A Climate-sensitive Analysis of Lodgepole Pine Site Index in Alberta

Xiao Jing Guo

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Science at
Concordia University
Montreal, Quebec, Canada

September 2005

Your file Votre référence
ISBN: 0-494-10212-8
Our file Notre référence
ISBN: 0-494-10212-8

ABSTRACT

A Climate-sensitive Analysis of Lodgepole Pine Site Index in Alberta

Xiao Jing Guo

Growth and yield models in forest management are derived from past observations, assuming implicitly that future growth conditions will be similar. Local observations of apparent changes in site index (SI: defined as the top height at 50 years breast height age) of lodgepole pine in Alberta during the 20th century raise serious questions about validity of this assumption.

As part of a joint program on climate change in Alberta by Canadian Forest Service and Laval University, this thesis aims at investigating the impacts from climate change on the site index based on a process-based forest growth model, StandLEAP. Data processing techniques, nonlinear regression and time series analysis aree conducted to obtain the necessary models.

The research involves the calibration of a climate-sensitive site index model. This model is then used to explain SI variability between 1901 and 2000 for each plot. A significant SI increment of 4 mm/year appears on average. This change is significant over 100 to 200 years, the time period used to check that the projected cut can be sustained by the forest over the long term. Over this time period, stand SI will change from .4 to .8 m, more than half of a site index class.

The results suggest that climate is an important factor affecting lodgepole pine productivity in Alberta, and have implications for future forest management under a warmer climate.

Acknowledgments

The research project reported in this thesis would not have been possible without the generous help of many persons, of which I am grateful and wish to express my gratitude. If I seem to have forgotten someone, please blame that on my forgetfulness, not lack of gratitude.

First of all, I would like to thank Dr. Xiaowen Zhou, my thesis director. His support and help were absolutely invaluable through my study at Concordia university. His great ideas in directing this thesis have inspired my interest, which will have a long-time influence in my future work.

I sincerely thank my thesis co-director Dr. Frédéric Raulier at Laval University. He has been my supervisor, and keeping me on the track, for more than two years with his brain-storming advices. Most of all I want to thank him for giving me some excellent ideas that turned out to be valuable contributions to the thesis.

Thanks to Dr. Jose Garrido and Dr. T.N. Srivastava, for sharing their excellent knowledge in class.

Thanks to Dr. Pierre Bernier, Ecoleap program leader and also my supervisor, for taking time out of his busy schedules, to offer suggestions and advices for my work.

I would also like to express my gratitude to Dr. David Price for writing and getting funded the initial project proposal, of which the present thesis is a substantial part. I would like to thank Dr. Dan McKenney for giving me access to his Canada-wide monthly climate data over the last century; and Marty Siltanen for the help in understanding the PSP data and patience in answering all my questions. I also acknowledge the Alberta Sustainable Resource Development for giving me access to their valuable provincial Permanent Sample Plot data.

Cordial thanks to Ann Marie Agnew at Concordia university, graduate program assistant, for her continuous help and encouragement during the writing of this thesis, and Marie-Claude Lambert, Statistician at CFS who has been offering her generous help to my work.

Finally, I would give my deep gratitude to my parents and my brother Ru Shan Guo, for their love and care, and especially for having brought me up to who I am today. Thanks Patrick Akedjo for his friendship and all the suggestions and help he has kindly offered to this work.

The study was supported by Foothills Model Forest, Alberta Forest Growth and Yield Association and Canadian Forest Service at Northern and Laurentian Forestry Centers (ECOLEAP and ECOLEAP-WEST projects).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Forest modelling

In 1889, the great forester Endres wrote, concerning the formulation of growth functions: "This may be an interesting mathematical exercise, but such an attempt to bring mathematics into the service of forestry science is not an enrichment of the results of scientific research."(Prodan 1968).

This opinion has been reversed in part by the development of detailed knowledge and the accumulation of vast quantities of data. The evaluation of complicated phenomena and biological processes may be carried out by methods of mathematics or statistics. The example of Malthus' law shows, however, that an attempt to solve biological problems by methods of mathematics alone might lead to wrong conclusions. Malthus (1766 - 1834) came to the conclusion that population increases geometrically while increase in food production follows a much slower arithmetic growth. This famous "law" did not prove to be correct, although it suggested a certain trend. It

could not claim the general validity of a mathematical method, since the data on which Malthus based his argument could not be represented by a mathematic relationship because of their complexity.

Growth and yield model is one of the oldest and broadest classes in forest modelling, dated back to the first yield tables in 18^{th} - 19^{th} centuries. They are intended to predict the expected yield of specific forest stands submitted to a given silvicultural regime. Most yield models share a common philosophy: the site-specific prediction of yield over time. Site index*[1] is the standard measure of productivity, a measure used only by foresters (Johnsen et al. 2001). Growth and yield models are developed from sample plot data.

A sample plot is usually a circular area about 1/50 to 1/10 ha (200 to 1000 m^2) in size, containing a small group of trees that are measured sometimes only once (Temporary Sample Plots - TSP) or sometimes at specific time intervals (Permanent Sample Plots - PSP), usually every 10 years in boreal and temperate forests. Information collected includes a biophysical site description, tree biometrical measurements (tree heights, diameters at breast height*), various form of stem defects and stand regenerations. The information is used to make forest management and harvest decisions. Permanent sample plots are used to monitor the rate of forest growth and hence establish the level of harvest that the forest can sustain.

[1]The terms with * are explained in Appendix A

1.2 Project: Investigating effects of climate on site index

Intergovernmental Panel on Climate Change (IPCC) was established by the World Meteorological Organization (WMO) and the United Nations Environment Programme (UNEP), regarding the problem of a potential global climate change. Based on its third assessment (Environmental Factsheet No. 11, December 2002)[2], the global average surface temperature has risen by 0.6 ($\pm$0.2)°C over the last hundred years. And it is very likely that 1990s was the warmest decade on record since 1861. As for future trend, average global temperatures are expected to rise by 1.4 to 5.8 °C from 1990 to 2100. Such rapid changes in a relatively short period of time should affect forests significantly, and especially the boreal forest (Stewart et. al. 1998), which accounts for a large portion of forest in Canada.

In Alberta, a climate change program was setup in 2003 in an attempt to develop a process-based model of lodgepole pine (Pinus contorta) growth that can be used to produce climate-sensitive growth and yield models[3].

Lodgepole pine is Alberta's provincial tree. It is the most common tree species in the Rocky Mountains and Foothills regions (Alberta Environmental Protection 1994). Although lodgepole pine comprises about 20% of the mature standing timber in Alberta, it accounts for approximately 40% of the annual harvest in the province (Huang et al., 2001).

[2] *http://www.acidrain.org*

[3] *http://www.fmf.ca/pa_cc.html*

In temperate and boreal forests, site index is usually defined as the stand top height* at 50 years of age (Huang et al. 2001). Logan T. and Price D. (2004)[4] studied the plots recently established and older plots growing on similar conditions in Alberta PSP, and found that the stand level site index has been increased. They also carried out stem analysis. The preliminary results suggested that, on average, trees from stands established more recently showed significantly faster mean height increment than older trees measured in the same region. Furthermore, this increase followed a gradual but consistent trend throughout the 20th Century.

Among a number of questions raised by these preliminary results, one possible factor could be that the observed regional climate warming in the past century has gradually improved growing conditions and/or lengthened growing seasons. Time-series of historical monthly climate data were interpolated (10 km resolution) to estimate changes in temperature at PSP locations during the period 1901 - 2000 (McKenney et al. 2000). Bernier P. (Canadian Forest Service, Laurentian Forestry Center) and Raulier F. (Laval university) focused on the application of "StandLEAP", a process-based model for forest productivity, to simulate the effects of climate on stand level. The purpose of this study is to detect if the effects of historical temperature and precipitation patterns can be observed on lodgepole pine growth in western Alberta.

[4] *http : //www.fmf.ca/CC/CC_Qn1.pdf* Climate Change Program Quicknote: Investigating Effects of Climate on Site index of Logdgepole Pine in Western Alberta.

1.3 Methods and tools

Project strategy

The project is part of the climate change program in Alberta, which aims at investigating the influence of climate change on the forest in Alberta. Forest management decisions are made according to the future forest yield. If forest yield is indeed sensitive to climate changes, climate changes should be an element to be considered before any forest decision can be made. Forest yield is sensitive to site fertility, an expression that summarizes the effects of temperature, rain, cloudiness, site slope and exposition, soil chemical fertility, texture and drainage. The introduction of a climate-sensitive site index thus represents an interesting contribution to measure the effects of climate changes on forest growth and yield and would help the forest industry to make appropriate decisions in order to maintain the sustainability of its forest resources. As top stand height at 50 years is used as a site index, it represents the cumulative effect of site productivity on stand growth and yield.

Climate changes affect forest growth at a time scale different than that at which forest management decisions are made. Forest management decisions are made over 100 to 200 years (once or twice the time a stand needs between the seed and harvest stages) in order to make sure that the wood cut by the industry will not deplete the natural resources available in a given territory (Davis et al. 2001). Climate changes mostly affect forest photosynthesis and respiration as they are sensitive to radiation, temperature, air dryness (expressed as water vapour pressure deficit) and available soil water. These effects occur at the scale of an hour to a week. Consequently, a

process-based growth model has to be used in order to summarize, on the long term, the effects of climate changes on stand growth and yield.

For this purpose, we used permanent sample plot data of Alberta to estimate a climate-sensitive site index model for pure lodgepole pine plots. To achieve this, we tried to reproduce the evolution of the mean annual increment (MAI) of forest standing aboveground biomass* with "StandLEAP" for 100 years with historical climate data constructed from weather station records. A climate-sensitive site index model was built based on the results. This model was then used to explain the average site index variability between 1901 and 2000. We hope to find if there exists an increasing trend of site index corresponding to what we have observed in the reality.

Statistical methods

To obtain a good model, we used a univariate analysis for preliminary analysis on main variables. We used Nonlinear Least Squares (NLS) regression method to estimate a climate-sensitive site index model and a univariate Box-Jenkins method to analyze average site index time series. The results are shown in Chapter 4.

Statistical tests (for example t-test, F-test, AIC, BIC etc.) and numerical methods (Newton, Marquart compromise, SUR etc.) were used for these estimation.

All the results were generated from the procedures of SAS©.

Chapter 2

Data and methodology

2.1 The data

The data which were used for this study include permanent sample plot data in Alberta and monthly climate data in those PSP locations.

2.1.1 Permanent sample plot data

The Permanent Sample Plot (PSP) data, collected by Alberta Land and Forest Division over the past four decades, were used for this study. The data were measured over 650 locations in western Alberta (Alberta land and forest division 2002).

Remeasurement of existing PSPs took place every 5 or 10 years depending on the stands and ages. There was only one plot in 339 locations and a cluster of three or four plots in each of the remaining locations. Altogether, there were 1751 plot measurements with some plots remeasured up to 6 times.

Aspect, elevation, slope and soil conditions were measured at plot level. DBH

and health status were recorded for each tree every time. As height was not as easy as DBH to measure, 221,241 over 1,079,877 trees were chosen for the height measurements (per tree measurement).

Re-measured PSP data are correlated temporally. These correlations violate the basic assumption of independent errors in least squares methods, which invalidates the corresponding t-test, F-test and confidence intervals (Kozak, 1997).

However, Gertner (1987) and Borders et al. (1987) found that the temporal correlation decreased as the measurement interval increased, and did not occur for non-overlapping intervals. Huang (1997) concluded that for prediction purposes, whether correlation was accounted for or not had little practical significance.

2.1.2 Climate data

Monthly climate data including maximum temperature, minimum temperature, and precipitation from 1901-2000 for Alberta PSP locations was based on the studies of McKenney et al. (2000). In their studies, they have re-mapped Canadian hardiness zone* through 1930 - 1990. The studies were based on the hardiness indices and zones using the data from roughly 1930-1960, for the period of 1961-1990, a thin plate spline interpolation method was used. A function with 3 variables (latitude, longitude, and elevation) was shown to give the best performance among several trials. Digital elevation models captured the spatial variation (such as elevation at any point, slope and aspect) in climate more accurately, and enabled the mapping for the hardiness formula at spatial resolutions of 1 km to 10 km. Standard errors of the temperature

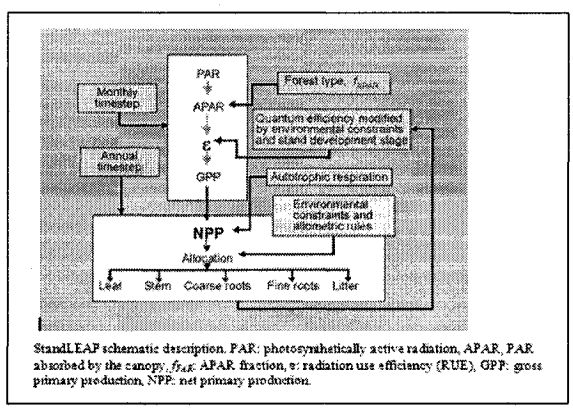
variables were about 0.5°C or less and 5 to 28% for rainfall.

2.2 Process-based growth model

Forest process models are mathematical representations of biological systems that incorporate our understanding of physiological and ecological mechanisms into predictive algorithms. Their use in research has developed rapidly in the past 20 years for two main reasons: Firstly, the steady gain in our understanding of forest biology and ecology has been coupled with great technological improvements in computers and softwares. Secondly, and more importantly, there is a great need to address questions posed at scales higher than those at which processes are being measured.

StandLEAP (Raulier et al. 2000) is a top-down radiation-use-efficiency (RUE) model that computes net primary productivity (NPP) of a forest stand from the fraction of absorbed photosynthetically active radiation (fPAR) (Figure 2.1). Derived from the 3-PG model of Landsberg and Gower (1997), it uses many of the same modifiers to constrain NPP as a function of specific limiting environmental conditions and stand properties including air temperature, soil water content and stand developmental stage. Transpiration is estimated via a water-use-efficiency (WUE) model (Dewar 1997) and is also constrained by limiting environmental conditions in a similar fashion as that for the RUE model. The time step is monthly and the results are summarized on a yearly basis. In order to validate results against permanent sample plots, StandLEAP also simulates stand dynamics through the computation of self-thinning and accrual of standing biomass.

StandLEAP is defined to operate at the canopy level and on a monthly time step. These spatial and temporal scales differ greatly from those at which direct and diffuse light is absorbed by the canopy elements (leaves, shoots, branches and stems) to drive photosynthesis. Hence, the process of scaling up from the leaf-level and hourly timesteps to the canopy and monthly intervals involves prior use of a more detailed process model (FineLEAP, Raulier et al. 2000, Bernier et al. 2001), parameterized from field measurements of growth processes. The choice of tree-level processes to be used in StandLEAP, the shapes of the functions representing them, and the values of the parameters used in these functions are all derived from simulations carried out with FineLEAP.



StandLEAP schematic description. PAR: photosynthetically active radiation, APAR, PAR absorbed by the canopy, f_{PAR}: APAR fraction, ε: radiation use efficiency (RUE), GPP: gross primary production, NPP: net primary production.

Figure 2.1: StandLEAP: Description

As any process-based model, StandLEAP integrated a large number of sub-models, and the correlations between different models can not always be identified. Even though each model were carefully calibrated and shown with a good result, we were not sure how the whole model might behave (as we discussed in Section 5.5, this is more interesting in reality). So validating the overall performance of the process

model was especially important.

We have done the validation of StandLEAP for lodgepole pine using Permanent Sampling Plots in Alberta, and the results was quite applausible as we can see from Figure 2.2:

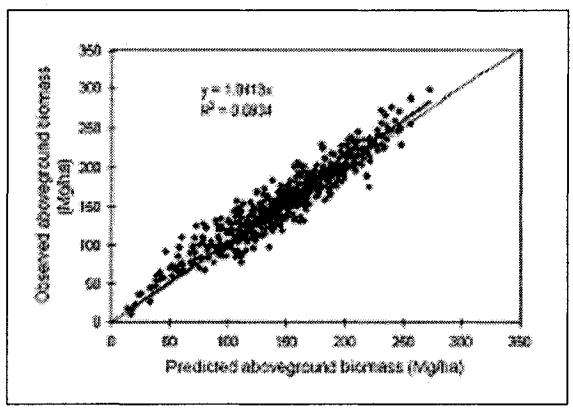


Figure 2.2: StandLEAP: Validation

2.3 Methodology

Data examining and cleaning is quite important to assure a successful analysis. One task is to recognize outliers, because we only need to work with data that contributes to the analysis. In this project, we investigated the growth and yield of merchantable lodgepole pines in pure stands. So all trees with DBH < 9.1 cm were removed. All plots in which basal area* of lodgepole pine took less than 75% of total stand basal area were removed also. Other criteria for cleaning data were to avoid plots where trees were damaged by insects or other natural disturbances and plots that had been commercially cut. By removing these anomalous plots, we can focus our study on

the natural growth of pure and undisturbed lodgepole pine stands. Of course, these factors are also important to the forest ecosystem. But we are not interested in them in this project. We will see the data cleaning in detail in Section 4.1. After we have the cleaned data, the analysis can be proceeded in two steps:

1. Calibrate a climate-sensitive site index model;

2. Analyze the average site index time-series.

2.3.1 Calibration of a climate-sensitive site index model

Since we focus on the results from StandLEAP, we need to prepare the inputs for StandLEAP. These include: density, biomass, soil, position and elevation, and climate. Except for biomass, the rest are directly measured or received from other organizations.

Biomass is a function of DBH and height as we can see from Lambert et al. (2005). All DBH were measured but heights were measured selectively. To keep the completeness of the data at plot level, we need the height-DBH model to proceed. The result of this model is shown in Section 4.2.

For the calibration of the climate-sensitive site index model, we run StandLEAP on climate normals (this can be thought of as an average monthly climate data) for 150 years. StandLEAP cannot estimate site index but is good at biomass production estimation. Since both site index and maximum annual increment in biomass are indices of site quality (Rondeux 1993), our initial idea is to develop a site index model with maximum annual increment. However, a large portion of plots in this study have

passed their ages at the maximum, we decide to model it in two steps: first, calibrate the evolution of Mean Annual Increment (MAI) in aboveground biomass predicted by StandLEAP as a function of mean stand age, and then calibrate a site index model based on the parameters of the last models. Site index values are calculated from GYPSY, a well-accepted model in Alberta (Huang S. 2001). The climate-sensitive site index model is shown in Section 4.3.

2.3.2 Analysis on average site index series

We run StandLEAP to estimate biomass for 100 years with the climate data from 1901-2000 interpolated by McKenney et al. (2000) (this time we used 100 years instead of 150 because the values after 100 year did not contribute much to the shape of the curve, Figure 4.11.) Then we calibrate MAI-age model by plot, which means, for each plot, we get a series of parameters under a particular climate year between 1901 and 2000. Using the parameters of climate-sensitive site index model from Section 4.3, we have the climate-sensitive site index values for each plot and each year. We then average the annual values of site index over all the plots to obtain an average site index series over the years 1901-2000. We use the univariate Box-Jenkins method to investigate the trend and other properties of the series.

Chapter 3

Statistical consideration

Data analysis procedures involve many practical considerations of statistical methods. In this chapter, we will discuss some techniques from the view of model fitting.

3.1 Nonlinear Least Squares model

Linear regression is a powerful method for analyzing data described by models which are linear in the variables. However, to obtain a more accurate relation between the response and the regressor, a nonlinear model is often required. In such cases, linear regression techniques must be extended, which introduces considerable complexity. A nonlinear model can be written as:

$$Y = f(X_1, X_2, ..., X_k, \beta_1, \beta_2, ..., \beta_p) + \varepsilon \tag{3.1}$$

where f is a nonlinear function of the k independent variables $X_1, X_2, ...X_k$ and the p coefficients $\beta_1, ..., \beta_p$, and assume that $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$, observations are independent of each other as linear regression. The criterion used for determining the

estimated values for the coefficients is the same as that used in the linear regression, i.e., minimizing the sum of squared errors (SSE). For N observations, SSE can be written as:

$$SSE = \sum_{i=1}^{N} [Y_i - f(X_{1i}, ..., X_{ki}, \beta_1, ..., \beta_p)]^2 = \sum_{i=1}^{N} \varepsilon_i^2.$$

If ε follows $N(0, \sigma^2)$, the least squares estimate β is also the maximum likelihood estimate of β. This is because the likelihood function for this problem can be written as

$$l(\beta, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{\varepsilon_i^2/2\sigma^2} = (2\pi\sigma^2)^{-n/2} e^{-SSE(\beta)/2\sigma^2}$$

so that if σ^2 is known, maximizing $l(\beta, \sigma^2)$ with respect to β is equivalent to minimizing $SSE(\beta)$ with respect to β.

3.1.1 Model specification

An important step in any nonlinear analysis is the specification of the model, which includes specifying both the model function and the characteristics of the disturbances.

The model function

Theoretically, any ecological considerations will lead to a model function. The analyst's job is then to find the simplest form of the model and the parameter estimates which provide an adequate fit of the model to the data, subject to the assumptions about the disturbance.

The disturbance term

All nonlinear estimation programs are based on specific assumptions about the disturbance term, usually that the disturbance is additive and normally distributed with zero mean, constant variance, and independence between different observations.

3.1.2 Starting values

One of the best things one can do to ensure a successful nonlinear analysis is to obtain good starting values for the parameters, values from which the convergence is quickly obtained. Several simple but useful principles for determining starting values can be used:

(1) Interpreting the model behavior :

- One of the advantages of nonlinear regression is that the parameters in the model function are usually meaningful. This meaning can be very helpful in determining starting values.
- Plotting a nonlinear model function using various values for the parameters is an beneficial exercise. In this way one becomes familiar with the function and how the parameters affect its behavior.
- Sometimes starting values can be obtained by considering the behavior near the origin or at other special values (such as limits).

(2) Interpreting derivatives of the model function: Sometimes rates of change

of the function at specified design values can be used to obtain parameter starting estimates.

(3) Transforming the model function: Transformations of the model function can often be used to obtain starting values. Log and reciprocal transform are two commonly used ways.

(4) Reducing Dimensions: Peeling is an example of the general technique of reducing dimensions. In this technique one estimates parameters successively, each estimated parameter making it easier to estimate the remaining ones. Consider the model

$$f(x) = \beta_1 + \beta_2 exp(-\beta_3 x),$$

where β_3 is positive. Then the limiting value of the response when $x \rightarrow \infty$ is β_1 and the value at $x = 0$ is $\beta_1 + \beta_2$. Depending on whether the data is increasing or decreasing, we can use y_{max} or y_{min} to get the starting value of β_1^0, and then use the difference $f(0) - \beta_1^0$ to get β_2^0. Once β_1^0 and β_2^0 are determined, we could substitute these values into the function and evaluate $(1/x)ln[(y - \beta_1^0)/\beta_2^0]$ at selected values of x to obtain β_3^0.

3.1.3 Computing method

To find the least squares estimate of $\hat{\beta}$ we need to differentiate Equation (3.1) with respect to β. This provides the p normal equations, which must be solved for $\hat{\beta}$.

$$\sum_{t=1}^{N} [Y_t - f(X_t, \beta)] \left[\frac{\partial f(X_t, \beta)}{\partial \beta_i} \right] = 0, \qquad i = 1, 2, ..., p.$$

A closed form solution generally does not exist. Thus, PROC NLIN, a procedure for fitting nonlinear model (SAS/STATS User's Manual, 1999) uses an iterative procedure: a starting value for β is chosen and continually improved until the SSE is minimized.

The iterative process begins at some point β_0. Then X and Y are used to compute a change in parameters $\triangle$ such that

$$SSE(\beta_0 + k\triangle) < SSE(\beta_0)$$

where k is for controlling the convergence precision. Four methods are implemented in PROC NLIN and they differ in how $\triangle$ is computed to change the vector of parameters.

$$\begin{aligned}
\text{Steepest descent} : \triangle &= X'e \\
\text{Gauss-Newton} : \triangle &= (X'X)^{-1}X'e \\
\text{Newton} : \triangle &= (G)^{-1}X'e \\
\text{Marquardt} : \triangle &= (X'X + \lambda diag(X'X)))^{-1}X'e
\end{aligned}$$

where

$$G = (X'X) + \sum_{i=1}^{n} H_i(\beta)e_i,$$

and $H_i(\beta)$ is the Hessian matrix of e:

$$[H_i]_{jk} = \left[\frac{\partial^2 e_i}{\partial\beta_j\partial\beta_k}\right]_{jk}$$

The details can be found in SAS/STATS User's Manual (1999).

3.1.4 Obtaining convergence

Obtaining convergence is sometimes difficult. When we have troubles, we should check the model function first with care. Then try different starting values to see if it helps. Changing the calculation method sometimes can give a good result. Sometimes convergence is not obtained because the model has too many parameters, PROC NLIN can give out warnings when there is collinearity detected. In such case, we should consider a simpler model instead.

3.1.5 Validating model

Once a model is fitted, an assessment of its validity using an independent data set is needed to see if the quality of the fit reflects the quality of prediction.

Graphical validity

Graph is one of the most important parts in model validation. In practice, two curves are most widely used:

1. Plots showing the observed vs. the fitted values;

2. Plots showing the predicted errors vs. predicted values.

When the model involves time factor, the trajectories of observed vs. predicted over time and errors over time should be plotted. Such a plot can show whether the prediction variance is changing across all the time range.

Statistical tests

Several of the most frequently used prediction statistics are: mean prediction error ($\bar{e}$) and percentage error ($\bar{e}\%$), mean absolute difference (MAD), mean square error of prediction (MSEP), relative error in prediction (RE%) and prediction coefficient of determination (R_p^2):

$$\begin{aligned}
\bar{e} &= \sum_{i=1}^{k} \frac{(y_i - \hat{y}_i)}{k} = \sum \frac{e_i}{k}, \qquad \bar{e}\% = 100 \times \frac{e_i}{\bar{y}}, \\
MAD &= \sum_{i=1}^{k} \frac{|y_i - \hat{y}_i|}{k}, \\
MSEP &= \sum_{i=1}^{k} \frac{(y_i - \hat{y}_i)^2}{k} = \sum_{i=1}^{k} \frac{e_i^2}{k}, \qquad RE\% = 100 \frac{\sqrt{MSEP}}{\bar{y}}, \\
R_p^2 &= 1 - \frac{\sum_{i=1}^{k} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{k} (y_i - \bar{y})^2}.
\end{aligned}$$

where k is the number of samples, and $\bar{y}$ is the average of the observed y values.

The validation statistics shown above provide "averaged" measures with respect to the overall model performance. Sometimes, the prediction performance may be needed at some critical ranges.

In fact, validation is a big topic in forest modelling, more options were discussed in Section 5.5.

3.1.6 Comparing models

In some situations there may be more than one function which could be used as a model, one model may give a superior fit to the data.

Nested models

To decide which is the simplest nested model to fit a data set adequately, Draper and Smith (1998) suggest an assessment of the extra sum of squares due to the extra parameters involved in going from the partial to the full model.

Letting S denote the sum of squares, ν denote the degree of freedom, and P denote the number of parameters, with subscripts f and p for the full and partial models and a subscript e for extra, the calculations can be summarized as in Table 3.1. To complete the analysis, we compare the ratio S_e^2/S_f^2 to $F(\nu_e, \nu_f; \alpha)$ and accept the partial model if the calculated mean square ratio is lower than the corresponding critical value. Otherwise, we retain the extra terms and use the full model.

Source	Sum of Squares	Degree of Freedom	Mean Squares	F ratio
Extra parameters	$S_e = S_p - S_f$	$\nu_e = P_f - P_p$	$s_e^2 = S_e/\nu_e$	s_e^2/s_f^2
Full model	S_f	$\nu_f = N - P_f$	$s_f^2 = S_f/\nu_f$	
Partial model	S_p	$N - P_p$		

Table 3.1: Data processing: Extra of squares analysis

Non-nested models

When trying to decide which of the several non-nested models is best, the first approach should be referred to scientific reasons, because the primary aim of data analysis is to explain or account for the behavior of the data, not simply to get the best fit.

The statistical analysis follows if the first approach does not work. The most important statistical analysis is the analysis of the residuals. Generally the model with the smallest mean square and the most random-looking residuals should be chosen.

3.2 Univariate time series analysis

In traditional regression analysis, it is assumed that various observations within a single data series are statistically independent, and this is the standard assumption about the error terms. But with a time-series data, like the tree height, our forecasts of tree height for next year are based on the height of the current year. We will start with the idea that the observations in a time series may be statistically related to other observations in the same series.

3.2.1 Univariate Box-Jenkins model

The time series modelling and forecasting was first developed in the late 60's, using lags and shifts in the historical data to uncover patterns and predict the future. It became quite popular following the publication of the book "Time Series Analysis: Forecasting and Control" by George Box and Gwilym Jenkins in 1976. In this book, they used the symbol ARIMA(p, d, q) to represent a large class of models which could describe the behavior of many observed time series. The acronym ARIMA stands for "Auto-Regressive Integrated Moving Average." "p" is the number of autoregressive terms, if lags of the variable appearing in the equation; and "q" is the number of

lagged forecast errors in the prediction equation, which is called "moving average" terms. "d" is the number of times the series to be differenced before arriving at a stationary series.

The general autoregressive-moving average process of order p and q is denoted by ARMA(p, q), and defined by

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q}.$$

It can be re-written with a back-shift operator B, such that $B^j y_t = y_{t-j}$,

$$(1 - \phi_1 B - ... - \phi_p B^p) y_t = (1 - \theta_1 B - ... - \theta_q B^q) \varepsilon_t.$$

where ε_t is zero-mean white noise.

If all $\theta_i = 0$, it is an auto-regressive model AR(p). The characteristic equation for an AR(p) process is:

$$1 - \phi_1 B - ... - \phi_p B^p = 0. \tag{3.2}$$

The stationarity of the process requires that all roots of the characteristic equation (3.2) must have absolutely value strictly greater than 1.

If all $\phi_i = 0$, it is a moving average model MA(q). The characteristic equation for an MA(q) process is:

$$1 - \theta_1 B - ... - \theta_q B^q = 0. \tag{3.3}$$

The MA(q) model is called invertible if the absolute values of the roots of the characteristic equation (3.3) are all strictly greater than 1. For an ARMA(p, q) model, the requirement for stationarity is the same as that for the AR(p) and the requirement for invertibility is the same as that for the MA(q).

The Procedure of Unvariate Box-Jenkins method

Box and Jenkins (1976) proposed a practical three-stage procedure for finding a good model.

Stage 1: Identification. At the identification stage, we tentatively select one or more ARIMA models by looking at two graphs derived from the available data. These graphs are called the estimated autocorrelation function (ACF) and the estimated partial autocorrelation function (PACF). We choose the models whose associated theoretical ACF and PACF look like the estimated ACF and PACF calculated from the data.

Stage 2: Estimation At this stage we tentatively select one or more models that seem likely to provide parsimonious and statistically adequate representations of the available data.

Stage 3: Diagnostic We perform tests to see if the estimated model is statistically adequate. If it is not satisfactory we return to the identification stage to tentatively select another model.

End of Box-Jenkins method — Random walk

The random walk is regarded as the simplest random process. y_t is determined by

$$y_t = y_{t-1} + \varepsilon_t,$$

with $E(\varepsilon_t) = 0$, $Var(\varepsilon_t) = \sigma^2$ and $E(\varepsilon_t \varepsilon_s) = 0$ for $t \neq s$. The forecast of one period ahead is given by

$$\hat{y}_{T+1} = E(y_{T+1}|y_T, ..., y_1).$$

But ε_{T+1} is independent of $y_T, ..., y_1$. Thus, the forecast of one period ahead is simply

$$\begin{aligned} \hat{y}_{T+1} &= E(y_T + \varepsilon_{T+1}|y_T, ..., y_1) \\ &= y_T + E(\varepsilon_{T+1}|y_T, ..., y_1) \\ &= y_T. \end{aligned}$$

The forecast of two periods ahead is

$$\begin{aligned} \hat{y}_{T+2} &= E(y_{T+2}|y_T, ..., y_1) \\ &= E(y_T + \varepsilon_{T+1} + \varepsilon_{T+2}|y_T, ..., y_1) \\ &= y_T + E(\varepsilon_{T+1} + \varepsilon_{T+2}|y_T, ..., y_1) \\ &= y_T. \end{aligned}$$

Similarly, the forecast k period ahead is also y_T. But the variance of k period ahead is

$$\begin{aligned} Var(y_{T+k}) &= Var(y_T + \varepsilon_{T+k} + \varepsilon_{T+k-1} + ... + \varepsilon_{T+1}|y_T, ..., y_1) \\ &= Var(y_T) + \sum_{i=1}^{k} Var(\varepsilon_{T+i}) \\ &= Var(y_T) + k\sigma^2. \end{aligned}$$

If a time series can be recognized as a random walk, there is nothing to do with the forecasting, then our analysis can be stopped.

3.2.2 Stationarity, transformation and unit root test

Univariate Box-Jenkins method can only deal with stationary time series. We define stationarity by classical probability theory.

Stationarity

A stochastic process is said to be strictly stationary if its properties are unaffected by a change of time origin. That is, if the joint probability distribution associated with n observations $y_1 ..., y_n$ made at any time is the same as that associated with n observations $y_{1+k}, ...y_{n+k}$. A weak stationarity of order f, is that the moments up to some order f depend only on time difference. For example, mean stationarity means that the expected value of the process is constant over time:

$$E(y_t) \equiv \mu, \qquad \forall t.$$

Similarly, variance stationarity means that the variance is temporally stable:

$$Var(y_t) = E[(Y_t - \mu)^2] \equiv \sigma_Y^2, \qquad \forall t,$$

and covariance stationarity is:

$$Cov(y_t, y_{t-s}) = E[(y_t - \mu)(y_{t-s} - \mu)] \equiv \gamma_s, \qquad \forall s,$$

which means that the autocovariance of two observations y_t and y_{t-s} depends only on the lag s, not on "where" they fall in the series.

In general, people are only interested in weak stationarity, and that is also what we will consider in here.

Homogeneous non-stationarity and differencing

Probably very few time series one encouters in practice are stationary, even in the weak sense. So their characteristics of the underlying stochastic process change over time. Fortunately, there is a big family of the non-stationary process which can be changed into stationary process by transformation. One commonly used transformation is differencing. We say that y_t is homogeneous nonstationary of order d if $\omega_t = \Delta^d y_t$ is a stationary series. Here Δ denotes differencing, i.e.,

$$\begin{aligned} \Delta y_t &= y_t - y_{t-1}, \\ \Delta^2 y_t &= \Delta y_t - \Delta y_{t-1}. \end{aligned}$$

Unit root tests for stationary

Suppose we have a variable Y_t, which has been observed growing over time. Then it can be described by the following equation:

$$Y_t = \alpha + \tau t + \rho Y_{t-1} + \varepsilon_t.$$

One possibility is that Y_t has been growing because it has a positive trend ($\tau > 0$ and $\rho = 0$), ie.

$$Y_t = \alpha + \tau t + \varepsilon_t, \tag{3.4}$$

or it follows a random walk with positive drift ($\alpha > 0$, $\tau = 0$ and $\rho = 1$). i.e.,

$$Y_t = \alpha + Y_{t-1} + \varepsilon_t. \tag{3.5}$$

In both cases, the series are non-stationary, and they can be made stationary by differencing. For Equation (3.4),

$$\begin{aligned} \Delta Y_t \equiv Y_t - Y_{t-1} &= \alpha + \tau t + \varepsilon_t - [\alpha + \tau(t-1) + \varepsilon_{t-1}] \\ &= \tau t + \varepsilon_t - \tau(t-1) - \varepsilon_{t-1} \\ &= \varepsilon_t - \varepsilon_{t-1} + \tau, \end{aligned}$$

which is stationary. For Equation (3.5), differencing yields:

$$\begin{aligned} \Delta Y_t \equiv Y_t - Y_{t-1} &= \alpha + Y_{t-1} + \varepsilon_t - Y_{t-1} \\ &= \alpha + \varepsilon_t, \end{aligned}$$

which is also stationary.

Dickey Fuller unit root tests

The problem with these two models is that differencing by itself cannot distinguish between (3.4) and (3.5). Obviously we can estimate Y_t of (3.4) by ordinary least square regression. But for (3.5), if ρ is indeed 1, the use of OLS in this manner can lead to incorrectly reject the random walk hypothesis, because the distribution of $\hat{\rho}$ is nonstandard under the null hypothesis of $\hat{\rho} = 1$. The distribution it follows is known as the "Dickey Fuller" (D-F) distribution, which was first derived in 1979. The D-F distribution is right-skewed (so the t-statistics will tend to be larger and negative), which means that we will tend to over-reject the null hypothesis if we use the standard t-distribution. Thus the "Dickey-Fuller" test for a unit root amounts

to estimating $\hat{\rho}$ and doing a standard-looking t-test for $H_0 : \hat{\rho} = 1$, but using a non-standard set of critical values. If "t" is less than the critical values, this is the evidence of nonstationary. Otherwise, it indicates a stationary series. Note that the D-F test requires that the ε's are white noise.

Augmented Dickey Fuller test

Suppose we have a series which is a stationary AR(p) process after the first difference.

$$\Delta Y_t = \sum_{j=1}^{p} \psi_j \Delta Y_{t-j} + \varepsilon_t,$$

if we estimate a standard D.F. test:

$$Y_t = \hat{\rho} Y_{t-1} + \varepsilon_t,$$

the term $\sum_{j=1}^{p} \psi_j \Delta Y_{t-j}$ gets lumped into the error ε_t. This induces an AR(p) structure in the ε's, and the standard D.F. test statistics will be wrong, We need to change the model a little bit to deal with the case, which is called augmented Dickey Fuller Test (ADF). It is based on estimating the test:

$$Y_t = \hat{\rho} Y_{t-1} + \sum_{j=1}^{p} \psi_j \Delta Y_{t-j} + \varepsilon_t, \tag{3.6}$$

The null hypothesis for this test is $\hat{\rho} = 1$, which means, a unit root exists, hence nonstationary. We usually use I(1) to represent this hypothesis.

Let $\hat{\tau} = \hat{\rho} - 1$, Equation (3.6) can be re-written as:

$$\Delta Y_t = \hat{\tau} Y_{t-1} + \sum_{j=1}^{p} \psi_j \Delta Y_{t-j} + \varepsilon_t, \tag{3.7}$$

The null hypothesis for Equation (3.7) is still I(1), which implies $\hat{\tau} = 0$. This testing equation is more practical because it is the usual t-statistic reported for testing

the significance of the coefficient for Y_{t-1}. The ADF t-statistic (column τ as appeared in the output of PROC ARIMA, a procedure for fitting ARIMA models, see SAS/ETS User's Manual 1999, Dickey 2005) and normalized bias statistic (column ρ as appeared in the output of PROC ARIMA) are based on least squares estimates of Equation (3.7) and are defined by

$$ADF_t = \frac{\hat{\tau}}{SE(\tau)},$$

and

$$ADF_n = \frac{N\hat{\tau}}{1 - \hat{\psi}_1 - ... - \hat{\psi}_p}.$$

An important practical issue for the ADF test is the specification of the lag length p. If p is too small, the remaining serial correlation in the errors will make the test bias; while large p value will reduce the power of the test statistic.

Ng and Perron (1995) suggest a data dependent lag length selection procedure: Start with an upper bound p_{max}, and test whether the coefficients of the last lags are significant. If they are, set $\hat{p} = p_{max}$ and perform the unit root test. Otherwise, reduce the lag length by one and repeat the procedure.

Schwert (1989) suggests a useful rule for determining p_{max}:

$$p_{max} = \left[12\left(\frac{N}{100}\right)^{1/4}\right],$$

where N is the number of data in the series.

Phillips-Perron test

The Phillips-Perron Test (1988) for a unit root adopts a strategy which is a little different. Rather than changing the model estimated, the PP test sticks with the model:

$$\Delta Y_t = \hat{\pi} Y_{t-1} + \varepsilon_t,$$

where ε_t may be heteroscedastic. By directly modifying the test statistics $t_{\hat{\pi}=0}$ and $T\hat{\pi}$, the PP test corrects any serial correlation and heteroscedasticity in the error of the test regression, where T is the number of data in the series. These modification statistics, denoted by Z_t and Z_π, are given by

$$Z_t = \sqrt{\frac{\hat{\sigma}^2}{\hat{\lambda}^2}} \cdot t_{\hat{\pi}=0} - \frac{1}{2}\left(\frac{\hat{\lambda}^2 - \hat{\sigma}^2}{\hat{\lambda}^2}\right)\left(\frac{T \cdot SE_{\hat{\pi}}}{\hat{\sigma}^2}\right),$$

and

$$Z_\pi = T \cdot \hat{\pi} - \frac{1}{2}\frac{T^2 \cdot SE_{\hat{\pi}}}{\hat{\sigma}^2}(\hat{\lambda}^2 - \hat{\sigma}^2),$$

The terms $\hat{\lambda}^2$ and $\hat{\sigma}^2$ are consistent estimates of the variance parameters:

$$\sigma^2 = lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} E[\varepsilon_t^2],$$

and

$$\lambda^2 = lim_{T\to\infty} \sum_{t=1}^{T} E[T^{-1} S_T^2],$$

where $S_T = \sum_{i=1}^{T} \varepsilon_t^2$.

Under the null hypothesis that $\hat{\pi} = 0$, the PP Z_t and Z_π statistics have the same asymptotic distributions as the ADF t-test statistics and normalized bias statistics.

3.2.3 Identification

Identification is clearly a critical stage in Box-Jenkins methods. The specific aim here is to obtain some idea of the values of p, d and q needed in the general ARIMA model and to obtain initial estimates for the parameters. The tentative model so obtained provides a starting point for the application of the more formal and efficient estimation methods. The basic tools are the sample autocorrelation function and the partial autocorrelation function.

ACF

Since it is usually impossible to obtain a complete description of a stochastic process, the autocorrelation function is extremely useful because it provides a partial description of the process for modeling purposes. We define the autocorrelation with lag k as:

$$\rho_k = \frac{E[(y_t - \mu_y)(y_{t+k} - \mu_y)]}{\sqrt{E[(y_t - \mu_y)^2]E[(y_{t+k} - \mu_y)^2]}} = \frac{Cov(y_t, y_{t+k})}{\sigma_{y_t}\sigma_{y_{t+k}}}.$$

For a stationary process the variance is a constant, say γ_0, then

$$\rho_k = \frac{\gamma_k}{\gamma_0}.$$

The ACF above is theoretical, in practice, we need to calculate the estimated sample autocorrelation function:

$$\hat{\rho}_k = \frac{\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}.$$

PACF

One problem in constructing autoregressive models is identifying the order of the underlying process. For moving average models this is less of a problem, since if the process is of order q the sample autocorrelations should all be close to zero for lags greater than q. (Bartlett's formula provides approximate standard errors for the autocorrelations, so that the order of a moving average process can be determined from significance tests on the sample autocorrelations.) Although some information about the order of an autoregressive process can be obtained from the oscillatory behavior of the sample autocorrelation function, much more information can be obtained from the partial autocorrelation function.

For an AR(p) process, the covariance with lag k is determined from

$$\gamma_k = E[y_{t-k}(\phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t)]. \tag{3.8}$$

Dividing both sides of (3.8) by γ_0, we obtain the following Yule-Walker equations:

$$\begin{aligned}
\rho_1 &= \phi_1 + \phi_2\rho_1 + ... + \phi_p\rho_{p-1}, \\
\rho_2 &= \phi_1\rho_1 + \phi_2 + ... + \phi_p\rho_{p-2}, \\
&\text{.....}, \\
\rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + ... + \phi_p.
\end{aligned}$$

If $\rho_1, \rho_2, ...\rho_p$ are known, the equations can be solved for $\phi_1, \phi_2, ...\phi_p$. Unfortunately, solution of the Yule-Walker equations requires knowledge of p, the order of autoregressive process, which we are looking for. Therefore, we solve the equations for successive values of p.

We start from $p = 1$, which leads to $\hat{\rho}_1 = \hat{\phi}_1$. Thus, if the calculated value $\hat{\phi}_1$ is significantly different from zero or not, we know that the autoregressive process is at least order 1. we denote this $\hat{\phi}_1$ as a_1.

When $p = 2$, solve the Yule-Walker equations, we get a new set of $\hat{\phi}_1$ and $\hat{\phi}_2$. If $\hat{\phi}_2$ is approximately zero, we can conclude that $p = 1$. While if $\hat{\phi}_2$ is significantly different from zero we can conclude that the process is at least order 2. we denote the value of $\hat{\phi}_2$ as a_2.

...

Repeat this process for successive values of p. Thus, we obtain a series of $a_1, a_2, ...$, we call this series the partial autocorrelation function and note that if the order of the autoregressive process is p, we should observe that $a_j \approx 0$ for $j > p$.

To test whether a particular a_j is zero, we can use the fact that the PACF series is approximately normally distributed, with mean zero and variance 1/T. Hence we can check whether it is statistically significant at, say, the 5 percent level by determining whether it exceeds $2/\sqrt{T}$ in magnitude.

Table 3.2 summarizes some of the behaviors regarding autocorrelation and partial autocorrelation, stationarity and invertibility for ARMA model:

3.2.4 Estimation and diagnostic

Goodness-of-fit statistics

The Akaike Information Criterion (AIC) (Akaike 1974) and Schwarz's Bayesian Criterion (SBC) (Schwarz 1978) are general tests for model specification. They can be

ARIMA	AR(p)	MA(q)	ARMA(p, q)
ACF	$\neq 0$ in general at lag k	0 for $k > q$	$\neq 0$ in general for all k
PACF	0 for $k > p$	$\neq 0$ in general for all k	$\neq 0$ in general for all k
Stationarity	restrictions for ϕ's	Stationary if $q < \infty$	restrictions for ϕ's
Invertibility	invertible if $p < \infty$	restrictions for θ's	restrictions for θ's

Table 3.2: ARMA models comparation

applied across a range of different areas, and are like F-tests in that they allow for the testing of the relative power of nested models. Each, however, does so by penalizing models which are over-parameterized. The AIC statistic is:

$$AIC(p) = -2ln(L) + 2k,$$

where L is the likelihood function and k is the number of free parameters in the model; similarly, the SBC statistic is calculated as:

$$SBC(p) = -2ln(L) + 2\ln(n)k,$$

where n is the number of residuals that can be computed from the time series.

The idea is to calculate these statistics for a range of different values of p, and then choose the model in which the statistic is the lowest. Note that the SBC statistic imposes a greater 'penalty' for larger numbers of parameters, this means that the model selected using the SBC statistic will always be at least as parsimonious as that chosen using AIC.

Check for White Noise residuals

Suppose that a time series Y_t is generated by the stationary ARMA(p, q) process

$$(1 - \phi_1 B - ... - \phi_p B^p) Y_t = (1 - \theta_1 B - ... - \theta_q B^q) \varepsilon_t \tag{3.9}$$

where $B^j Y_t = Y_{t-j}$ and ε_t is zero-mean white noise. An integral part of the methodology of Box and Jenkins (1976) for fitting models (3.9) involves "diagnostic checks" on the adequacy of representation of an initially identified model to a series of n observations. One such check, developed by Box and Pierce (1970), contemplates general alternatives within the autoregressive-moving average class of models. Denote the residuals from the fitted model by $\hat{\varepsilon}_t$, with autocorrelations

$$\hat{r}_k = \frac{\sum_{t=k+1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}, \qquad k = 1, 2, ...$$

Box and Pierce showed that, under the hypothesis of correct model specification, provided that m is moderately large, the statistic

$$Q = n \sum_{k=1}^{m} \hat{r}_k^2,$$

is asymptotically distributed as χ^2 with $(m - p - q)$ degrees of freedom. Tests of model adequacy based on this statistic are generally called portmanteau tests.

It has been shown by Davies, Triggs and Newbold (1977) that, for sample sizes commonly found in practice, the actual significance levels of Q can be considerably lower than those predicted by asymptotic theory. However, a simple modification, studied in detail by Ljung and Box (1978),

$$Q' = n(n+2) \sum_{k=1}^{m} \frac{\hat{r}_k^2}{n-k},$$

appears to have a distribution very much closer to the asymptotic χ^2. It would seem preferable, then, to base tests of model adequacy on Ljung-Box statistics.

The final step of Box-Jenkins method is forecasting based on the valid estimation equations. The forecasts are generated using forecasting equations consistent with the method used to estimate the model parameters. Usually, the confidence interval will be given together with the forecasting.

Chapter 4

Results

Data should be cleaned before any analysis based on the purpose of the project.

Biomass is required by StandLEAP, which is the function of height and DBH, so a height-DBH model should be calibrated first.

Climate-sensitive site index model is the key model in this study developed from the results of StandLEAP.

Average site index series is analyzed using univariate Box-Jenkins method to detect any changes of the site index over the last century.

4.1 Data cleaning

Diameter at Breast height (DBH) is the most frequent measurement made by a forester. This has traditionally been the "sweet spot" on a tree where measurements are taken. Many calculations are made to determine growth, volume, yields etc. based on DBH.

We first take a look at the basic statistics of DBH (unit: cm), Figure 4.1.

Analysis Variable : DBH

N	Mean	Std Dev	Minimum	Maximum
717025	19.7911938	8.7582353	9.1000000	662.7000000

Figure 4.1: Basic statistics of DBH

We find extreme values - for instance, trees with DBH of 6m is out of imagination. Since we are working with permanent sampling plots, we can find the DBH of the same tree at another measurement, so we are quite sure that it is a transcription mistake.

Figure 4.2 shows the data set after cleaning: Figure 4.1 and Figure 4.2 are gener-

Analysis Variable : DBH

N	Mean	Std Dev	Minimum	Maximum
274181	16.9505108	6.0523741	9.1000000	67.7000000

Figure 4.2: Basic statistics of DBH after cleaning

ated from PROC MEANS (SAS/STATS User's Manual 1999). PROC UNIVARIATE is another univariate procedure which gives more details. Parts of its output are listed in Figure 4.3 and Figure 4.4.

From Quantiles (Figure 4.3) and histogram (Figure 4.4), we can see that, the distribution of DBH is not normal. There are some big trees forming the tail starting around 37.5cm. It is natural for a forest. But since the 99 quantile is 34.6cm from Figure 4.3, we can be assured that the bigger trees would not have much influence in our analysis. Another point we should notice is that the data is left censored at

Quantile	Estimate
100% Max	67.7
99%	34.6
95%	28.7
90%	25.5
75% Q3	20.5
50% Median	15.6
25% Q1	12.2
10%	10.3
5%	9.7
1%	9.1
0% Min	9.1

Figure 4.3: Graphs of DBH: Quantiles

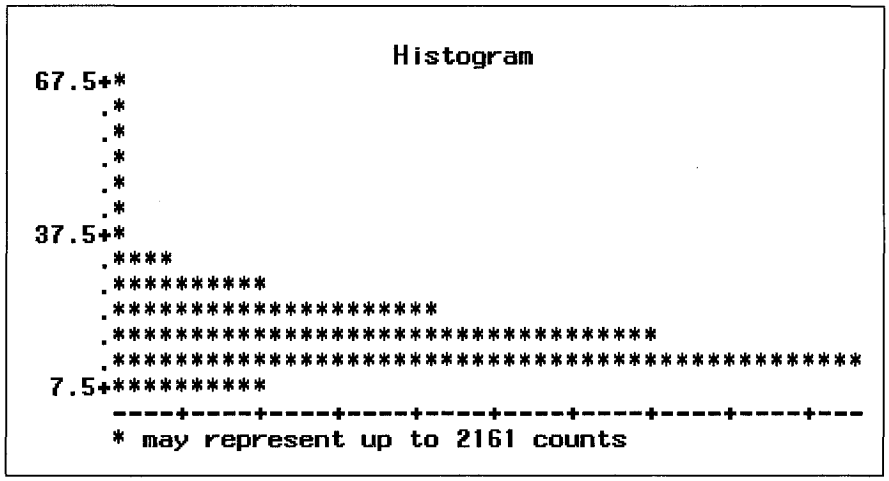


Figure 4.4: Graphs of DBH: Histogram

9.1cm, the minimum DBH for a merchantable tree.

4.2 Height-DBH model

The relation between height and diameter at breast height(DBH) is important and is quite often modelled, since height is costly to measure in the field. In practice, DBH is measured for all the trees in a plot and people only carefully subsample a few trees for the height measurement which are supposed to represent the whole plot. Consequently, we need to estimate a height-DBH relation for individual trees. Here we introduce two models which have been widely used in the literature and compare the behaviors of the models.

The data for height-DBH curve are shown in Figure 4.5.

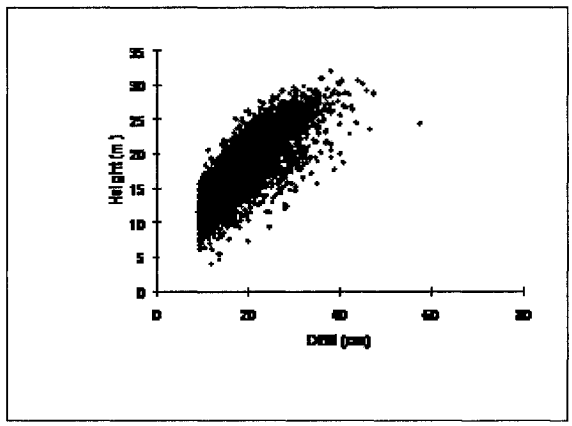


Figure 4.5: H-D: Height-DBH curve

Model (a) (equation (4.1)) was studied by Bégin and Raulier (1995) in Quebec, a

nonlinear equation considering average height (H) and DBH (D):

$$H = 1.3 + \frac{D}{\frac{\bar{D}}{\bar{H}-1.3} + \beta_2 * (D - \bar{D})}, \tag{4.1}$$

where $\bar{D}$ and $\bar{H}$ represent the average DBH and the average height at plot level. Equation (4.1) can be rewritten as:

$$\frac{D}{H - 1.3} - \frac{\bar{D}}{\bar{H} - 1.3} = \beta_2 * (D - \bar{D}), \tag{4.2}$$

Since DBH is the diameter at 1.3-meter height of a tree, the inverse of the ratio $\frac{D}{H-1.3}$ is called "form factor" in forest management. Equation (4.2) shows that we can use the difference from the mean in DBH on the RHS to explain the ratio difference from a "mean tree" appearing on LHS. We use the nonlinear fitting procedure PROC NLIN (SAS/STATS User's Manual 1999) to fit the model and the results are shown in Figure 4.6 and Figure 4.7:

NOTE: An intercept was not specified for this model.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	1	17597210	17597210	8500236	<.0001
Error	52686	109071	2.0702		
Uncorrected Total	52687	17706280			

Figure 4.6: H-D(a): ANOVA

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
beta2	0.0293	0.000097	0.0291	0.0295

Figure 4.7: H-D(a): Parameters Estimates

Model (b) was studied by Huang S. (1994) in Alberta, an exponential equation with three parameters (a, b, c):

$$H = 1.3 + a * (1 - e^{-b*D})^c.$$

We use the same procedure to fit the model and the results are shown in Figure 4.8 and Figure 4.9.

NOTE: An intercept was not specified for this model.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	3	17329556	5776519	807833	<.0001
Error	52684	376724	7.1506		
Uncorrected Total	52687	17706280			

Figure 4.8: H-D(b): ANOVA

The NLIN Procedure

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
a	29.7605	0.2854	29.2012	30.3199
b	0.0503	0.00137	0.0476	0.0530
c	1.1578	0.0184	1.1216	1.1939

Approximate Correlation Matrix

	a	b	c
a	1.0000000	-0.9813477	-0.9233978
b	-0.9813477	1.0000000	0.9784506
c	-0.9233978	0.9784506	1.0000000

Figure 4.9: H-D(b): Parameters Estimates

The fitting procedure is relatively easy due to the large amount of data. We will see which one is more appropriate to our data. Two methods are presented here to compare the estimates: graphing techniques and statistics test.

Figure 4.10 shows the graphs for both models: observed height vs predicted height and residuals vs. predicted height. It is obvious that both models are unbiased, while model (a) has less variability than model (b);

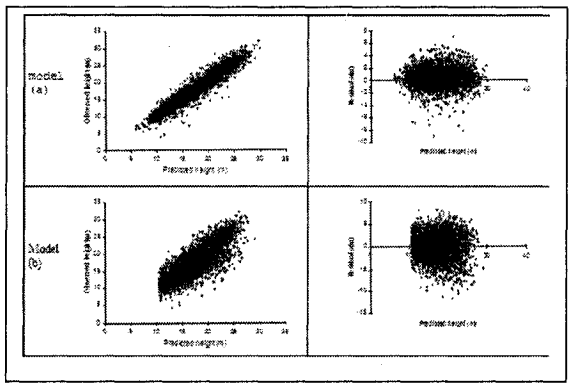


Figure 4.10: H-D: Validate models

Statistics tests: the mean square error for model (a) is 2.0702 as shown in Figure 4.6 and for model (b) is 7.1506 in Figure 4.8.

Model (a) being superior to model (b) for PSP data indicates that trees with same DBH may have different height due to the plots they were located. Trees with a given DBH in the bigger average DBH (hence higher average height) plots will be higher than those with the same DBH value but in smaller average DBH plots.

4.3 Climate-sensitive site index model

Forest site productivity is sensitive to local climate. We try to find some parameters to represent this sensitivity through the use of a process-based model, StandLEAP, which provides stand biomass accrual as an output with climate input. Since climate is plot-specific, these parameters were calibrated at the plot level. Maximum annual increment in biomass is a good indicator of site index, but since about 2/3 of the plots in studies have passed their age with maximum increment, we used the curve of the Mean Annual Increment (MAI) as a function of mean stand age instead. MAI is defined as aboveground merchantable biomass/mean stand age. We simulate the evolution of aboveground biomass for a certain number of years (here we tried 150), as we can see from Figure 4.11, the shape of the curve is quite regular. For this effect, we use a power equation

$$MAI = a * age^b * c^{age}$$

to model MAI and the results show that over 345 plots in studies, 294 plots converge with a pretty good fit (R^2 are 86% or higher). The good fit quality at this step is quite important. Otherwise, we would need to find other equations. After that, we use the parameters we got before (a, b, c at plot level) for calibrating a site index

model.

The site index (SI) model is defined as a function of the parameters and the stand density factor (SDF*)

$$SI = p_0 + P_a * log(a) + P_b * b + P_s * SDF + P_{sa} * (SDF * log(a))$$

We removed parameter c, because b and c were strongly correlated (-0.93).

From Figure 4.13, Figure 4.14 and Figure 4.15, we can see the site index model is unbiased, with good fitness for the statistics and the residuals are homogeneous, around 75% of the variance is explained by the model. We accept the model for the following analysis.

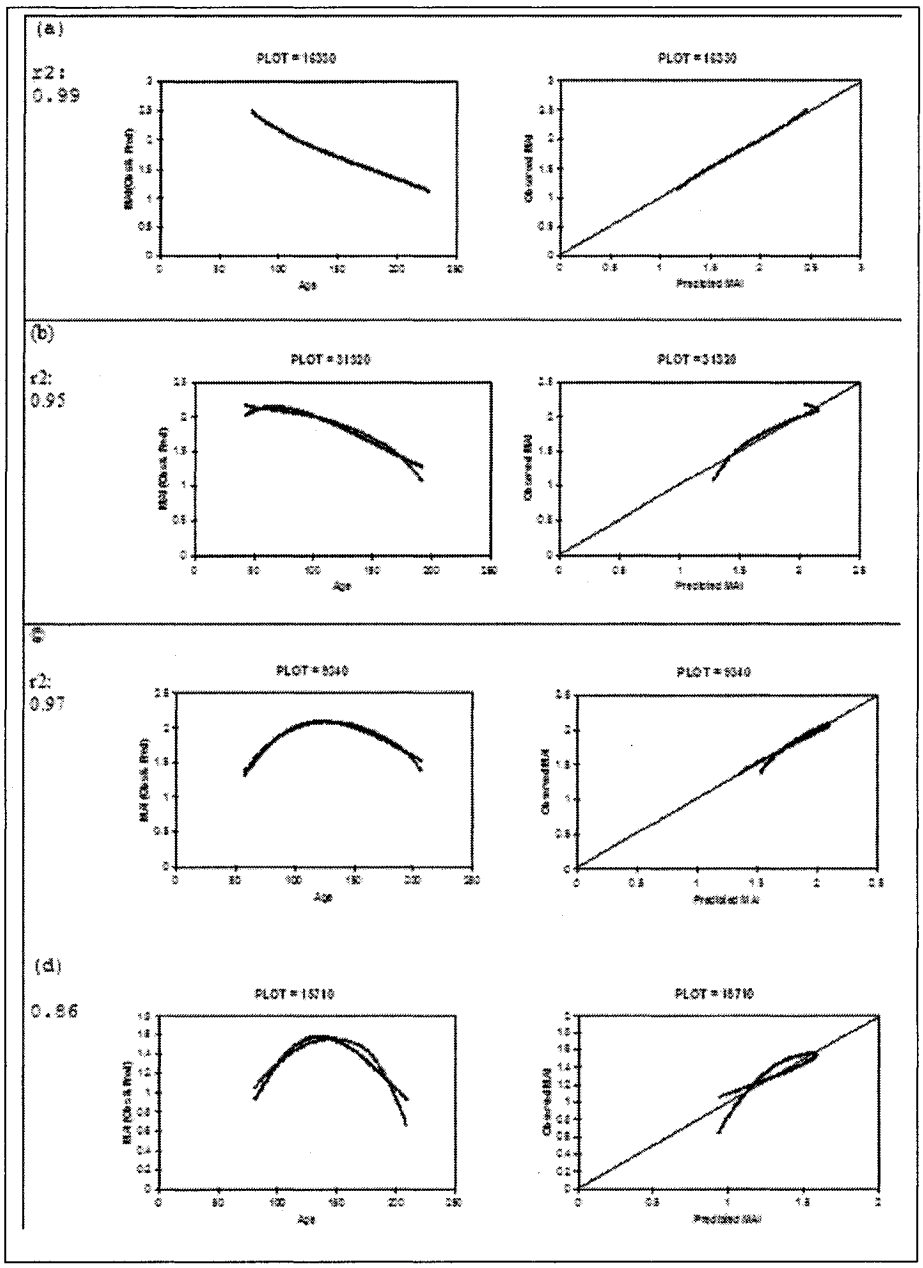


Figure 4.11: SI: Fit of MAI models by plot

Nonlinear SUR Parameter Estimates

Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|	Label
P0	14.62558	0.3510	41.66	<.0001	P0
pa	4.27248	0.1810	23.60	<.0001	pa
Pb	15.71256	0.7041	22.31	<.0001	Pb
Ps	-0.00114	0.000100	-11.40	<.0001	Ps
Psa	-0.00009	9.644E-6	-9.72	<.0001	Psa

Figure 4.12: SI: Parameter Estimates

Nonlinear SUR Summary of Residual Errors

Equation	DF Model	DF Error	SSE	MSE	R-Square	Adj R-Sq
si_bh	5	289	571.0	1.9758	0.7482	0.7447

Figure 4.13: SI: Goodness of fit

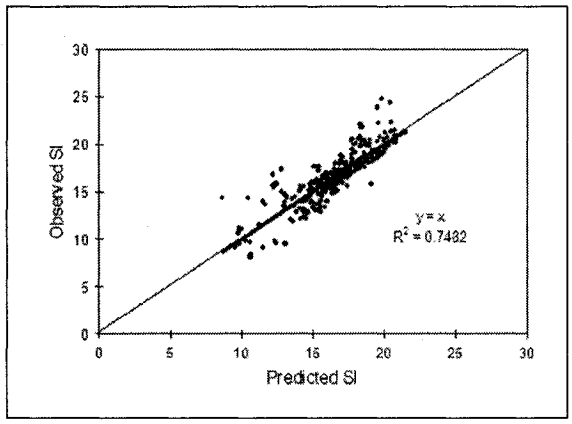


Figure 4.14: SI: Observed vs. Predicted

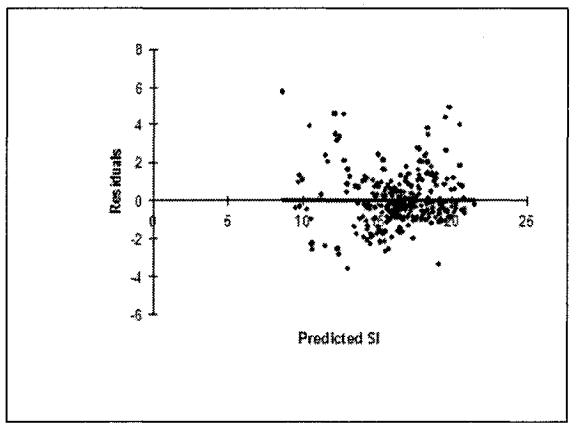


Figure 4.15: SI: Residuals vs. Predicted

4.4 Average site index series

We simulate aboveground biomass for 100 years with StandLEAP under each of the climate year (1901 - 2000). In the following step, we estimate parameters a, b, c in the model $MAI = a * age^b * c^{age}$ for each plot and year. We then use the climate-sensitive site index model built in Section 4.3 to calculate, plot by plot, the evolution of site index as a function of climate. After averaging site index over all the plots, we get a series of average site index per climate year. This series gives a rough idea of the character of the area (Figure 4.16). The curve of average site index±standard deviation is also given in the figure to show the variability of the data. The linear trend is obtained from a linear regression.

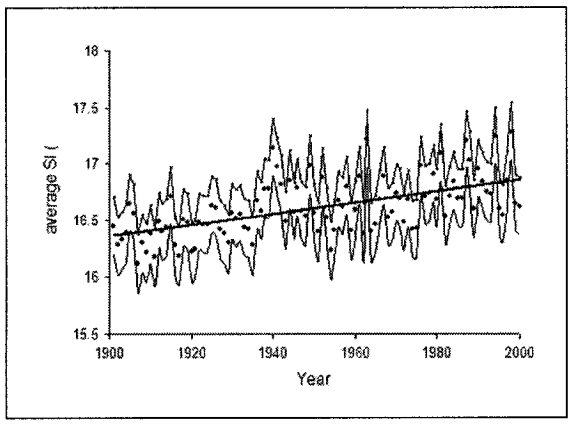


Figure 4.16: Average site index series

All figures given below are generated by PROC ARIMA, a procedure developed directly from the Box-Jenkins methods for fitting ARIMA models (SAS/ETS User's Manual 1999).

4.4.1 Identification

First, we can have a general idea of the data we are going to work with and the basic statistics in Figure 4.17.

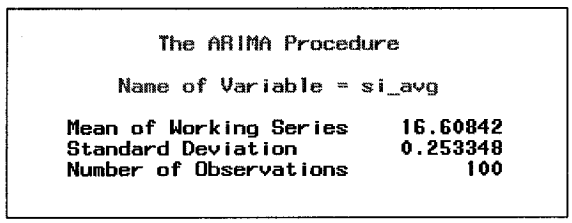


Figure 4.17: Identification: Descriptive Statistics

We start by examining the estimated ACF and PACF for the undifferenced data in Figure 4.18 and Figure 4.19.

Estimated ACF falls to zero slowly (Figure 4.18), indicating that the mean of the data is nonstationary.

We then use augmented Dickey-Fuller test and Phillips-Perron test to detect stationarity.

We use the single mean type in Figure 4.20 because we are testing the mean stationarity. And we failed to reject the nonstationary hypothesis at 5% level by $\tau - test$. This is consistent to what the ACF showed. But Phillips-Perron test gave different results (Figure 4.21).

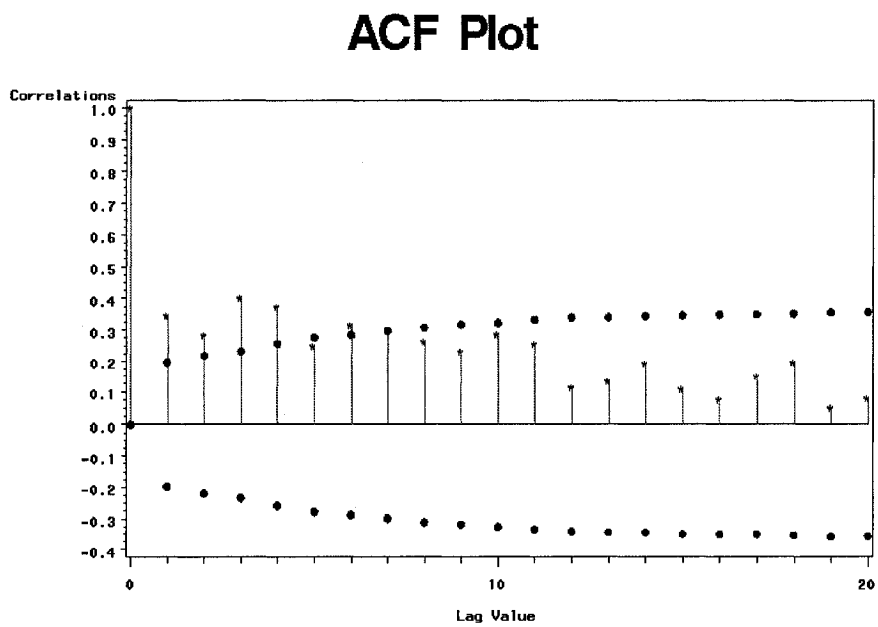


Figure 4.18: Identification: Autocorrelation Function

However, by considering all the tests, it is hard to accept that the site index series is stationary.

We use a Ljung-box test to see if the series can be associated to a white noise. The null hypothesis is that none of the series up to a given lag is significantly different from 0. If it is true, then the series is a white noise and no ARIMA model is needed for the series.

P-value up to lag 24 of the autocorrelation is significant at a .0001 level (Figure 4.22). So the null hypothesis is rejected very strongly, as expected.

Differencing is applied afterwards to check if it can change the series into stationary.

We can see that after the differencing, the ACF falls to zero very quickly, a good

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.34486	|*******|
2	0.18557	|****|
3	0.30354	|******|
4	0.19080	|****|
5	0.01828	| |
6	0.10696	|**|
7	0.05077	|*|
8	0.04756	|*|
9	0.00439	| |
10	0.07741	|**|
11	0.04052	|*|
12	-0.13730	|***|
13	-0.06850	|*|
14	0.01104	| |
15	-0.03014	|*|
16	-0.03104	|*|
17	0.04390	|*|
18	0.12809	|***|
19	-0.06376	|*|
20	-0.01724	| |
21	0.04773	|*|
22	-0.08429	|**|
23	-0.04383	|*|
24	-0.04695	|*|

Figure 4.19: Identification: Partial Autocorrelation Function

Augmented Dickey-Fuller Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	3	0.0271	0.6870	0.56	0.8363		
	4	0.0213	0.6857	0.47	0.8155		
	5	0.0182	0.6849	0.46	0.8129		
Single Mean	3	-11.8970	0.0765	-2.29	0.1766	2.81	0.3641
	4	-10.3824	0.1131	-2.02	0.2791	2.16	0.5261
	5	-7.8098	0.2168	-1.71	0.4236	1.58	0.6721
Trend	3	-47.3949	0.0003	-3.84	0.0185	7.39	0.0264
	4	-64.3488	0.0003	-3.85	0.0179	7.43	0.0254
	5	-71.4586	0.0003	-3.65	0.0306	6.70	0.0446

Figure 4.20: Identification: Augmented Dickey-Fuller Test

sign for a stationary series (Figure 4.23).

Notice that the PACF (Figure 4.24) after lag 3 falls within the two standard deviation region. So we can guess the value of p is 3, if it is a AR(p) model.

Figure 4.25 and Figure 4.26 show that both augmented Dickey-Fuller test and Phillips-Perron test indicate a stationary series after the first differencing.

Figure 4.27 indicates that the white noise hypothesis is still rejected. In case of a white noise (p-value $>$ 0.05), we could have concluded that the first-differenciated

Phillips-Perron Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau
Zero Mean	3	0.0069	0.6825	0.08	0.7071
	4	0.0070	0.6825	0.09	0.7080
	5	0.0078	0.6827	0.11	0.7153
Single Mean	3	-68.2858	0.0009	-6.99	<.0001
	4	-75.6316	0.0009	-7.20	<.0001
	5	-81.1897	0.0009	-7.37	<.0001
Trend	3	-97.9023	0.0003	-9.61	<.0001
	4	-101.626	0.0001	-9.63	<.0001
	5	-102.581	0.0001	-9.64	<.0001

Figure 4.21: Identification: Phillips-Perron Test

The ARIMA Procedure

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	69.21	6	<.0001	0.345	0.282	0.402	0.371	0.247	0.312
12	110.32	12	<.0001	0.297	0.260	0.230	0.285	0.254	0.110
18	126.31	18	<.0001	0.133	0.191	0.108	0.074	0.150	0.196
24	130.68	24	<.0001	0.049	0.078	0.152	0.018	-0.007	0.045

Figure 4.22: Identification: White Noise Test

site index time series followed a random walk process. No ARIMA model would then have been necessary for estimating it.

4.4.2 Estimation and diagnostic check

After getting an stationary series, we use the Box-Jenkins method to determinine the value of p and q in an ARIMA(p, d, q), here $d = 1$ as we discovered before.

The basic idea to find tentative models using Box-Jenkins method is to try a series of p and q values. If the set of (p, q) can pass the statistical tests, we accept it as a tentative model. Then we analyze the tentative models in detail and compare the statistics to select the best model. We implement four methods to tentative determine the values of p and q of an ARMA models — ESACF, SCAN, ODQ and CORNER.

The Extended Sample Autocorrelation Function (ESACF) and The Smallest CANon-

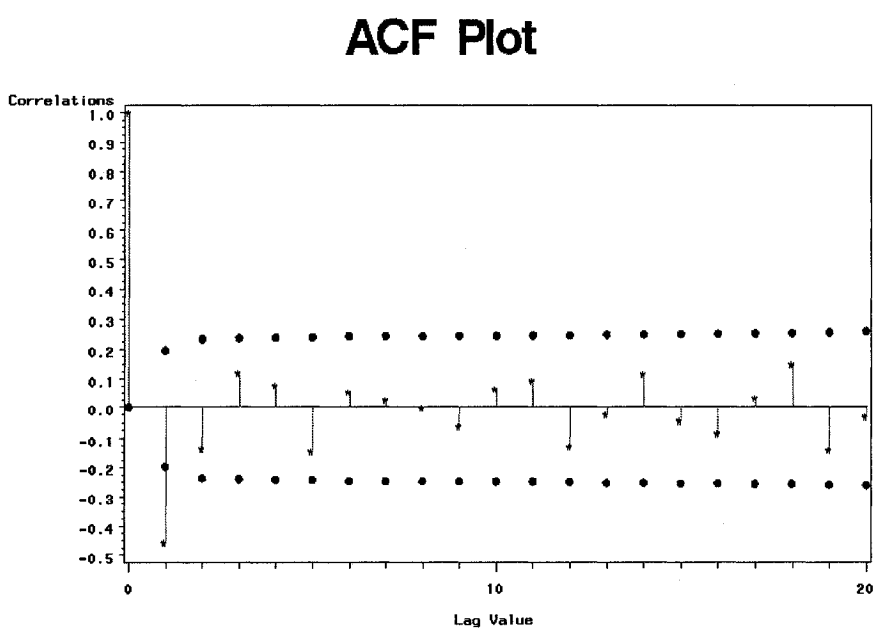


Figure 4.23: Identification(1st difference): Autocorrelation Function

ical (SCAN) methods were proposed by Tsay and Tiao (1984, 1985). The procedures were built in PROC ARIMA.

The Order Determination Quantity (ODQ) was proposed by Zhang H.M and Wang P. in 1994. It discussed the case separately for $ODQ > 0$ and $ODQ < 0$ instead of minimizing. The corner method was proposed by Beguin et al. in 1980. The procedures of ODQ and corner methods were implemented by Dominique Ladiray[1].

By checking convergence and residuals, two models pass the tests: ARIMA(1,1,1) and ARIMA(3,1,0). Table 4.1 Compares the statistics. ARIMA(1,1,1) wins for smaller variance, AIC and SBC.

The estimation of an ARIMA(1,1,1) is given in Figure 4.28. the parameter es-

[1] *http : //www.unige.ch/ses/sococ/eda/sas/MacrosSAS.pdf* Dominique LADIRAY : Diverses Macros SAS: Analyse Exploratoire des Donnees, Analyse des Series Temporelles .

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.45725	**********| .
2	-0.44044	**********| .
3	-0.27663	******| .
4	-0.07810	. **| .
5	-0.15125	.***| .
6	-0.10131	. **| .
7	-0.09834	. **| .
8	-0.06172	. *| .
9	-0.12719	.***| .
10	-0.09363	. **| .
11	0.09724	. |** .
12	0.03409	. |* .
13	-0.04433	. *| .
14	0.00590	. | .
15	-0.00487	. | .
16	-0.08310	. **| .
17	-0.16292	.***| .
18	0.03164	. |* .
19	-0.01573	. | .
20	-0.08773	. **| .

Figure 4.24: Identification(1st difference): Partial Autocorrelation Function

Augmented Dickey-Fuller Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	3	450.9296	0.9999	-7.56	<.0001		
	4	183.1439	0.9999	-6.73	<.0001		
	5	111.7496	0.9999	-6.06	<.0001		
Single Mean	3	434.7644	0.9999	-7.55	<.0001	28.50	0.0010
	4	179.5716	0.9999	-6.72	<.0001	22.59	0.0010
	5	108.6366	0.9999	-6.08	<.0001	18.51	0.0010
Trend	3	434.7866	0.9999	-7.50	<.0001	28.16	0.0010
	4	179.9844	0.9999	-6.68	<.0001	22.34	0.0010
	5	108.6309	0.9999	-6.05	<.0001	18.30	0.0010

Figure 4.25: Identification(1st difference): Augmented Dickey-Fuller Test

timated for MA1 is almost 1. This can be dangerous, since PROC ARIMA tends to converge on the invertibility boundary if the data are differenced and a moving average model is fit (SAS/ETS User's Manual 1999). Recall that the condition for a MA(1) to be invertible is $\left|\frac{1}{\theta_1}\right| > 1$, we should be very careful with a coefficient value around 1. We try other computational methods, unconditional least squares (ULS) and maximum likelihood (ML). But the results do not improve. We cannot even get convergence with ULS. We estimate the ARIMA(3,1,0) model with three methods. The results show that there is very little differences in parameter estimates among

Phillips-Perron Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau
Zero Mean	3	-118.142	0.0001	-22.25	<.0001
	4	-117.523	0.0001	-22.64	<.0001
	5	-115.021	0.0001	-24.59	<.0001
Single Mean	3	-118.088	0.0001	-22.16	<.0001
	4	-117.458	0.0001	-22.56	<.0001
	5	-114.942	0.0001	-24.52	<.0001
Trend	3	-118.128	0.0001	-22.03	<.0001
	4	-117.503	0.0001	-22.42	<.0001
	5	-114.984	0.0001	-24.37	<.0001

Figure 4.26: Identification(1st difference): Phillips-Perron Test

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	27.88	6	<.0001	-0.457	-0.139	0.117	0.076	-0.145	0.051
12	31.72	12	0.0015	0.020	-0.006	-0.063	0.061	0.093	-0.130
18	37.22	18	0.0049	-0.027	0.113	-0.045	-0.087	0.030	0.147
24	45.05	24	0.0057	-0.141	-0.033	0.163	-0.083	-0.060	0.059

Figure 4.27: Identification(1st difference): White Noise Test

the methods, and the residuals from each method are white noises. So it is safer to take ARIMA(3,1,0) model.

Figure 4.29 to Figure 4.31 show the results of estimation for ARIMA(3,1,0) model. Figure 4.29 gives the parameters and test statistics of the estimation for ARIMA(3,1,0) model using the conditional method. Figure 4.30 shows the residuals check using Ljung-box statistic and the white noise hypothesis cannot be rejected. Figure 4.31 gives the form of the model and the estimated mean is 0.004323 m.

4.4.3 Forecasting

Finally, the 50-year ahead forecasting is given by Figure 4.32 together with a 95% confidence interval. The increasing trend is obvious and will continue for a certain period of time.

Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.0037225	0.0009773	3.81	0.0002	0
MA1,1	0.99999	0.02360	42.38	<.0001	1
AR1,1	0.06077	0.10566	0.58	0.5665	1

Constant Estimate	0.003496
Variance Estimate	0.046503
Std Error Estimate	0.215646
AIC	-19.8514
SBC	-12.066
Number of Residuals	99

* AIC and SBC do not include log determinant.

Figure 4.28: Estimation: ARIMA(1,1,1) Parameter Estimates

Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.0043229	0.0081644	0.53	0.5977	0
AR1,1	-0.80932	0.09874	-8.20	<.0001	1
AR1,2	-0.67462	0.11252	-6.00	<.0001	2
AR1,3	-0.30446	0.10141	-3.00	0.0034	3

Constant Estimate	0.012054
Variance Estimate	0.050433
Std Error Estimate	0.224574
AIC	-10.8565
SBC	-0.47602
Number of Residuals	99

* AIC and SBC do not include log determinant.

Figure 4.29: Estimation: ARIMA(3,1,0) Goodness of fit

ARIMA	Variance	STD	AIC	SBC
(3,1,0)	0.050433	0.224574	-10.8565	-0.47602
(1,1,1)	0.046503	0.215646	-19.8514	-12.066

Table 4.1: Estimation: models comparation

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.84	3	0.1842	-0.022	-0.070	-0.109	-0.142	-0.087	-0.035
12	8.23	9	0.5111	-0.049	0.026	0.042	0.081	0.079	-0.111
18	11.91	15	0.6860	-0.083	-0.005	-0.066	-0.052	0.018	0.126
24	15.52	21	0.7962	-0.040	0.027	0.100	-0.069	-0.101	-0.025

Figure 4.30: Estimation: ARIMA(3,1,0) Residuals check

Model for variable si_avg

Estimated Mean	0.004323
Period(s) of Differencing	1

Autoregressive Factors

Factor 1: 1 + 0.80932 B**(1) + 0.67462 B**(2) + 0.30446 B**(3)

Figure 4.31: Estimation: ARIMA(3,1,0) Models for average site index

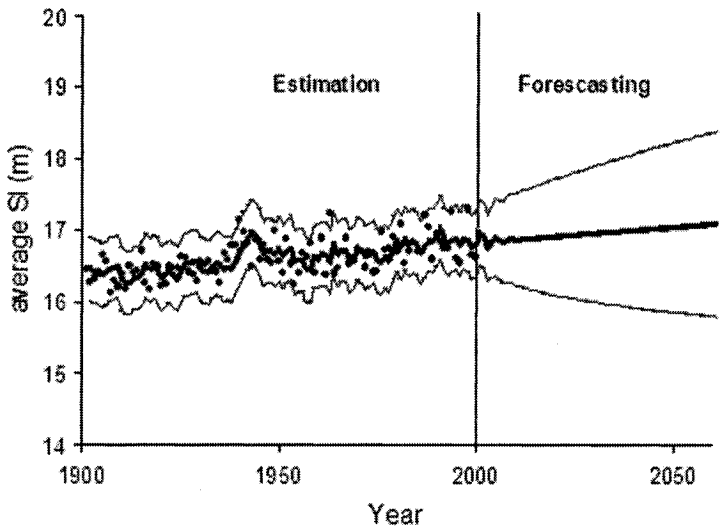


Figure 4.32: Forecasting: ARIMA(3,1,0) Models for average site index

Chapter 5

Discussions and future work

5.1 Height-DBH model

The height-DBH model studied by Bégin and Raulier (1995) gives a reasonable explanation in forest science. and it has passed the main statistical tests. By graphing, it looks unbiased, and has less variability than model (b). It seems to be a good model. However, if we carefully look at the model (Equation (4.1)),

$$H = 1.3 + \frac{D}{\frac{\bar{D}}{\bar{H}-1.3} + \beta_2 * (D - \bar{D})},$$

we can find that the average height appearing on RHS as a regressor is a function of the response, height. This violates the assumption that the regressor should be observed independently of the response.

In fact, if we add one more parameter (β_1) to replace the problematic item,

$$H = 1.3 + \frac{D}{\beta 1 + \beta_2 * (D - \bar{D})},$$

it also gives a pretty good estimate (Figure 5.1). Though it has greater mean squared error than Equation (4.1), it is more statistically acceptable.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	2	17355521	8677760	1303423	<.0001
Error	52685	350759	6.6577		
Uncorrected Total	52687	17706280			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
beta1	1.1471	0.000771	1.1455	1.1486
beta2	0.0292	0.000163	0.0289	0.0295

Figure 5.1: H-D(revised): Estimated parameters

Another minor problem with H-D model is, that the observations within a plot are not independent, but this problem can be negligible due to two facts: Only a few trees were selected for the height measurements in each plot. Compared to the large amount of data in the whole data set, the within plot dependence among trees is small; We could have kept only one tree in a plot to remove the intrinsic dependence. But the sampling itself could also have brought a bias to the model.

However, despite of all these practical issues, Bégin and Raulier (1995) have shown that, the one-parameter model ((Equation (4.1)) is better than the two-parameter model (Equation (5.1)) through jack-knife procedure.

5.2 Climate-sensitive site index model

Let us review the climate-sensitive site index model developed in section 4.3. Figure 4.14 is put here for convenience.

Site index values vary from 8 m to 22 m in our model. This variability is compa-

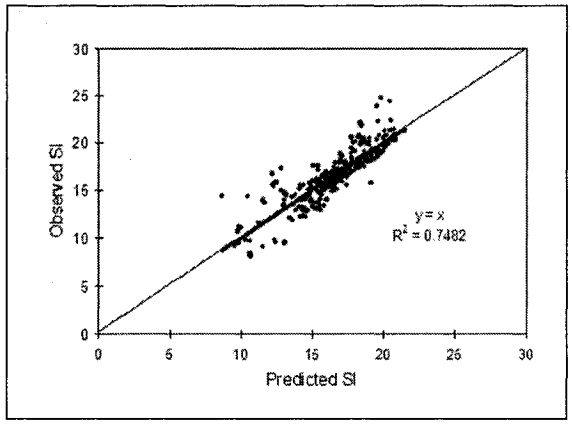


Figure 5.2: SI: Observed vs. Predicted

rable to the values appearing in Huang et al. (2001)'s yield tables for lodgepole pine stands in Alberta. In Huang's tables, fifteen site index classes are given (6 m - 25 m), the difference between the site index classes is by 1 meter from 10 m to 20 m. This is to say that the data used for calibrating our climate-sensitive site index model has covered most of the observed range for the pure logdepole pine plots in Alberta PSP, hence insuring some reliability. The wide range of values assure us to have included sufficient information in the model.

The range problem may happen if for any reasons the data was missing for a period of value range. This is possible due to lack of measurement of age or other variables for StandLEAP. An extreme case could have happened if we only have data at the two ends and we could not have drawn right conclusion on such a narrow range.

5.3 Some problems with unit root tests

The ADF and PP tests are asymtotically equivalent but may differ substantially in finite samples due to the different ways they correct for serial correlation for high order ARMA models. Schwert (1989) found out that if Δy_t has an ARMA representation with a large and negative MA component, the ADF and PP tests are severely size distorted (reject I(1) null much too often when it is true) and that the PP tests are more size distorted than the ADF tests. Perron and Ng (1996) suggested useful modifications to the PP tests to reduce this size distortion.

In general, the ADF and PP tests have very low power against I(0) alternatives that are close to being I(1). That is, unit root tests cannot distinguish highly persistent stationary processes from nonstationary processes very well.

5.4 Time series analysis methods

5.4.1 Box-Jenkins method

The univariate Box-Jenkins approach is not the only way to analyze time series. We choose it as the principle method because it has three advantages over many other traditional methods (Pankratz 1983):

1. The concepts associated with univariate Box-Jenkins models are derived from a solid foundation of classical probability theory and mathematical statistics.

2. Box and Jenkins have developed a strategy that guides the analyst in choosing one or more appropriate models out of the large family of ARIMA models.

3. It can be shown that an appropriate ARIMA model produces optimal univariate forecasts. That is to say that no other standard single-series model can give forecasts with a smaller mean-squared forecast error.

However, the univariate Box-Jenkins method has its own restrictions, like any other statistical methods, even though it has been more widely used than the others:

Short-term forecasting This is because most ARIMA models place heavy emphasis on the recent past rather than the distant past;

Data type The univariate Box-Jenkins method only applies to stationary data spaced at a discrete equally time interval. and the observation in the series are assumed to be sequentially related.

Sample size Building an ARIMA model requires an adequate sample size, Box-Jenkins method requires at least about 50 observations.

And also, as we have known before, if the time series is not correlated, no ARIMA model can fit the series.

5.4.2 Spectral analysis

In spectral analysis, a data series can be expressed by the finite Fourier transform. The Fourier transform decomposition of the series y_t can be written as:

$$y_t = \frac{a_0}{2} + \sum_{k=1}^{q} [a_k cos(2\pi\omega_k t) + b_k sin(2\pi\omega_k t)]$$

where

N is the number of observations in the series;

q is the number of frequencies in the Fourier decomposition:

q = [N/2], where [.] denotes the integer part;

a_0 is the mean term: $a_0 = 2\bar{y}$;

a_k is the cosine coefficient;

b_k is the sine coefficient;

ω_k is the Fourier frequencies: $\omega_k = \frac{k}{N}$.

The amplitude periodogram J_k is defined as:

$$J_k = \frac{N}{2}(a_k^2 + b_k^2) \qquad k = 1, 2, ...q. \tag{5.1}$$

If we allow the frequency ω to vary continuously in the range 0 to 0.5 cycle, the definition of periodogram in (5.1) is referred as sample spectrum (5.2).

$$I(\omega) = \frac{N}{2}(a_\omega^2 + b_\omega^2). \qquad 0 \leq \omega \leq \frac{1}{2}. \tag{5.2}$$

The autocovariance and sample spectrum are in fact transformable, as shown in Box and Jenkins (1994).

$$I(\omega) = 2\left[\hat{\gamma}_0 + 2\sum_{k=1}^{N-1}\hat{\gamma}_k cos(2\pi\omega k)\right] \quad 0 \leq \omega \leq \frac{1}{2}$$

where $\hat{\gamma}_0$ is the variance of the series and $\hat{\gamma}_k$ is the autocovariance at lag k defined in Section 3.2.3.

Another application of the spectral analysis is to identify periodicities in the time series using periodograms J_k. This can be done by PROC SPECTRA, one of the procedures used for spectral analysis. More details can be found in SAS/ETS User's Manual (1999).

A preliminary result showed that there is a 3.4-year cycle existing in the increment of average site index series (Figure 5.3). To explain how and why this cycle is formed may involve more exploration in the climate-related studies.

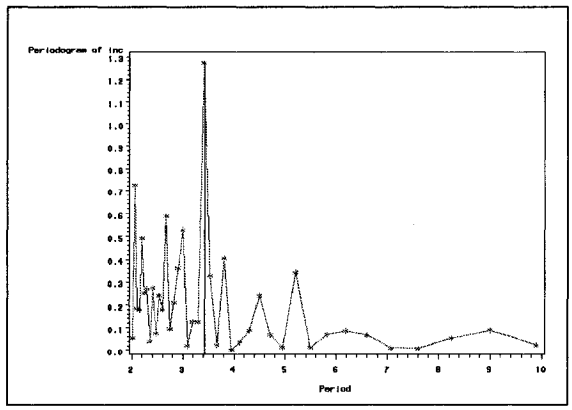


Figure 5.3: Spectra: Periodogram vs. Period

Other empirical methods like spline or interpolation have applications in some area of time series analysis. But compared to Box-Jenkins method, they are derived in more of an ad hoc or intuitive way.

5.5 Two general issues in forest modelling

Due to the increasing applications of models in decision making, model credibility is becoming increasingly important in forest management. Model validation and

usefulness are two general questions which have called more and more attentions to both model builders and model users.

5.5.1 Model validation

Model validation is one of the most effective ways to enhance model credibility.

Basically, the goals of model validation is to ensure that model predictions reflect the most likely outcome in reality and gain sufficient confidence about a model. The idea behind a model validation seems simple, while it is in fact one of the most complex topics associated with model building. Huang et al. (2003) studied in detail model validation in growth and yield models and suggested a "comprehensive" approach. It involves

1. independent validation data;
2. graphical validation;
3. calculation of validation statistics;
4. examination of the biological and theoretical validity;
5. environmental factor;
6. consideration on the practicality and operability of a model;
7. a third party validation.

The "comprehensive" method may seem applausible, but it cannot always be carried out completely due to practical reasons. Usually, graphical validation and

statistical tests are the two most commonly used methods. Other options are also used from time to time by modelers.

"Cross-validation" , a method that reserves a certain portion of data for validation, seems to be a reasonable method. It mostly validates the sample scheme. But it is not recommended for model fitting (Kozak et al 2003).

Checking the assumptions of the statistical tests before applying them is important. Otherwise, we may get conflicting answers from different statistical tests with the same goal. But it is not always easy to decide which test to take because of the limitation of our knowledge on the data. A "smart" way could be to selectively choose the tests that best suit our purposes. This can increase the confusion during validating.

5.5.2 Utilizing models: What kind of models are really expected?

More and more foresters recognize that forest models play a crucial role in forest management decision making. However, some models can only be used for research even though they have been well validated. Many models emphasize modelling for prediction instead of modelling for understanding, which results in a large gap between models in theory and models in practice.

What kind of models are really expected in reality? According to Aramo (2003): modelers are called on to develop models that address more of the operational concerns, and to build models that can solve real world problems. Simple and practical

models are more appreciated than complex and idealistic models. The ideal models should be relatively easy to use, have reasonable statistical testing values, and are fairly consistent and robust under the wide range of conditions that one may encounter in practice.

Decision makers care more about the overall performance of a model. Good individual models do not necessarily result in good overall performance, partial due to an incomplete understanding of the interactions among the model components.

Chapter 6

Conclusion

Most of growth and yield models currently in use assume that past growth conditions, soils, climate and disturbance regimes, stay relatively stable in the foreseeable future. This seems unlikely, however, for lodgepole pine in the Permanent Sample Plots in Alberta, either by observations or the analysis we made in this study. The climate warming could be partially responsible for the average site index increase. As a result, traditional models might provide unreliable results in forecasting future yields.

The increasing trend we observed in Section 4.4 is important, though the annual increment of 4.323 mm may seem small comparing to a height of 16 meters. To gain 1 meter of increment will take over 200 years (1 meter is the difference of site index class of the growth and yield tables in Huang S. (2001)). But if we consider the time scale, 100 to 200 years, on which a forest management plan is made, or that an average lodgepole pine can live up to 400 to 600 years, the influence is significant.

The period for the accumulation of the increase can most probably be shorten considering that the climate data we used for the analysis were interpolated from 1930

to 1990, and the temperature could increase more rapidly in the following century based on IPCC assessment report. So it is logical to expect that the influence from climate changes in reality will be stronger than what was predicted by our analysis.

The present study is an important attempt to discover the regional climate influence on the forest productivity. We believe that with the ongoing developments in climate data interpolation techniques, more detailed processed-based models will be developed to capture future growth trend under a warmer climate.

Appendix A

Glossary of terms

A few terms definition in forest management we have used in this text.

Basal Area	the area of the cross-section of tree stems. The value for a tree is $\pi * (DBH/2)^2$.
Biomass	the dry weight of all organic matter in a tree.
Breast height	the standard height, 1.3 m above ground level.
Breast height age	total age less the year that a tree takes to attain the height of 1.3 meters.
DBH	the stem diameter of a tree measured at breast height, 1.3 meters above the ground.
Stand density factor	the stand density at a reference breast height age of 50 years.
Top height	the average height of the 100 largest diameter trees per hecta.
Site index	the top height at 50 years breast height age.
Hardiness zone	The Plant Hardiness Zones map outlines the different zones in Canada where various types of trees, shrubs and flowers will most likely survive. It is based on the average climatic conditions of each area.

References

[1] Akaike, H. (1974), A new look at the statistical model identification. *IEEE Transaction on Automatic Control AC-19, 716-723.*

[2] Amaro A. et. al (2003), Modelling forest systems. *CABI Publishing.*

[3] Alberta Land and Forest Division (2002), Permanent sample plot (PSP) field procedures manual. *Alberta Sustainable Resource Development, Land and Forest Division.*

[4] Bégin J. and Raulier F. (1995), Comparasion de différentes approches, modèles et tailles d'échantillons pour l'établissement de relations hauteur-diamètre locales. *Can J. For. Res. 25, 1303 - 1312 .*

[5] Bates D.M. and Watts D.G. (1988), Nonlinear regression analysis and its applications. *John Wiley Sons.*

[6] Beguin J.M., Gouriéroux C. and Monfort A. (1980), Identification of a mixed autoregressive-moving average process: the corner method. *Time Series (ed. Anderson O.D.) Amsterdam, North Holland, 423-31.*

[7] Bernier P.Y., Raulier F., Stenberg P. and Ung C-H. (2001), The importance of needle age and shoot structure on canopy net photosynthesis of Balsam fir (Abies balsamea): a spatially-inexplicit modelling analysis. *Tree Physiol. 21, 815-830.*

[8] Borders, B.E., Bailey, R.L. and Clutter, M.L. (1987), Forest growth models: parameter estimation using real growth series. Ek, A.R., Shifley, S.R., Burk, T.E.(Eds), Forest Growth Modelling and Prediction. *USDA Forest Service General Technical Reports NC-120, 660-667.*

[9] Box, G.E.P. and Jenkins G.M. (1976), Time series analysis: Forecasting and control. *Holden-Day.*

[10] Box, G.E.P. and Jenkins G.M. (1994), Time series analysis: Forecasting and control. *Prentice-Hall.*

[11] Chen Y. (2002), On the robustness of Ljung-Box and Mcleod-Li Q Test: A simulation study. *Economics Bulletin Vol.3, No.17, 1-10.*

[12] Davies N., Trigg, C.M. and Newbold P. (1977), Significance levels of the Box-Pierce portmanteau statistic in finite samples. *Biometrika 64, 517 - 522.*

[13] Davies N. and Newbold P. (1979), Some power studies of a portmanteau test of time series model specification. *Biometrika 66, 1, 153-155.*

[14] Davis, L.S., Norman Johnson, K., Bettinger, P.S. and Howard T.E. (2001), Forest management: to sustain ecological, economic, and social values. *McGraw-Hill.*

[15] Dewar R.C. (1997), A simple model of light and water use evaluated for Pinus radiata. *Tree Physiol. 17, 259-265.*

[16] Dickey D.A. (2005), Stationarity issues in time series models. *SUGI 30*

[17] Drapper N.R. and Smith H. (1998), Applied regression analysis. *John Widey Sons.*

[18] Fuller W.A. (1996), Introduction to statistical time series.*John Wiley, New York.*

[19] Gallant A.R. (1987), Nonlinear statistical models. *John Wiley Sons.*

[20] Gertner, G. (1987), Approximating precision in simulation projections: an efficient alternative to Monte Carlo methods. *For. Sci. 33 (1), 230-239.*

[21] Gordon D. Nigh (2004), Climate and productivity of major conifer species in the interior of British Columbia, Canada. *Forest Science 50 (5), 659-671.*

[22] Hall R.J., Price D.T., Raulier F., Arsenault E., Bernier P.Y., Case B.S. and Guo X. (2005), Integrating remote sensing and climate data with process models to map forest producivity: Ecoleap-west.

[23] Hamilton, J. (1994), Time series analysis. *Princeton University Press, New Jersey.*

[24] Huang S. (1994), Ecologically based individual tree volume estimation for major Alberta tree species, Report #1 *Alberta Environmental Protection Pub. No.: T/288*

[25] Huang, S. (1997), Development of a subregion-based compatible height-site index-age model for black-spruce in Alberta. *Forest Management Research Note No. 5. Pub. No. T/352. Edmonton, Alberta, P.55.*

[26] Huang S., Morgan D.J., Klappstein G., Heidt J., Yang Y. and Greidanus G. (2001), GYPSY: A growth and yield projection system for natural adn regenerated lodgepole pine stands within Alberta and ecologically based, enhanced forest management framework: yield tables for seed-origin in natural and regeneratd lodgepole pine stands. *Alberta Sustainable Resource Development, Technical Report Publication No. T/485, Edmonton, Alberta.*

[27] Huang S., Yang Y. and Wang Y. (2003), A critical look at procedures for validating growth and yield models *CABI.*

[28] Johnsen K., Samuelson L., Teskey R. McNulty S. and Fox T. (2001), Process models as tools in forestry research and management *Forest science, 47 (1).*

[29] Klugman S. A. (2003), Estimation, evaluatin and selection of actuarial models.

[30] Kozak, A. (1997), Effects of multicollinearity and autocorrelation on the variable-exponent taper functions. *Can. J. For. Res. 27 (5), 619-629.*

[31] Kozak A. and Kozak R. (2003), Does cross validation provide additional information in the evaluation of regression? *Canadian Journal of Forest Research 33, 976-987.*

[32] Lambert M.-C., Ung C.-H., and Raulier F. (2005), Canadian national above-ground biomass equations. *Can. J. For. Res. In Press.*

[33] Landsberg J.J. and Gower S.T. (1997), Applications of physiological ecology to forest management. *Academic Press, San Diego.*

[34] Ljung G.M. and Box G.E.P. (1978), On a measure of lack of fit in time series models. *Biometrika 65, 2, 297-303.*

[35] McKenney D.W., Hutchinson M.F., Kesteven J.L. and Venier L.A. (2000), Canada's plant hardiness zones revistied using modern climate interpolation techniques. *Can. J. Plant Sci. 81, 129-143.*

[36] Ng, S., and Perron, P. (1995), Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association 90, 266-281.*

[37] Pankratz, A. (1983), Forecasting with univariate Box-Jenkins models:concepts and cases *New York: John Wiley Sons.*

[38] Phillips, P.C.B., and Perron, P. (1988), Testing for unit roots in time series regression. *Biometrika 75, 335-346.*

[39] Pindyck R.S. and Rubinfeld D. L. (1998), Econometric models and economic forecsts.*Mcgraw-Hill.*

[40] Prodon M. (1968), Forest biometrics. *Pergamon Press.*

[41] Raulier F., Bernier P.Y. and Ung C-H. (2000), Modeling the influence of temperature on monthly gross primary productivity in a sugar maple stand. *Tree Physiol. 20, 333-345.*

[42] Rouduex, J. (1993), La mesure des arbres et des peuplements forestiers *Les Presses Agronomiques de Gembloux.*

[43] SAS institute (1999), SAS/STATS user's manual v8.

[44] SAS institute (1999), SAS/ETS user's manual v8.

[45] Schwarz, G. (1978), Estimating the dimension of a model *Annals of Statistics, 6, 461-464.*

[46] Sen A. and Srivastava M. (1990), Regression analysis: theory, methods, and applications. *Springer-Verlag.*

[47] Stewart R.B., Wheaton, E, and Spittlehouse D. (1998), Climate change: implicagtions for boreal forest.*SRC Publication No. 10442-4D98.*

[48] Tsay R.S. and Tiao G.C. (1984), Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models. *Jounal of the American Statistical Association 79, 385, 84-96.*

[49] Tsay R.S. and Tiao G.C. (1985), Use of canonical analysis in time series model identification. *Biometrika, 72, 2, 299-315.*

[50] Yang Y., Titus S.J. and Huang S. (2003), Modeling individual tree mortality for white spruce in Alberta. *Ecological modelling 163, 209-222.*

[51] Yang Y., Monserud R.A. and Huang S. (2004), An evaluation of diagnostic tests and their roles in validating forest biometric models. *Canadian Journal of Forest Research 34, 619-629.*

[52] Zhang H.M. and Wang P. (1994), A New way to estimate orders in time series. *Journal of Time Series 15 (5), 545-559.*