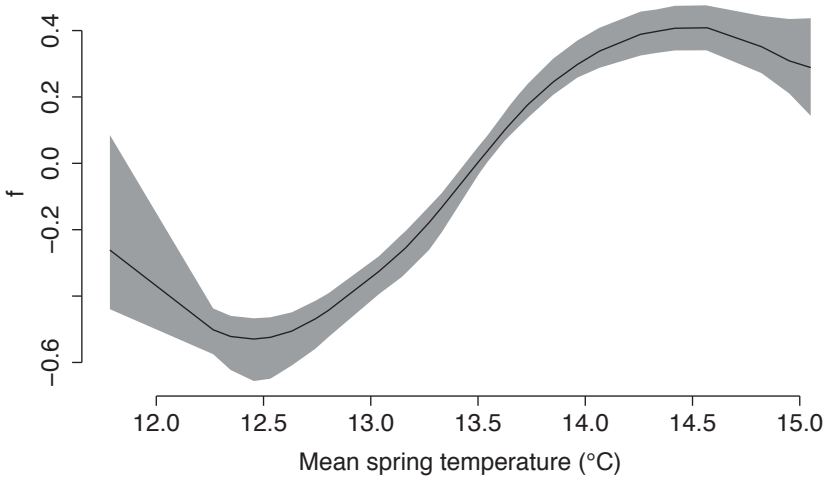

Figure 7: Posterior information about the function $f(.)$. The solid line corresponds to the posterior estimated median and the gray area represents the 95 % credibility interval.

and the second graph to a residuals histogram. We can note with Bayesian Generalized Additive Model residuals seem more gaussian than with response functions.

5 Conclusion

From a statistical perspective, the choice of the explanatory climatic variables and their time scales to explain tree growth could be viewed like a selection variables or a selection model problem. In the light of the past studies, we opted to model the relation between tree growth and climatic factors with a Generalized Additive Model and we proposed a variables selection method for this type of models in a bayesian context. The choice of a Generalized Additive Model is interesting because, contrary to classical dendrochronological methods, it allows the tree growth-climate link is not linear and provides more flexibility to the modelling. To select variables we work with the possible models marginal distributions and we choose the model with the more important posterior probability. The difficulty of this procedure is to calculate the full marginal distributions because they do not have explicit expression. To solve this problem we we propose to use a method developed by (Chib and Jeliazkov, 2001).

Our analysis of *Pinus halepensis Mill* tree ring density series gave interesting results. Our new model selects less variables than response functions (Spring temperatures for Generalized Additive Model and Spring temperatures and Summer precipitation for response functions), but the residuals seem best.

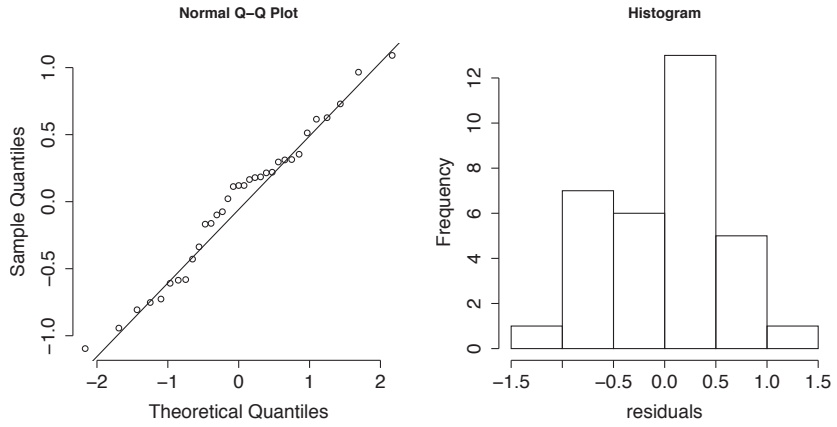

Figure 8: Residuals for response functions. The first graph corresponds to a residuals normal Q-Q plot and the second graph to a residuals histogram.

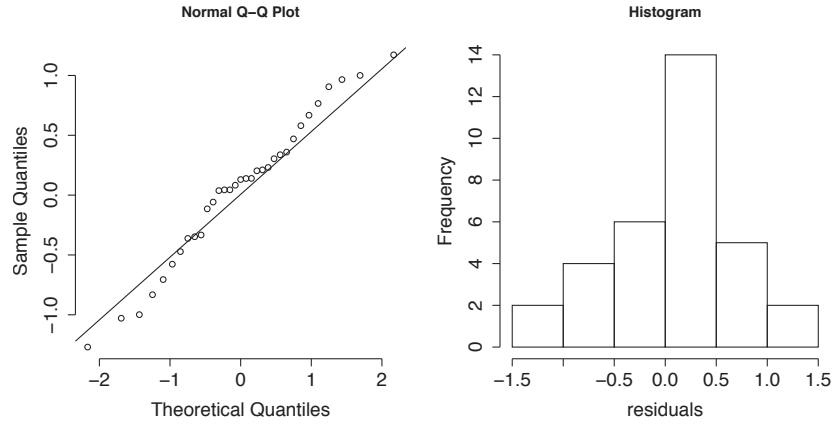

Figure 9: Residuals for Bayesian Generalized Additive Model. The first graph corresponds to a residuals normal Q-Q plot and the second graph to a residuals histogram.

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.

J.-J. Boreux, P. Naveau, O. Guin, L. Perreault, and J. Bernier. 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, 2009.

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

P. Brown, M. Vannucci, and T. Fearn. Multivariate bayesian variable selection and prediction. *J. Royal Statist. Soc. Series B*, pages 627–641, 1998.

G. Casella and E. Moreno. Objective bayesian variable selection. Technical report, University of Florida, 2004.

G. Celeux, J.-M. Marin, and C. Robert. Sélection bayésienne de variables en régression linéaire. *Journal de la Société Française de Statistique*, 147(1):59–79, 2006.

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

H. Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 1:17–36, 1996.

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.

W. Cui and E. George. Empirical bayes vs. fully bayes variable selection. *Journal of Statistical Planning and Inference*, 138:888–900, 2008.

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.

J. Dupuis and C. Robert. Bayesian variable selection in qualitative models by kullback-leibler projections. *J. Statist. Plann. Inference*, pages 77–94, 2003.

1305

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.

E. George. The variable selection problem. *J. American Statist. Assoc.*, 95:1304– 1308, 2000.

E. George and R. McCulloch. Variable selection via gibbbs sampling. *J. American Statist. Assoc.*, 88:881–889, 1993.

E. George and R. McCulloch. Approaches to bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

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

O. Guin, P. Naveau, and J.-J. Boreux. Extracting hidden trends in tree rings with a semi-parametric bayesian hierarchical model. *submitted*, 2010.

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.

R. Kohn, M. Smith, and D. Chan. Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11:313–322, 2001.

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

F. Liang, R. Paulo, G. Molina, M. Clyde, and J. Berger. Mixtures of g-priors for bayesian variable selection. *J. American Statist. Assoc.*, 103(481):410–423, 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.

T. Mitchell and J. Beauchamp. Bayesian variable selection in linear regression. *J. American Statist. Assoc.*, 83:1023–1032, 1988.

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.

A. Nicault, C. Rathgeber, L. Tessier, and A. Thomas. 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, 2001.

D. J. Nott and P. J. Green. Bayesian variable selection and the swendsen-wang algorithm. *J. Comput. Graph. Statist.*, 13:1–17, 2004.

R. Philips and I. Guttman. A new criterion for variable selection. *Statist. Prob. Letters*, 38:11–19, 1998.

U. Schneider and J. Corcoran. Perfect sampling for bayesian variable selection in a linear regression model. *J. Statist. Plann. Inference*, 126:153–171, 2004.

M. Smith and R. Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75:317–343, 1996.

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.

Chapitre 6

Conclusions et perspectives

L'objectif principal de cette thèse était d'analyser les forces et les faiblesses de quelques méthodes classiques de reconstructions climatiques, à partir notamment de données sur la croissance des cernes d'arbres, et de développer de nouvelles méthodes permettant une amélioration des résultats obtenus et de leur précision. Pour cela nous avons fait le choix de la statistique bayésienne qui nous a semblé un outil intéressant pour l'estimation des incertitudes liées à ces reconstructions. Dans cette conclusion, nous résumerons rapidement les principales avancées méthodologiques puis nous donnerons quelques pistes concernant les prolongements possibles à aborder et les questions qui restent ouvertes.

Développements méthodologiques Comme nous avons pu le voir, la grande majorité du travail effectué au cours de cette thèse porte sur des questions méthodologiques, avec en particulier le développement de deux modèles : un modèle bayésien hiérarchique semi-paramétrique permettant l'estimation de tendances cachées dans un jeu de plusieurs séries chronologiques et un modèle bayésien additif généralisé incluant le problème de la sélection de variables. Ces développements peuvent trouver de nombreux champs d'application autre que celui développé dans ce manuscrit et être adaptés suivant les besoins.

Concernant la dendroclimatologie, et plus largement les reconstructions climatiques, trois thèmes de développements méthodologiques sembleraient intéressants à approfondir. Tout d'abord, comme nous l'avons vu, les reconstructions climatiques s'effectuent en deux étapes : extraction d'un signal commun climatique à l'aide du modèle décrit dans la section 3.1, puis estimation de la relation entre ce dernier et

certaines variables climatiques à l'aide d'un modèle additif généralisé (chapitre 5). Cependant, lorsqu'on a estimé cette relation et qu'on a sélectionné les variables climatiques expliquant au mieux le signal commun extrait, on a utilisé la médiane *a posteriori* obtenue pour ce dernier et ainsi, on n'a pas tenu compte des erreurs commises lors de l'extraction. Une solution relativement simple serait de traiter le problème de la même manière que dans le chapitre 4 et de prendre comme loi *a priori* pour le signal à expliquer une approximation normale de la loi *a posteriori* obtenue en sortie du modèle décrit dans la section 3.1. Cependant, en faisant une telle approximation, nous introduisons une source d'erreur supplémentaire qui n'est pas prise en compte dans notre reconstruction. Il serait donc intéressant de réfléchir à la manière dont on peut introduire dans notre modèle la loi *a posteriori* exacte du signal commun estimé à l'aide de notre modèle bayésien hiérarchique semi-paramétrique.

Une seconde piste intéressante à approfondir serait l'aspect spatial du problème. Jusque là nous nous sommes limités à une étude temporelle, mais les dendrochronologues disposent d'une importante base de données réparties sur de nombreux sites et dans de nombreuses régions du monde. Si l'on reprend l'exemple des 21 sites échantillonnés par Nicault au cours de sa thèse et décrits dans le section 4.2, on peut noter l'apparition d'une structure spatiale. En effet, pour chacun des 21 sites nous avons extrait un signal commun à l'aide de notre modèle bayésien hiérarchique semi-paramétrique (section 3.1) et nous avons calculé la corrélation entre les médianes *a posteriori* des différents signaux obtenus. La Figure 6.1 représente la distance entre les sites en fonction de ces corrélations. On voit qu'il semble exister un lien entre les deux : plus les sites sont proches plus les signaux extraits ont des profils similaires. D'autre part, si on cherche à reconstruire le climat d'une région donnée, il semble pertinent d'utiliser des données sur plusieurs sites de cette région plutôt que de se limiter à un site unique. Une idée pourrait donc être d'ajouter une couche supplémentaire à notre modèle d'extraction. Plus précisément, nous pourrions introduire dans notre modèle un signal commun "régional" qui dépendrait des signaux communs extraits pour chacun des sites, avec une structure spatiale qui pourrait dépendre de la dis-

tance entre les sites. Afin de mener à bien notre réflexion sur le sujet, nous pourrions également nous inspirer des travaux de Hooten and Wikle (2007) et de Li et al. (2010) qui ont eux aussi cherché à intégrer une dimension spatiale dans leur analyse.

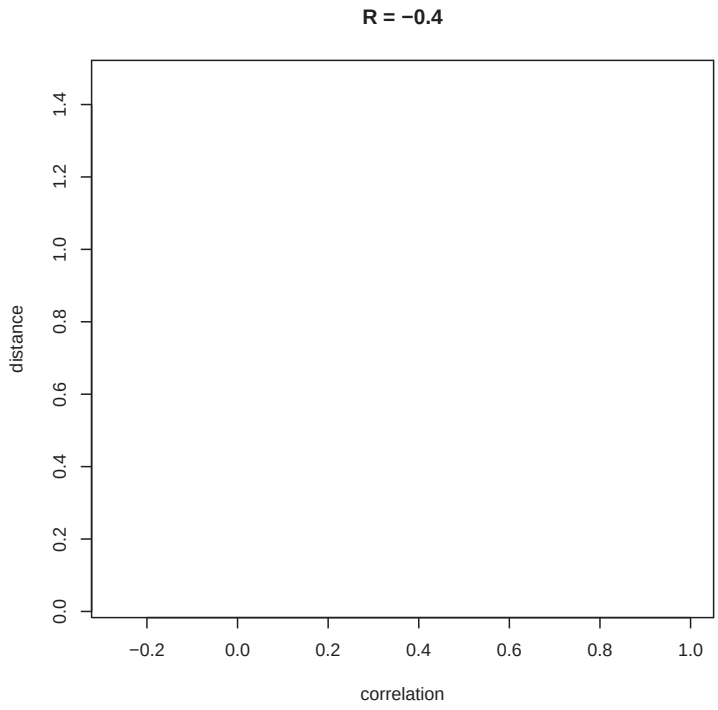

Figure 6.1 Distance entre les sites échantillonnés en Provence calcaire en fonction des corrélations entre les médianes *a posteriori*.

Pour finir, aujourd'hui un nouvelle stratégie concernant les reconstructions climatiques consiste à faire "l'inverse" de ce que nous avons fait dans ce manuscrit, c'est-à-dire de produire un ensemble de proxies simulés à partir de facteurs environnementaux (Hughes et al., 2010 ; Hugues and Ammann, 2009 ; Guiot et al., 2009). Dans ce contexte, la capacité de notre modèle bayésien hiérarchique semi-paramétrique (section 3.1) à simuler des mesures de cernes d'arbres synthétiques serait intéressante à explorer. Une possibilité pourrait être de contraindre la fonction f (autrement dit le signal commun supposé climatique) avec des facteurs environnementaux, puis de simuler des séries de croissance de cernes d'arbres à l'aide de notre modèle. L'une des

difficultés principales d'une telle approche réside dans la forte homogénéité des effet individuels de l'âge.

Analyse des données Au cours de cette thèse, nous nous sommes principalement intéressés à des questions méthodologiques, mais la qualité des reconstructions climatiques ne dépendent pas seulement de ces dernières. En effet, la qualité et le choix des données utilisées pour l'analyse ne sont pas moins importants. Tout d'abord, dans les exemples qui illustrent les méthodes développées nous avons fait le choix, relativement « arbitraire », d'utiliser des données de densités de cernes d'arbres. Or, il pourrait être intéressant de se demander si cette mesure est la plus adaptée pour répondre à nos objectifs. En effet, si l'on reprend par exemple le jeu de données de la section 3.1, correspondant au site Les Pennes-Mirabeau, on dispose de données portant sur d'autres paramètres que la densité des cernes. On peut donc extraire un signal commun pour chacun des jeux de mesures dont on dispose (densité du bois initial, densité du bois final, densité totale, largeur du bois initial, largeur du bois final, largeur totale). La figure 6.2 représente les signaux ainsi obtenus. On voit que les signaux extraits sont quelque peu différents les uns des autres, ce qui entraine bien sûr des reconstructions climatiques différentes. Une étude cherchant à déterminer à partir de quel type de mesure et dans quel contexte nous obtenons des résultats le plus proche possible de la réalité serait donc intéressante à mener.

Pour finir, il semble qu'il serait également intéressant de mener une réflexion sur le choix des variables climatiques ou environnementales explicatives. En effet, dans les exemple utilisés au cours de cette thèse nous avons travaillé à l'aide de données de précipitations ou de températures ; ce choix est d'ailleurs un choix classique en dendroclimatologie. Cependant, on peut se demander si le signal commun extrait est bien relié à ces données et non pas à d'autre paramètres environnementaux. Cette réflexion nous vient de résultats récents obtenus par Boreux et présentés à l'université de Montréal lors de la conférence « Méthodes statistiques en météorologie et chan-

gement climatique »(2011). En effet, ce dernier à repris le jeu de données portant sur des aires de cernes d'arbres au Québec et décrit dans la section 2.2, lui a appliqué le modèle de la section 3.1 et a comparé les résultats obtenus à la moyenne des débits des mois de mai et juin observés au bassin de Caniapiscau. La figure 6.3 représente la médiane *a posteriori* obtenue à partir des données de cernes d'arbres et les débits moyens des mois de mai et juin observés au bassin de Caniapiscau. Ces deux séries semblent particulièrement corrélées (avec une corrélation de 0.5), ce qui nous laisse à penser que l'information environnementale enregistrée par les arbres de cette région est fortement liée aux débits, variable intéressante puisqu'elle intègre en quelque sorte plusieurs facteurs environnementaux et climatiques. Ainsi, il serait peut être intéressant d'étendre nos recherche à de nouvelles variables explicatives.

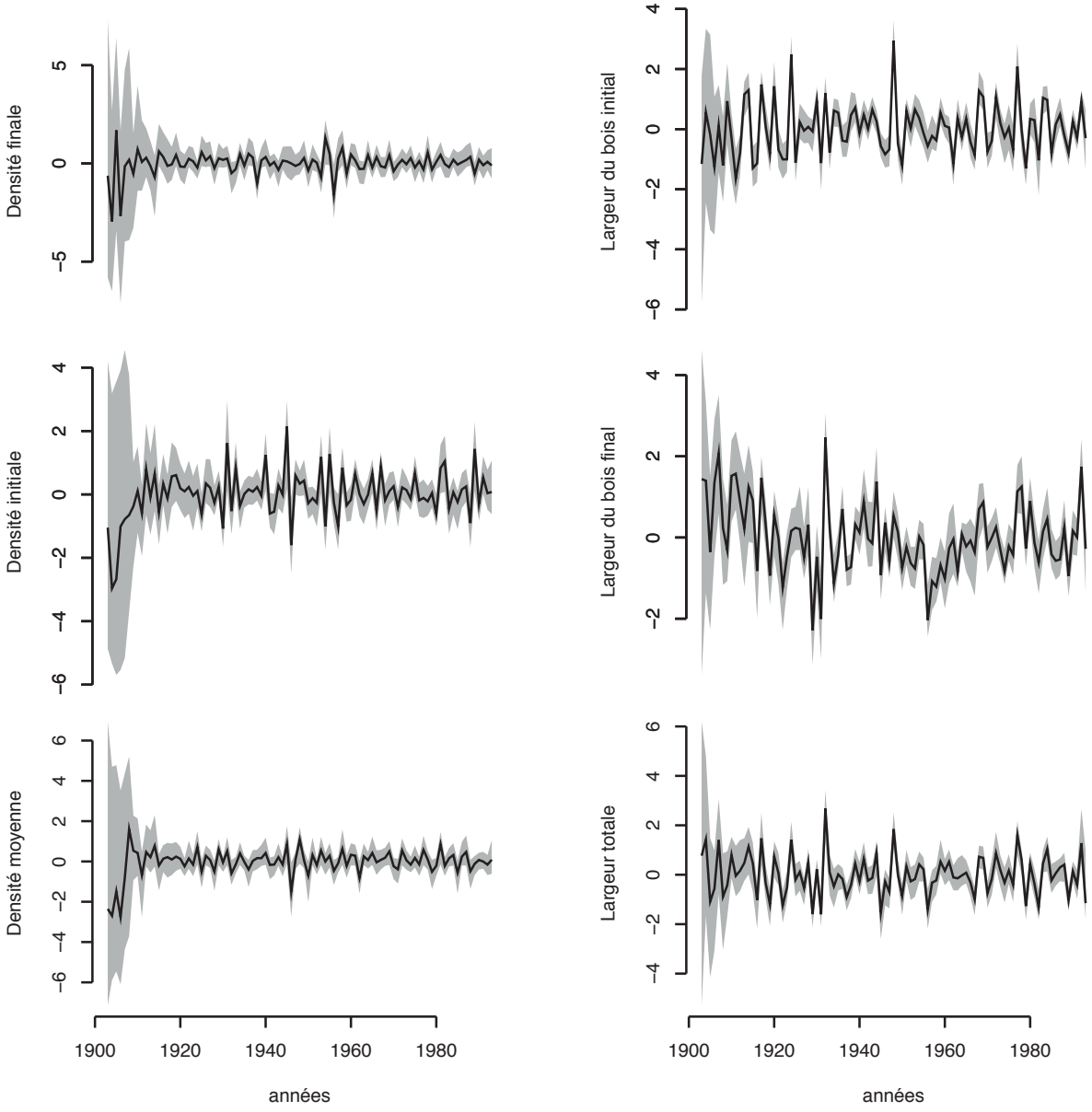

Figure 6.2 Médiane *a posteriori* pour chaque type de mesures de cernes d'arbres pour le site Les Pennes-Mirabeau. La surface grisée corresponds aux intervalles de crédibilité à 95%.

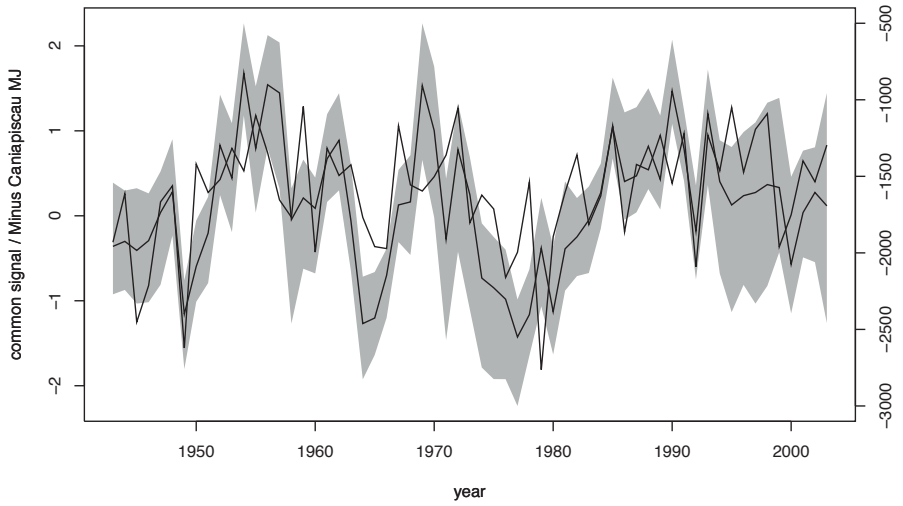

Figure 6.3 Médiane *a posteriori* obtenue à partir des aires de cernes d'arbres au Québec (courbe noire) et intervalle de crédibilité à 95%. Le courbe en pointillés correspond à l'opposée des débits moyens des mois de mai et juin observées au bassin de Caniapiscau.

Annexe A

Extraction d'un signal commun haute fréquence à partir de cernes d'arbres à l'aide d'un modèle bayésien hiérarchique

Cet article a été publié dans Climate of the Past, 5, 607–613, 2009 (http://www.clim-past.net/5/607/2009/)

Résumé : L'une des hypothèse de base de la dendrochronologie est que les cernes d'arbres peuvent être vus comme des proxies climatiques, c'est-à-dire qu'elles sont supposées contenir une information cachée sur le climat passé. D'un point de vu statistique, ce problème d'extraction peut être compris comme la recherche d'une variable cachée qui représente le signal commun contenu dans un ensemble de séries de largeurs de cernes d'arbres. Les techniques classiques utilisées en dendrochronologie ont été appliquées très largement afin d'estimer le comportement moyen de cette variable latente. Cependant, une quantification précise des incertitudes associées à la distribution de la variable cachée reste difficile car elle dépendent de l'espèce des arbres, de facteurs régionaux et des méthodes statistiques utilisées. Afin de modéliser les erreurs commisent tout au long de la procédure d'extraction nous proposons et étudions un modèle bayésien hiérarchique dont l'objectif est l'extraction d'un signal haute fréquence inter-annuel. Notre méthode est appliquée à des séries de cernes d'Epinette noires (*Picea mariana*) enregistrées au nord du Québec et comparée aux techniques de moyennes classiques utilisées par les dendrochronologues (Cook and Kairiukstis, 1992).

Clim. Past, 5, 607–613, 2009
www.clim-past.net/5/607/2009/

Climate of the Past

Extracting a common high frequency signal from Northern Quebec black spruce tree-rings with a Bayesian hierarchical model

J.-J. Boreux[1], P. Naveau[2], O. Guin[2], L. Perreault[3], and J. Bernier[1]

[1]The University of Liège, Arlon, Belgium
[2]Laboratoire des Sciences du Climat et de l'Environnement, IPSL-CNRS, France
[3]Institut de Recherche d'Hydro-Quebec, Montréal, Canada

Received: 4 January 2009 – Published in Clim. Past Discuss.: 4 March 2009
Revised: 13 August 2009 – Accepted: 15 August 2009 – Published: 14 October 2009

Abstract. One basic premise of dendroclimatology is that tree rings can be viewed as climate proxies, i.e. rings are assumed to contain some hidden information about past climate. From a statistical perspective, this extraction problem can be understood as the search of a hidden variable which represents the common signal within a collection of tree-ring width series. Classical average-based techniques used in dendrochronology have been applied to estimate the mean behavior of this latent variable. Still, depending on tree species, regional factors and statistical methods, a precise quantification of uncertainties associated to the hidden variable distribution is difficult to assess. To model the error propagation throughout the extraction procedure, we propose and study a Bayesian hierarchical model that focuses on extracting an inter-annual high frequency signal. Our method is applied to black spruce *(Picea mariana)* tree-rings recorded in Northern Quebec and compared to a classical average-based techniques used by dendrochronologists (Cook and Kairiukstis, 1992).

1 Introduction

1.1 Dendrochronology

In our changing climate, the search for accurate information about the past remains essential to understand and link past, present and future climate variations. Direct measurements are missing beyond the length of instrumental records and proxies are necessary to reconstruct chronologies of past temperatures and precipitation for a given region and/or period for which direct observations are unavailable. One of the most widely known proxy consists in tree-ring widths that possess good skill in representing climate information at the interannual to decadal time scale. An overview on this topic can be found in Cook and Kairiukstis (1992). The fundamental assumption in dendroclimatology is that a climatic signal can be hidden into tree-ring growths. Since the pioneering work of Douglass (1936), dendrochronologists have developed various methods to extract such common signals for different species. A required step in dendrochronology, called standardization, is classically needed to transform ring-width series, that are non-stationary due to tree aging processes, into relative tree-ring indices with unit mean and a constant variance. This can be accomplished by dividing each measured ring width by its expected value, i.e. the growth trend is modelled as a regression function of tree ages. Then a common signal is derived by averaging the ensemble of such tree-ring indices across series for each year. Several methods exist to calculate indices averages (e.g., Melvin et al., 2007). Esper et al. (2002) noticed that the low-frequency climate component can be highly sensitive to the standardization method. Recently, Nicault et al. (2009) proposed a neural network approach to remove the age effect and to estimate regional growth curve via explanatory variables such as tree age and their productivity. They developed a standardization procedure to preserve long-term fluctuations.

In contrast with these past methods, our goal in this paper is neither to reconstruct a series of temperatures or precipitation, nor to propose novel regression schemes based on well-chosen explanatory variables as in Nicault et al. (2009). We prefer to focus on the problem of extracting a common inter-annual high frequency signal from a given tree species and region, without regressing on possible predictants. One

Correspondence to: J.-J. Boreux
(jj.boreux@ulg.ac.be)

reason for such a choice is based on the intrinsic difficulties in linking tree-ring growths to specific explanatory variables and in interpreting these relationships. Depending on the tree species under study, it is not always clear to dendrochronologists, even today, what are the precise contributions of precipitation, temperatures, soil and hydrological characteristics, competition and other factors, to tree-ring growths. This is particularly true for black spruces in Northern Quebec. Long and reliable instrumental records of precipitation and temperature are not available for this region. By bypassing this selection, our strategy is to let the raw data "speak" for themselves. Of course, our extracted common signal could be interpreted with respect to local measurements of temperatures, precipitation and other hydroclimatological variables, whenever such information would be available. Hence, independently of the estimation step, explanatory variables could be employed in a validation scheme. To some extent, our extraction strategy could be thought within a "blind experiments" framework. The latter is classically used in medical studies to remove the experimenter bias. In dendrochronology, the "experimenter" could be viewed as the dendroclimatologist who can select his/her favorite explanatory variables (precipitation, temperatures, soil information, etc.). In our statistical modeling, neither climatological nor environmental data (besides tree rings) are used in the analysis. Consequently, our hidden signal should not be influenced by the choice of the "experimenter". Ideally, the extracted common signal should be then linked to climate variables by independent researchers who did not participate in our data analysis. Such a reasoning about design experiments is very common in many research fields (medical studies, nuclear physics, etc.) but it is rarely a topic of interest in climatology. Here we believe that our blind extraction represents an advantage because it can reduce the subjective link between the extracted common tree signal and the climate.

1.2 Bayesian Hierarchical Modeling

Assessing uncertainties in any statistical dendrochronological procedure has to be carefully addressed. To tackle this important statistical issue, we opt to work within a Bayesian Hierarchical Modeling (BHM) framework. The main idea of BHMs is to statistically model a complex 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. Schematically, the prior corresponds to a probability distribution representing knowledge or belief about an unknown quantity a priori, that is, before any data have been observed. Then, in the light of relevant data, the prior probability is updated via Bayes' theorem and becomes the posterior. For an introduction to such models, see e.g. Gelman et al. (2003). In environmental sciences, BHM has become more and more popular during the last two decades. For example, 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. Cooley et al. (2007) estimated extreme precipitation return levels by combining BHM and extreme value theory. Concerning dendrochronology, Hooten and Wikle (2007) recently investigated with a BHM shifts in the spatio-temporal growth dynamics of shortleaf pine.

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

What is the interest of BHMs for dendrochronologists? The choice of the Bayesian paradigm allows the use of unobserved variables in a hierarchical structure, while easily modeling uncertainties at each different level of this structure. In particular, expert information can be integrated via probability densities (the priors). In other words, past knowledge, even diffuse or imperfect, from scientists can be taken advantage of. More precisely, each parameter of a Bayesian hierarchical model can be viewed as a random variable and hence, a dialogue with dendrochronologists can be engaged to set the prior distribution of this random variable. If the expert has no prior knowledge then the distribution is set to be very wide (a diffuse prior), otherwise the uncertainty of the parameter can be reduced by using knowledge from past studies. In a following step, the incoming data (tree-ring areas here) are used to update all the parameters of our model. The Bayes' theorem provides the mathematical formula to perform this updating, i.e. to derive the posterior distributions. In summary, one can see the above Bayesian strategy as an assembly of elementary parts. Its modular character makes it possible to replace prior uncertainty knowledge (set by experts) by posterior distributional information, throughout the incoming data. In this sense, it is an evolutionary construction.

The paper is organized as follows. Section 2 describes the data and the regional characteristics of the site under study. The details of our latent model are presented in Sect. 3. A short discussion about our application is proposed in Sect. 4. Perspectives are given in the conclusion.

2 Data and region of interest

To extract a common tree signal, the dendrochronologist has to make a series of important decisions about the tree species, the region of interest and the sampling procedure

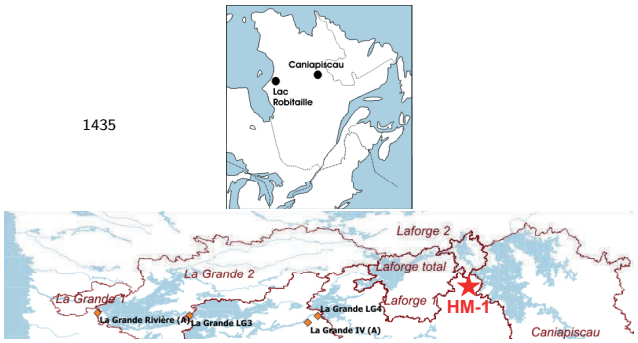

Fig. 1. The upper panel corresponds to Northern Quebec. The lower panel is a zoom near the Caniapiscau region and the red star called HM-1 represents the site from which fifteen trees have been sampled.

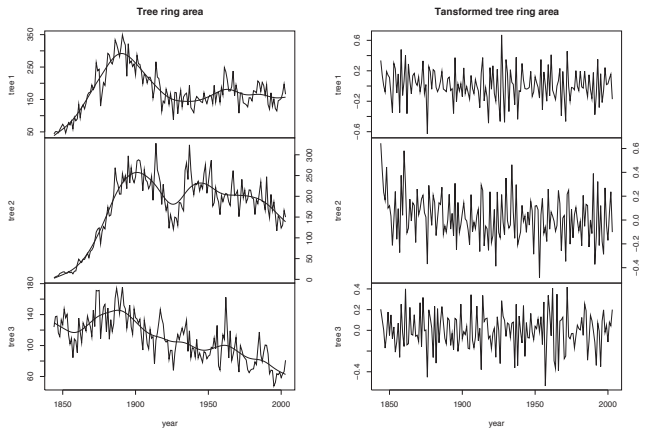

Fig. 2. Temporal behavior of three ring area time series (randomly chosen from a set of fifteen trees) over the period 1846–2003. The left panels correspond to the measured tree ring areas with a fitted cubic spline trend. The right panels indicate the log difference of the same ring areas, see Eq. (1).

(e.g., George et al., 2008). Concerning the region choice, Hydro-Quebec, one of the founding agencies involved in this project, has had a strong interest in Northern Quebec because of its hydro-electrical capacities. With this constraint in mind, a mesic site, i.e. with a moderate supply of moisture, close to lake Hurault (54°15′ N, 70°47′ W) was chosen, see the red star called HM-1 in the right panel of Fig. 1. This site has the advantages to belong to a climatic homogenous region and of being far away from most human activities. The black spruce *(Picea mariana)* was selected because it is a widespread species in Northern Quebec. Fifteen trees covering a period of 158 years were sampled. These trees were carefully chosen by an expert who removed singular individuals (sick trees, dominated trees, etc.). Each tree provided a ring width series from which annual growth ring areas were estimated. This transformation from ring width to ring area diminishes the geometrical effect impact, basically older trees have thiner rings. The last ring of all sampled trees, albeit missing rings, should correspond to the calendar year. Hence the youngest tree determines the common period length of all trees.

To illustrate the type of dendrochronological times series under study, Fig. 2 shows the temporal behavior of three ring area series, randomly chosen from fifteen trees. The right panels represent those three ring area series. From these three right panels, it is clear that each tree has a different trend and it seems difficult to find a common hidden signal in the low frequency domain. In addition, the variability around the cubic-spline trend in the right panels seems to be stronger after 1880 for trees 1 and 2. This example illustrates the high complexity of separating tree ring areas into their individual growth component and their common hidden component in the low frequency part of these signals. Different techniques (e.g. working with residuals after fitting a reference growth curve) exist to deal with this important issue. In this paper we do not address directly this issue. Instead we apply a simple non-parametric transformation to remove trends and to work with stationary time series. This implies that we only focus on inter-annual high frequencies in tree rings. The simple non-parametric transformation is defined as

$$Y_{ts} = \log X_{ts} - \log X_{t-1,s}, \quad (1)$$

with t=2, . . ., T and s=1, . . ., S and where X_{ts} represent the measured annual ring area produced during year t by tree s and T is the length of the temporal sequence and S the number of trees. Transformation (1) is extensively used in finance (Gencay et al., 2002). Besides its simplicity of implementation, this log-difference has the advantage of removing any smooth (i.e. polynomial) trend, see the right panels of Fig. 2. In addition the change of variability aforementioned in trees 1 and 2 is less pronounced in the right panels of Fig. 2. The drawbacks of using (1) are that, if present, the low frequency part of a possible common signal has been removed and that the time unit t in Y_{ts} does not correspond to a year anymore but to a one-year increment. The latter has to be kept in mind when interpreting our results. The former implies that our model described below will only focus on the high frequency part of a possible common signal.

Concerning the interpretation of Y_{ts} defined as a log-difference between two consecutive ring area values, the following simple facts need to be recalled. Whenever the relative ratio of two inter-annual consecutive ring areas from the same tree is close to one, then Y_{ts} is close to zero. If this relative ratio is very large (i.e. the ring area from year t is much larger than the one formed during year $t-1$), then Y_{ts} has to strongly positive. Conversely, a negative Y_{ts} represents a large decrease in ring areas between two consecutive years. As exemplified by Fig. 3, working with Y_{ts} instead of the raw ring areas X_{ts} allows us to remove long-term trends, to

focus on the inter-annual relative variability and to work with time series that can be assumed to be stationary and Gaussian. One drawback is that we have lost the absolute value of X_{ts}, i.e. working with the couple $(X_{ts}, X_{t-1,s})$ is equivalent to analyzing the couple $(aX_{ts}, aX_{t-1,s})$ for any $a>0$, independently of the value of a. Keeping in mind this drawback and those advantages, the correlation meaning in Y_{ts} and Z_t can be viewed as the short term memory in the relative log-transfom rate between two consecutive ring areas.

Before closing this section we would like to emphasize that our detrending choice represented by (1) is not unique and others techniques could be used to provide stationary signals. For example, we could have worked with the residuals obtained from the cubic spline fit shown in the left panels of Fig. 2.

3 An additive latent model

The random variable Y_{ts} defined by Eq. (1) is assumed to follow an additive model with a latent variable Z_t

$$Y_{ts} = \mu_s + \lambda_s Z_t + \epsilon_{st}, \quad (2)$$

with $t=2, \ldots, T$ and $s=1, \ldots, S$, and where μ_s corresponds to the mean level of tree s, Z_t represents the hidden regional signal common to all trees and ϵ_{st} describes local fluctuations of tree s during year t. Tree-to-tree variations captured by ε_{st} can be due to reserves accumulated by tree s and other factors that are not directly linked to environmental causes, the latter ones should be represented by Z_t. For each calendar year t, the product $\lambda_s Z_t$ measures how the hidden factor Z_t contributes to the growth of tree s. We assume that Z_t and ε_{st} are independent processes. With respect to the BHMs described in Sect. 1.2, the random variables Y_{ts} corresponds to the data layer and Z_t belongs to the process layer.

Before describing the probabilistic structure within Z_t and ϵ_{st}, it is advantageous to rewrite model (2) with obvious vectorial notations

$$\mathbf{Y}_s = \mu_s \mathbf{1} + \lambda_s \mathbf{Z} + \boldsymbol{\epsilon}_s, \quad (3)$$

where $\mathbf{1}$ is the unit vector of length $T-1$. Each tree s may have a temporal memory that should depend on the hydrological stress or other conditions that are particular to this tree location. Although these tree-to-tree effects can be complex, to keep the inference simple and the risk of over-parametrization low, we opt for a simple zero-mean Gaussian auto-regressive process of order one for $\boldsymbol{\epsilon}$, i.e. $\boldsymbol{\epsilon}_s = \phi_s \boldsymbol{\epsilon}_{-s} + \mathbf{V}_s$. The notation $\boldsymbol{\epsilon}_{-s}$ corresponds to $\boldsymbol{\epsilon}_s$ shifted by one year, i.e. $\boldsymbol{\epsilon}_{-s} = \left(\varepsilon_{s1}, \ldots, \varepsilon_{s(T-1)}\right)'$, ϕ_s represents the auto-regressive coefficient of tree s, and the random vector $\boldsymbol{V_s}$ of length $T-1$ follows a zero-mean multivariate Gaussian distribution with *precision* $\eta_s \times \mathbf{I}$ where $\mathbf{I}$ is the identity matrix of size $T-1$. In other words, all components of vector $\boldsymbol{V_s}$ correspond to a standardized normal independent random noise.

To allow the common regional factor Z_t to have a short year-to-year memory, we assume that the latent Z_t can be modeled as a zero-mean Gaussian auto-regressive process of order one, i.e. $\mathbf{Z} = \rho \mathbf{Z}_- + \mathbf{U}$ where $\mathbf{Z}_- = \left(Z_1, \ldots, Z_{(T-2)}\right)'$ and $\boldsymbol{U}$ represents a zero-mean multivariate normal vector of length $T-1$ with *precision* $\tau \times \mathbf{I}$.

Our full model counts $2+4\,S$ parameters, namely (ρ, τ) and $\boldsymbol{\theta}_s = (\lambda_s, \mu_s, \phi_s, \eta_s)$ with $s=1, 2, \ldots, S$. We assume that the priors distributions $[\rho, \tau]$, $[\boldsymbol{\theta}_1], \ldots,$ and $[\boldsymbol{\theta}_S]$ are mutually independent. By writing the joint distribution as a product of conditional distributions with a marginal distribution, the prior for (ρ, τ) can take the following form $[\rho, \tau] = [\rho|\tau]\,[\tau]$. In a classical way, we assume that the precision parameter τ follows a gamma distribution with two hyperparameters that must be fixed to reflect prior beliefs. In our application, a diffuse prior is chosen by setting the two gamma parameters to zero.

The choice of the auto-regressive coefficient prior $[\rho|\tau]$ is more delicate. Classically, it is assumed that auto-regressive processes are a priori stationary. This implies that auto-regressive coefficients have to belong to the interval $[-1, 1]$. As Bayesian statisticians, we defend the idea that the underlying characteristics of the hidden process Z_t should not be imposed but arise form the data via the Bayes' rule or via prior knowledge. For this reason, we assume that $[\rho|\tau]$ follows a zero-mean Gaussian distribution with a precision proportional to τ. This multiplicative factor must be fixed between zero and one, mainly to degrade the precision a little. In our application, we work with a diffuse prior by equaling the multiplicative factor to zero.

Concerning the prior of the random vector $\boldsymbol{\theta}_s = (\lambda_s, \mu_s, \phi_s, \eta_s)$, we assume conditional independence, i.e. $[\boldsymbol{\theta}_s] = [\lambda_s|\eta_s]\,[\mu_s|\eta_s]\,[\phi_s|\eta_s]\,[\eta_s]$ where the variable η_s follows a gamma distribution with two hyperparameters (set to zero in our application). The distributions $[\lambda_s|\eta_s]$, $[\mu_s|\eta_s]$ and $[\phi_s|\eta_s]$ are assumed to be diffuse Gaussian priors in this paper. As for the auto-regressive coefficient of Z_t, this means that the auto-regressive coefficient of ϵ_{st} are not a priori assumed to be in the interval $[-1, 1]$.

To compute the posteriors of the latent vector Z_t and of the $2+4\,S$ parameters, we implement the Gibbs sampler described in the Appendix. The Bayesian inference was carried out with the open source R statistical software (R Development Core Team, 2009). Our programs are available upon request.

4 Results and discussion

The solid line in Fig. 3 shows the estimated posterior median value of the common factor Z_t over the period 1846-2003. The shaded area corresponds to the 90% credible regions (CR). Note that the value of Z_t and λ_s are estimated up to a constant because it is always possible in (2) to multiply Z_t by a constant and divide the λ_s by the same constant

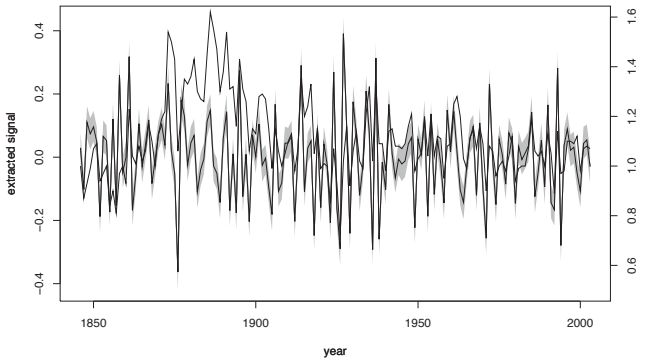

Fig. 3. The solid line corresponds to the estimated posterior median value of the common signal Z_t from (2) over the period 1846–2003. The shaded area corresponds to the 90% credible regions. The dashed line represents the so-called tree-growth index which is an arithmetic mean of ratios over all trees. Each ratio is derived by dividing ring thickness over a temporally smoothed tree signal for each tree (e.g., Cook and Kairiukstis, 1992).

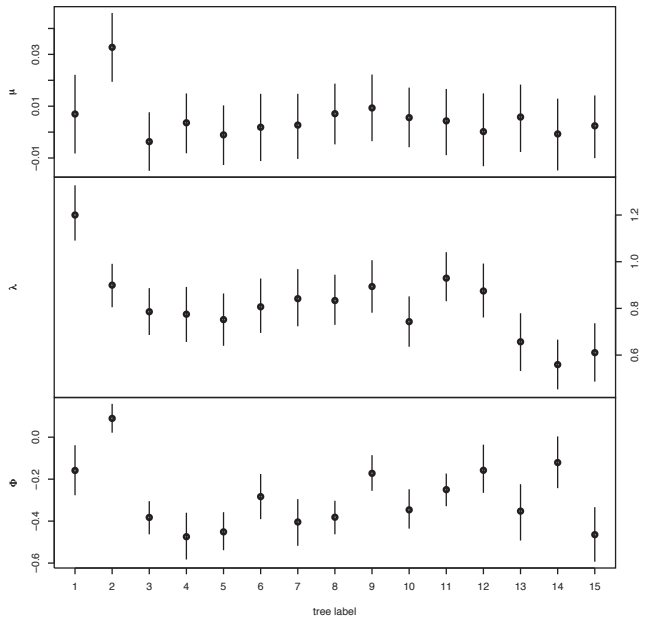

Fig. 4. For each of the fifteen trees, the posterior mean of μ_s, λ_s and ρ_s are represented by circles in the top, middle and bottom panels, respectively. Vertical bars correspond to the 90% credible regions.

without being able to identify this multiplicative factor. In Fig. 3, we compare our BHM results with a classical technique employed by dendrochronologists. The output of this procedure is represented by the dashed line, a so-called tree-growth index which is an arithmetic mean of ratios over all trees. Each ratio is derived by dividing ring thickness over a temporally smoothed tree signal for each tree (e.g., Cook and Kairiukstis, 1992). Up to a constant (this explains the two different scales for the y-axis), the classical tree-growth index behaves similarly to Z_t by staying in the CR over a long time period. From about 1875 to 1900, there is a discrepancy between Z_t and the classical tree-growth index, the latter producing higher values during this period. Although fairly localized in time, this difference indicates that this classical technique by not providing confidence intervals shows its limitations. Still, this comparison between the two extracted signals makes us believe that our BHM approach is capable of providing meaningful outputs for dendrochronologists because they do not contradict past results and offer another statistical approach to this community of scientists.

Concerning the memory within Z_t, the posterior distribution of the autoregressive coefficient ρ indicates a negative correlation because its 25%, 50% and 75% posterior quantiles are equal to -0.40, -0.36 and -0.32, respectively. In addition to CRs, our methods allow the practitioner to derive a finer analysis of her/his tree ring data. For example, an analysis tree-by-tree can be undertaken. For each of the fifteen trees, Fig. 4 displays the posterior mean and 90% CRs of the parameters μ_s, λ_s and ϕ_s, respectively. The mean posterior value of μ_s mostly oscillates around zero for all trees. Overall, each tree but tree 2 appears to have a mild negative inter-annual memory, all autoregressive coefficients (but tree 2) shown in the bottom panel of Fig. 4 have a ϕ_s posterior median around -0.4. The central panel clearly points out tree 1 which seem to contribute the most to Z_t.

To check the quality of our estimation, Fig. 5 displays for trees 1, 2 and 3 (shown in Fig. 2), the observed Y_{ts} versus the naive estimate $\hat{Y}_{ts}$ obtained by plugging our median posterior parameter values in (2) without noise. As expected, the relationships appear to be linear. The same result holds for the other trees.

5 Conclusions

To summarize our findings, we have implemented a hierarchical Bayesian model to estimate a common hidden signal in high frequency component of trees. This latent signal should be viewed as a representation of the regional pressure affecting black spruce trees over our studied area in Northern Quebec. The hierarchical structure provides another way to model the temporal structure associated to tree memories at the regional and tree-to-tree levels. This model attempts to quantify the contribution of a high frequency common hidden signal to each tree growth. This could help selecting trees with regard to a possible climatological interpretation in a reconstruction context. Compared with past approaches, our hidden signal was strongly correlated to the estimate obtained with the most traditional procedure. This confirms a past method derived by dendrochronologists, while bringing the benefits of a BHM approach. As a further step in this analysis, it would be of interest to integrate low frequency

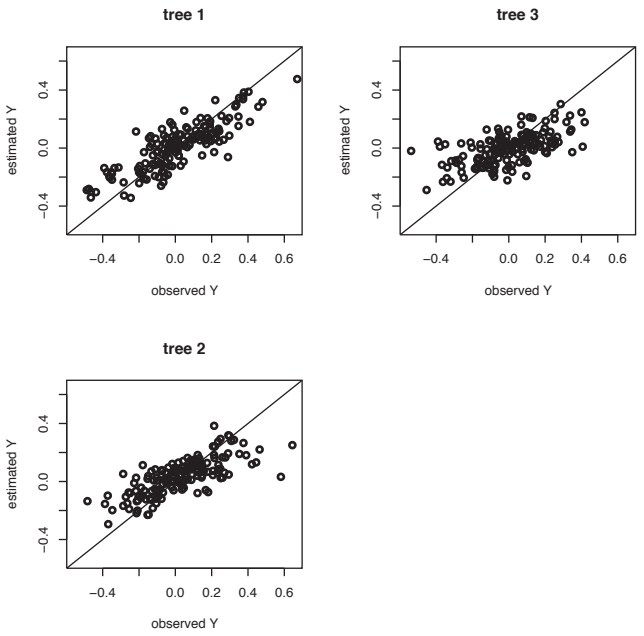

Fig. 5. For each of the three randomly chosen trees described in Fig. 2, the observed Y_{ts} versus its estimator from model (2) is plotted.

in Eq. (2). One possibility is to bypass transformation (1) by making the term μ_s in (2) varying in time. For example, μ_{ts} could be modeled by Bayesian splines. Besides the complexity of such an approach, the main difficulty is our limited sample size (fifteen trees). Another aspect is the handling of missing values and consequently avoid the limitation brought by the age of the youngest tree. In addition, ongoing field trips should provide a much larger sample of tree rings and allows us to extend our BHM procedure in future research. In this context, our present work should rather be viewed as an addition of a simple statistical procedure to the mathematical toolbox of dendroclimatologists rather than a comprehensive study of black spruce trees in Northern Quebec.

The combination of the additive model described by (2) and the Bayesian paradigm allows the practitioner to easily generate the full posterior distribution of the hidden signal, and consequently it is possible to simulate realizations of the relative log-transform ratio between two consecutive ring areas. Such simulations could help simulating the way tree growth in response to climatic forces that drives the common inter-annual variations. Another interesting perspective of the Bayesian approach resides in the possibility to compute predictive posterior densities for future years. Since we can derive the posterior density of the extracted signal, the predictive posterior $[z_{t+1}|\mathbf{y}_t, \theta]$ for an unobserved year $t+1$ can obtained by computing the hidden state posterior density at time $t+1$.

Appendix A

Gibbs sampling procedure

- *Step 0:* Initialize the vector $\mathbf{Z}|\rho, \tau, z_0$ of length T from multivariate normal distribution with mean $z_0\mathbf{B_0}$ and variance $\tau^{-1}\mathbf{B}\mathbf{B}^T$ where

$$\mathbf{B}_0^t \equiv \left[\rho, \rho^2, \ldots, \rho^T\right] \text{ and } \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ \rho & 1 & 0 & \ldots & 0 \\ \rho^2 & \rho & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^T & \rho^{T-1} & \ldots & \rho & 1 \end{bmatrix}$$

- *Step 1:* Draw the precision $\tau|\mathbf{z}, z_0, \rho$ from a gamma distribution with parameters $a+\frac{T+1}{2}$ and $(b+\frac{1}{2}\sum_{t=1}^{T}(z_t-\rho z_{t-1})^2)^{-1}$ where a and b are prior parameters (e.g. a=b=0)

- *Step 2:* Draw the correlation coefficient $\rho|\mathbf{z}, z_0, \tau$ from a normal distribution with mean $\frac{k_\rho m_\rho + \sum_{t=1}^{T} z_{t-1}z_t}{k_\rho + \sum_{t=1}^{T} z_{t-1}^2}$ and precision $\left[\tau\left(k_\rho + \sum_{t=1}^{T} z_{t-1}^2\right)\right]^{-1/2}$ where k_ρ and m_ρ are prior parameters (e.g. k_ρ=m_ρ=0)

- *Step 3:* For s=1, 2, ..., S.

  - *Step 3.1:* Let ψ_s represent μ_s, λ_s or φ. Draw $\psi_s|\mathbf{y}_s, \mathbf{z}, z_0, y_{0s}$ from a normal distribution with mean $\frac{k_\psi m_\psi + \mathbf{f}_s^T\mathbf{g}_s}{k_\psi + \mathbf{g}_s^T\mathbf{g}_s}$ and correlation $[k_\psi + \mathbf{g}_s^T\mathbf{g}_s\eta_s]^{-1/2}$ where k_ψ et m_ψ are prior parameters which are invariant from tree to tree (e.g. k_ψ=m_ψ=0). The vectors $\boldsymbol{f_s}$ and $\boldsymbol{g_s}$ depend on handling parameters.
  - *Step 3.2:* Draw precision $\eta_s|\mu_s, \lambda_s, \varphi_s, \mathbf{y}_s, \mathbf{z}, z_0, y_{0s}$ from a gamma distribution with parameters $c+\frac{T+3}{2}$ and $[d+0.5\mathbf{v}'\mathbf{v}]^{-1}$ where c and d are prior parameters (e.g. c=d=0)

- *Step 4.:* Draw vector $\mathbf{U}|\tau, \eta_s, \mathbf{L}_s, \mathbf{R}_s$ from a multivariate normal distribution with mean ω and covariance Ω^{-1} and set $\mathbf{z}$=$z_0\mathbf{B_0}$+$\mathbf{Bu}$. The mean ω and matrix Ω^{-1} relate vector $\boldsymbol{L}$ and matrix $\mathbf{R}$ which depend on previous parameters.

- *Step 5:* Return to step 1.

Note that when the Gamma hyper parameters are theoretically equal to zero, this means that they are set to a very small value in real computations, e.g. a=b=.001 in our case.

Acknowledgements. This study is founded by the Hydro-Quebec, the CRSNG and OURANOS. The authors would like to thank Joel Guiot and Delphine Grancher for interesting discussions about the statistical aspects of dendrochronology, Yves Bégin (INRS-ETE) and his colleagues for their data and their scientific expertise and James Merleau for his helpful suggestions. The ANR Escarcel, AssimilEx and AQCWA projects and the MAIF foundation are also acknowledged by Philippe Naveau and Ophélie Guin. Finally, this paper is dedicated to the memory of Dominique Joly.

Edited by: H. Goosse

Publication of this paper was granted by
EDD (Environnement, Développement Durable)
and INSU (Institut des Sciences de l'Univers) at CNRS.

References

Berliner, L., Wikle, C., and Cressie, N.: Long-lead prediction of Pacific SSTs via Bayesian Dynamic Modeling, J. Climate, 13, 3953–3968, 2000.

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

Cooley, D., Nychka, D., and Naveau, P.: Bayesian spatial modeling of extreme precipitation return levels, J. Am. Stat. Assoc., 102(479), 824–840, 2007.

Cook, E. R. and Kairiukstis, A.: Methods of dendrochronology, Kluwer Academic Publishers, 1992.

Douglass, A. E.: Climatic cycles and tree-growth, Carnegie Institution of Washington publication 289, vol. 3, 1936.

Esper, J., Cook, E. R., and Schweingruber, F. H.: Low-Frequency Signals in Long Tree-Ring Chronologies for Reconstructing Past Temperature Variability, Science, 295, 2250–2254, 2002.

Gencay, R., Selcuk, F., and Whitcher, B.: An Introduction to Wavelets and Other Filtering Methods in Finance and Economics, Academic Press, San Diego, 2002.

Gelman, A., Carlin, J., Stern, H., and Rubin, D.: Bayesian Data Analysis, 2nd edn., Chapman & Hall, 2003.

George, S., Meko, D. M., and Evans, M. E.: Regional tree growth and inferred summer climate in the Winnipeg River basin, Canada, since AD 1783, Quaternary Res., 70, 158–172, 2008.

Gelman, A.: Inference and monitoring convergence, in: Markov Chain Monte Carlo in Practice, edited by: Gilks, W. R., Richarson, S., and Spiegelhalter, D. J., Chapman and Hall, 1996.

Hooten, M. B. and Wikle, C. K.: Shifts in the spatio-temporal growth dynamics of shortleaf pine, Environ. Ecol. Stat., 14(3), 207–227, 2007.

Nicault, A., Guiot, J., Edouard, J.-L., and Brewer, S.: Preserving long-term fluctuations in standardisation of tree-ring series by Adaptive Regional Growth Curve (ARGC), Dendrochronologia, in press, 2009.

Melvin, T. M., Briffa, K. R., Nicolussi, K., and Grabner, M.: Time-varying-response smoothing, Dendrochronologia, 25, 65–69, 2007.

R Development Core Team: R: A Language and Environment for Statistical Computing, ISBN 3-900051-07-0, http://www.R-project.org, R Foundation for Statistical Computing, 2009.

Calculs pour les reconstructions climatiques bayésiennes

Soit $\boldsymbol{P} = (\boldsymbol{P}_O, \boldsymbol{P}_N)^T$ une série temporelle de mesures pour une variable climatique quelconque. $\boldsymbol{P}_N$ correspond aux données enregistrées sur la période récente et $\boldsymbol{P}_O$ aux valeurs passées que l'on cherche à reconstruire. De même $\boldsymbol{f} = (\boldsymbol{f}_O, \boldsymbol{f}_N)^T$ correspond au signal commun extrait à partir d'un ensemble de séries de cernes d'arbres pour la période récente et passée. On suppose la relation entre $\boldsymbol{P}$ et $\boldsymbol{f}$ linéaire ce qui nous donne la régression suivante,

$$\begin{pmatrix} \boldsymbol{P}_O \\ \boldsymbol{P}_N \end{pmatrix} = a \begin{pmatrix} \mathbf{1}_m \\ \mathbf{1}_{n-m} \end{pmatrix} + b \begin{pmatrix} \boldsymbol{f}_O \\ \boldsymbol{f}_N \end{pmatrix} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n). \tag{B.1}$$

On cherche à estimer les paramètres a, b, $\boldsymbol{f}$, $\boldsymbol{P}_O$ et σ^2 à l'aide de l'estimation bayésienne.

A partir de l'équation (B.1) on peut déduire la vraisemblance de notre modèle. En effet, on sait que $\boldsymbol{P}$ suit une loi normale multivariée

$$[\boldsymbol{P}|a, b, \boldsymbol{f}, \sigma^2] = (2\pi)^{-n/2}\sigma^{-n} \exp(-\frac{1}{2\sigma^2}(\boldsymbol{P} - a\mathbf{1}_n - \boldsymbol{f})^T(\boldsymbol{P} - a\mathbf{1}_n - \boldsymbol{f})).$$

On en déduit donc la vraisemblance de notre modèle de régression linéaire,

$$\begin{aligned}[\boldsymbol{P}_N|a, b, \boldsymbol{f}, \boldsymbol{P}_O, \sigma^2] &= (2\pi)^{-n/2}\sigma^{-n} \exp[-\frac{1}{2\sigma^2}(\boldsymbol{P}_N - a\mathbf{1}_{n-m} - \boldsymbol{f}_N)^T(\boldsymbol{P}_N - a\mathbf{1}_{n-m} - \boldsymbol{f}_N) \\ &\quad -\frac{1}{2\sigma^2}(\boldsymbol{P}_O - a\mathbf{1}_n - \boldsymbol{f}_O)^T(\boldsymbol{P}_O - a\mathbf{1}_n - \boldsymbol{f}_O)].\end{aligned}$$

Distribution *a posteriori* de a Par définition on a la distribution *a posteriori* de a

$$[a|b, \boldsymbol{f}, \boldsymbol{P}_O, \boldsymbol{P}_N, \sigma^2] \propto [\boldsymbol{P}_N|a, b, \boldsymbol{f}, \boldsymbol{P}_O, \sigma^2][a]$$

On suppose que a suit *a priori* une loi normale

$$a \sim \mathcal{N}(\mu_a, \tau_a).$$

Moyennant quelques calculs assez simples, on obtient la distribution *a posteriori* suivante

$$a|b, \boldsymbol{f}, \boldsymbol{P}_O, \boldsymbol{P}_N, \sigma^2 \sim \mathcal{N}(\mu_a^*, \tau_a^*),$$

avec

$$\begin{aligned}\tau_a^* &= (1/\tau_a + 1/\sigma^2)^{-1} \\ \mu_a^* &= \tau_a^*(\mu_a/\tau_a + 1/\sigma^2(\boldsymbol{P} - b\boldsymbol{f})^T\mathbf{1}_n).\end{aligned}$$

Distribution *a posteriori* de b De la même manière que pour a, on suppose que b suit *a priori* une loi normale $\mathcal{N}(\mu_b, \tau_b)$ et par le même type de calculs que précédemment on obtient la distribution *a posteriori* de b

$$b|a, \boldsymbol{f}, \boldsymbol{P}_O, \boldsymbol{P}_N, \sigma^2 \sim \mathcal{N}(\mu_b^*, \tau_b^*),$$

avec

$$\begin{aligned}\tau_b^* &= (1/\tau_b + 1/\sigma^2\boldsymbol{f}^T\boldsymbol{f})^{-1} \\ \mu_b^* &= \tau_b^*(\mu_b/\tau_b + 1/\sigma^2\boldsymbol{f}^T(\boldsymbol{P} - a\mathbf{1}_n)).\end{aligned}$$

Distribution *a posteriori* de $\boldsymbol{f}$ Par définition on a la distribution *a posteriori* de $\boldsymbol{f}$

$$[\boldsymbol{f}|a, b, \boldsymbol{P}_O, \boldsymbol{P}_N, \sigma^2] \propto [\boldsymbol{P}_N|a, b, \boldsymbol{f}, \boldsymbol{P}_O, \sigma^2][\boldsymbol{f}]$$

On suppose que la distribution de $\boldsymbol{f}$ est proche d'une loi normale multivariée, distribution choisie comme loi *a priori*

$$\boldsymbol{f} \sim \mathcal{N}_n(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f).$$

On en déduit la distribution *a posteriori* de $\boldsymbol{f}$

$$\boldsymbol{f}|a, b, \boldsymbol{P}_O, \boldsymbol{P}_N, \sigma^2 \sim \mathcal{N}_n(\boldsymbol{\mu}_f^*, \boldsymbol{\tau}_f^*),$$

avec

$$\begin{aligned}\boldsymbol{\Sigma}_f^* &= (b^2/\sigma^2 \boldsymbol{I}_n + \boldsymbol{\Sigma}_f^{-1})^{-1} \\ \boldsymbol{\mu}_f^* &= \boldsymbol{\Sigma}_f^*(\boldsymbol{\mu}_f \boldsymbol{\Sigma}_f^{-1} + b/\sigma^2(\boldsymbol{P} - a\mathbf{1}_n)).\end{aligned}$$

Distribution *a posteriori* de $\boldsymbol{P}_O$ On suppose que la loi *a priori* de $\boldsymbol{P}_O$ est une loi normale multivariée,

$$\boldsymbol{P}_O \sim \mathcal{N}_m(\boldsymbol{\mu}_O, \boldsymbol{\Sigma}_O).$$

La distribution *a posteriori* de $\boldsymbol{P}_O$ est donc

$$[\boldsymbol{P}_O|a, b, \boldsymbol{f}, \sigma^2, \boldsymbol{P}_N] \propto [\boldsymbol{P}_N|a, b, \boldsymbol{f}, \boldsymbol{P}_O, \sigma^2][\boldsymbol{P}_O],$$

1470 c'est-à-dire que

$$\boldsymbol{P}_O|a, b, \boldsymbol{f}_O, \sigma^2 \sim \mathcal{N}_m(\boldsymbol{\mu}_O^*, \boldsymbol{\Sigma}_O^*)$$

avec

$$\begin{aligned}\boldsymbol{\Sigma}_o^* &= \sigma^2(\mathbf{I}_m + \sigma^2 \boldsymbol{\Sigma}_O^{-1})^{-1} \\ \boldsymbol{\mu}_O^* &= \boldsymbol{\Sigma}_O^* \boldsymbol{\Sigma}_O^{-1} \boldsymbol{\mu}_O + (1/\sigma^2)\boldsymbol{\Sigma}_O^*(a\mathbf{1} + b\boldsymbol{f}_O).\end{aligned}$$

Annexe C

Liste des publications

V. Bellassen, G. Le Maire, O. Guin, J.F. Dhôte, P. Ciais, and N. Viovy. Modelling forest management within a global vegetation model – part 2 : Model validation from a tree to a continental scale. *Ecol. Model.*, 222 :57–75, 2011.

J.-J. Boreux, P. Naveau, O. Guin, L. Perreault, and J. Bernier. 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, 2009.

O. Guin, J. Merleau, and P. Naveau. Bayesian variables selection for generalized additive models applied to climatic reconstructions. *Rapport technique*, 2011a.

O. Guin, P. Naveau, and J.-J. Boreux. Extracting hidden trends in tree rings with a semi-parametric bayesian hierarchical model. *Soumis*, 2011b.