Growth Analysis Using Nuisance Baseline

Ken-ichi Kamo\textsuperscript{1*}, Tetsuji Tonda\textsuperscript{2}, Kenichi Satoh\textsuperscript{3}

Abstract: In the growth analysis, when the research focus is environmental factor, the longitudinal growth part is essential but not our main interest. In such a situation, by regarding age dependent growth behavior as baseline, we can reconstruct the models to include a nuisance baseline. Such an approach makes it possible only to estimate parameters of interest (environmental factors) without information about the nuisance baseline. After estimating the main parameters, we can graph the baseline trend, non-parametrically. In this paper, the growth model was generated using data on Sugi (Cryptomeria japonica) at Hoshino village in Japan. The results from this study showed an inverse relationship between altitude and tree growth, without modeling for longitudinal growth, i.e., higher altitude resulted in less tree growth. In practice the results can be explained by our knowledge based on two established facts. The first is based on the direction of water and nutrients flow from a higher elevation and accumulates at a lower elevation, and the second can be explained by the fact that trees are suppressed by strong winds at higher elevations. By using this data, we compared our model with traditionally used parametric growth models which are constructed without nuisance baseline. In statistical terms, both the variance of residual and the standard error for parameters in the proposed model were found to be the smallest among the other parametric alternatives. This implies that the estimate is stable in our model that is one of the advantage for statistical inference aspect.

Keywords: growth analysis, nuisance baseline, regression models

1. Introduction

The objective in this paper is to construct a new statistical approach for modeling the growth behaviour of a forest stand. A review of literature shows that growth behaviour is affected by several external factors (Ryan and Yoder 1997; Yin \textit{et al.}, 2004). For example, Nelder experimental plot (Nelder, 1962) suggests that, density affects growth behaviour (Imada, 1997). Such external factors can be integrated as explanatory variables. To investigate the relationship between these factors and the growth behaviour, there are several approaches to modelling the growth behaviour of a forest stand. In order to estimate the relationships among variables, regression analysis is a powerful statistical process for. In the regression framework, there are various approaches to modelling depending on the availability of data or situation. For example, logistic regression is a powerful statistical way of modeling a binomial outcome (e.g. it can a value of 0 or 1 like harvesting or not harvesting), with one or more explanatory variables. When more than two possible discrete outcomes are considered, multinomial logistic regression is employed (Kleinbaum and Klein, 2010). Poisson regression is a form of regression analysis used to model count data and contingency tables. (Cameron and Trivedi, 2013). Here we focus on the regression models for growth analysis. Since traditionally used growth models are parametric type, we need to assume many parts a priori, for example growth function. The more assumption set, the more risk for misleading the result when the assumption is not suitable. In order to solve such problem, we apply nuisance baseline concept to growth analysis. For explaining the overview of our model, we would start from the simplest modeling approach.

Since growth behavior is longitudinal, time (tree age) is considered as the main factor. From this viewpoint, the simplest growth model can be constructed by using a nonlinear growth function. Several types of growth functions have been proposed in literature, depending on their derivation process or concept. Previous studies by Zeide (1993) introduced twelve types of growth functions. For example, when Chapman-Richards function (Richards, 1958) is adopted, the growth model is set as:

\[ E[y(t)] = \alpha(1 - e^{-\beta t})^\gamma \]

where \( y(t) \) is the growth amount at age \( t \). Here \( \alpha, \beta \) and \( \gamma \) are the positive parameters determining the shape of growth function; \( \alpha \) controls the upper bound, \( \beta \) controls the speed (growth rate), and...
Growth Analysis Using Nuisance Baseline

\( \gamma \) controls the curvature of growth curve. This model can describe the sigmoid shape for growth amount against age.

Although, time (age) should be the main factor under consideration, there may be other external factors, for example, environmental factors. As an illustration, let \( x_{i,j} \) be the \( j \)-th kind of explanatory variable in \( i \)-th sample and \( y_i \) be the growth amount of \( i \)-th sample \( (i = 1, \ldots, n \text{ and } j = 1, \ldots, k) \), then the model is constructed as:

\[
E[y_i] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j}
\]

where \( \beta_0, \ldots, \beta_k \) are the unknown parameters. Since model [2] does not include time related element; therefore, this model cannot be applied to repeated measurement. What model should be suitable when both longitudinal growth and external factors are observed? One simple solution is to use a multiple linear regression model, defined as:

\[
E[y_i(t)] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} + \beta_{k+1} t
\]

But, there lies a problem with this solution approach. The nonlinear nature of the growth function, suggests that the linear assumption on longitudinal behavior is not suitable. Another solution approach is to combine the two models [1] and [2], that is,

\[
E[y_i(t)] = \alpha(1 - e^{-\beta t})^\gamma + \sum_{j=1}^{k} \beta_j x_{i,j}
\]

The model [4] can be regarded as the extension version of [2], where the intercept term \( \beta_0 \) in [2] changes to the varying coefficient for age \( t \). Here we summarize the advantage and disadvantage for the above four traditional models (Table 1).

### Table 1. Summary for alternative models.

| Formula | Advantage | Disadvantage |
|---------|-----------|--------------|
| [1] \( E[y(t)] = \alpha(1 - e^{-\beta t})^\gamma \) | Sigmoid shape expression | No explanatory variable |
| [2] \( E[y_i] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} \) | Explanatory variables | Cannot apply to repeated measurement, Linear only |
| [3] \( E[y_i(t)] = \beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j} + \beta_{k+1} t \) | Both explanatory variables and time dependent variable | Linear only |
| [4] \( E[y_i(t)] = \alpha(1 - e^{-\beta t})^\gamma + \sum_{j=1}^{k} \beta_j x_{i,j} \) | Both explanatory variables and time dependent variable, Sigmoid shape expression | Need to fix growth function in advance |

In models [1] and [4], we need to assume the form of growth function a priori, in this case, Chapman-Richards growth function. If Chapman-Richards setting is suitable, these models can be
used without a problem. If the wrong growth function is selected, results may be skewed or invalid, underscoring the importance of proper model selection. This implies that it is prudent an optimal growth function is selected. One approach is to use information criterion for variable selection in the regression models (Kamo and Yoshimoto, 2013ab).

Based on the aforementioned background information, we propose a new approach. We try to estimate only coefficients of interesting explanatory variables by regarding the longitudinal growth as nuisance baseline. The idea for nuisance baseline appears in several regression models. One example is Cox proportional hazard model used in survival analysis (Cox, 1984; Kamo, 2012). This model handles hazard ratio, and our main interest is in estimating the coefficients for explanatory variables. The coefficients can be estimated without using the form of baseline hazard function. Other examples can be found in the field of epidemiological, i.e., age-period-cohort (APC) model (see Holford,1992; Tonda et al., 2015). As Keyes et al., (2010) pointed out, since APC model has an identification problem, the cohort effect is often regarded as nuisance baseline. In this paper, we apply these concepts of nuisance baseline to the growth analysis.

2. Materials and Methods

2.1. Methods

Let $y_i(t)$ be the target growth amount of $i$-th sample at age $t$ ($i = 1, 2, \ldots, n$). By assuming that $y_i(t)$ is affected by both age and external factors, then the model is constructed as:

$$E[y_i(t)] = \beta_0(t) + \sum_{l=1}^{k} \beta_l x_{i,l}$$

where $x_{i,l}$ is the $l$-th explanatory variable in $i$-th sample. Although model [5] can be seen as general form of [4], the difference between [4] and [5] is that $\beta_0(t)$ in [5] is set as nuisance baseline (i.e. no concrete form is set). Let the observed age be $t_j$ ($j = 1, \ldots, m$), which means [5] can be rewritten by using matrix notation as:

$$y = X\beta + \epsilon = (U V) \begin{pmatrix} \lambda \\ \mu \end{pmatrix} + \epsilon$$

where $y=(y_1(t_1), y_2(t_1), \ldots, y_i(t_j), \ldots, y_n(t_m))'$ and $X=\{x_{i,l}\}$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_{mn})'$. Here $X$ and $\beta$ are partitioned into two parts as $X = (U V)$ and $\beta=(\lambda' \mu')'$. The matrix $U$ and parameter $\lambda$ corresponds to the nuisance part. The $nm \times m$ matrix $U$ is defined as:

$$U = \begin{pmatrix} 1_n & 0 \\ \vdots \\ 0 & 1_n \end{pmatrix}$$

where $1_n$ is an $n$-dimensional vector of ones. On the other hand, since $V$ and $\mu$ correspond to the main variable part, then our target is the estimation for $\mu$. The error $\epsilon$ can be assumed to be distributed according to the normal distribution, i.e., $\epsilon \sim N(0, \sigma^2 I)$, so that the maximum likelihood estimator (MLE) of $\beta$ is estimated as:

$$\hat{\beta} = \begin{pmatrix} \hat{\lambda} \\ \hat{\mu} \end{pmatrix} \left( V'U V'V^{-1} + U'y \right)^{-1} V'y$$

The $\hat{\mu}$-related part can be extracted from [8] as:

$$\hat{\mu} = \left(V'(I - P_U)V\right)^{-1} V'(I - P_U)y$$

where $P_U = U(U'U)^{-1}U'$ is the projection matrix. The key point in [9] is that $\hat{\mu}$ does not include the information for nuisance parameter $\lambda$. The covariance of $\hat{\mu}$ is also derived as:

$$\text{cov} \{\hat{\mu}\} = \sigma^2 \left( V'(I - P_U)V \right)^{-1}$$
The derivation for [9] and [10] are shown in Appendix 1.

The remaining parameters $\lambda$ and $\sigma^2$ are posteriorly estimated as:

\[ \hat{\lambda} = (U'U)^{-1}U'(y - X\hat{\mu}) \]

and

\[ \hat{\sigma}^2 = \frac{1}{mn - k - m}(y - X\hat{\mu})'(I - P_U)(y - X\hat{\mu}) \]

by using $\hat{\mu}$. All of the MLEs [9], [11] and [12] are unbiased estimators. The proof for the unbiasedness is shown in Appendix 2.

2.2. Materials

The aforementioned methods were applied to the growth data for Sugi (Cryptomeria Japonica). The data was collected during the period of 2002-2003 at Hoshino village, in Japan (Yoshimoto et al., 2012). Figure 1 shows the location of Hoshino village. For growth projections, three variables were considered: diameter at breast height (DBH), tree height and tree volume. Figure 2 shows the longitudinal behaviour of these projections. On the other hand, eight kinds of external variables were measured: three spatial variables (latitude, longitude, altitude) and five variables for surrounding (DBH when data was collected, DBH of the most neighborhood, density, occupancy rate for DBH (Maeda et al., 2004), distance weighted DBH (Yoshimoto et al., 2005)). In this analysis, DBH was set as the response variable and altitude as the explanatory variable (e.g. Ishida, 2003). For the response variable, the reason for choosing DBH was due to the complicated growth behaviour (far from linear). Since the growth behaviour of DBH was considered hard to reproduce through mathematical modeling, it was considered suitable for checking the performance of the model. On the other hand, the reason for setting altitude as the explanatory variable is due to the fact that latitude and longitude effects are not expected to be monotone and linear, numerically. Moreover, all of the surrounding variables are not suitable for explanatory variables, because they are expected to vary during the tree growth interval (about 20 years). For example, density may vary as a result of thinning or mortality (Kraft et al., 2010). In model [5], the value of explanatory variable in each stand is tacitly assumed not to change with age during tree growth interval. As a result, the model in this case study should be set as:

\[ E[DBH_i(t_j)] = \beta_0(t_j) + \beta_1 \times \text{Altitude}_i \]

where $DBH_i(t_j)$ is the DBH of the $i$-th sample at age $t_j$; $\text{Altitude}_i$ is the altitude of the $i$-th sample; $\beta_0(t)$ corresponds to longitudinal growth as a function of age $t$; and $\beta_1$ is the parameter of interest.

Figure 1. Location of Hoshino village in Japan.
Figure 2. Observed growth amount. (a): DBH, (b): Height, (c): Volume.
3. Results

In model [13], the MLE of $\beta_1$ is estimated as $\hat{\beta}_1 = -0.63$ without estimating $\beta_0(t)$. Hence, model [13] is rewritten as:

$$E[DBH_i(t)] = \beta_0(t) - 0.63 \times \text{Altitude}_i$$

The standard error of $\hat{\beta}_1$ is 0.052. By using $\hat{\beta}_1$, the nuisance baseline $\beta_0(t)$ and the residual variance $\sigma^2$ can be estimated posteriorly. The MLE of variance is, $\hat{\sigma}^2 = 1.73$ and the estimated nuisance baseline for longitudinal growth is shown in Figure 3. Figure 4 shows the deviations of the DBH of the sample trees from the baseline. It can be seen from this figure that once the DBH of the sample tree, exceeds the baseline, it does not get below the baseline. Figure 5 shows the plot of the differences in DBH from baseline, against the altitude with the estimated regression line. From this figure, it can be deduced that, an increase in altitude, leads to less growth.

![Figure 3. Estimated baseline growth. (The dotted line show each sample growth and the continuous line shows the estimated baseline growth.)](image)

4. Discussion

In the growth modeling with explanatory variables, the handling for the longitudinal growth part is difficult and sensitive problem, because wrong setting leads the results incorrect. To solve such problem, we apply the concept for nuisance baseline to the growth analysis. Owing to the concept of nuisance baseline, we can estimate main parameter without any assumption for longitudinal behavior. The resultant model, [14], means that higher altitude leads to less growth. There are two possible reasons for this result. The first possible explanation is the flow of water and nutrients. Both water and nutrients flow from a higher elevation and accumulate at a lower elevation. This means that the lower areas are rich in nutrients and water. Sands and Mulligan (1990) pointed out that such conditions are ideal for tree growth and therefore accelerates tree growth. The second possible explanation is that, in higher elevations, trees are suppressed by strong winds. King (1986) revealed that such condition impedes the tree growth. These two reasons explain why $\beta_1$ is negative.

In order to check the performance of the method, we compare our model with other alternative models such as [2], [3] and [4] using the same data. For the coefficient for altitude, estimated standard error of parameter and residual variance are summarized in Table 2. Similar estimates
Figure 4. The difference of DBH from baseline DBH for each tree.

Figure 5. Plot of the difference of DBH from baseline DBH against altitude, and estimated regression line.
of these parameters were obtained using models [3], [4] and [5]. Model performance can be judged based on the values of the standard error and residual variance. A model is considered better, if it has a lower standard error and lower residual variance. Among models [2], [3], [4] and [5], both indexes for model [5] are the smallest, and therefore considered the most suitable model among the four models.

Table 2. Comparison of the regression coefficients for altitude in different models [2]-[5].

| Longitudinal change Effect of altitude | Estimate | Standard Error | Residual variance |
|--------------------------------------|----------|----------------|-------------------|
| Model for $\beta_0(t)$               |          |                |                   |
| [2] Constant                         | -0.548   | 0.211          | 28.584            |
| [3] Linear                           | -0.626   | 0.068          | 2.948             |
| [4] Chapman-Richards                 | -0.628   | 0.053          | 1.831             |
| [5] Proposed model                   | -0.630   | 0.052          | 1.730             |

Although the proposed method possesses several advantages as mentioned above, there are a few problems when it comes to its practical use. These are problems that need further discussions. To begin with, for all practical purposes when it comes to growth analysis, the performance of a model is judged by its predictive power. For example, forest managers rely on the accuracy of growth models to determine how much and when to harvest, in their forest management plans. However, because the growth part of the model is handled as nuisance, it is difficult to construct the model for future extrapolation. For this reason, not just longitudinal baseline, but also spatial baseline, cannot be used to extrapolate for the prediction.

Next, we discuss a specific problem in this case study. For the explanatory variables, not just altitude but information on several variables were collected in the field. However, except for altitude, none of those variables was suitable for use as explanatory variables. It is expected that the external data should be longitudinal, from the data collection viewpoint, or varying coefficient model should be adopted from the theoretical viewpoint.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N. and Csaki, F. (eds), 2nd International Symposium on Information Theory. Akadimiai Kiado, Budapest, pp. 267–281.

Cameron, A.C. and Trivedi, P.K. (2013) Regression analysis of count data (Econometric society monographs), Cambridge University Press, Cambridge (U.K.), New York.

Cox, D.R. and Oakes, D. (1984) Analysis of Survival Data, Chapman & Hall, New York.

Holford, T.R. (1992) Analysing the temporal effects of age, period and cohort. Stat. Methods. Med. Res., 1: 317–337.

Imada, M., Kunisaki, T., Mizoue, N., and Teraoka Y. (1997) Optimum planting density for Japanese Oak (Quercus mongolica var. grosseserrata) based on spacing experiment with systematic design, J. For. Res. 2: 89–93.

Ishida, M. (2003) Distribution and Dynamics of Beech (Fagus crenata) Forests in Toyama Prefecture, Proceedings of the Institute of Statistical Mathematics, 51(1): 59–72 (in Japanese).

Kamo, K. (2012) Survival analysis and tree failure: Results from a tree-pulling experiment, FORMATH 11: 195–209 (in Japanese with English abstract).

Kamo, K. and Yoshimoto, A. (2013a) Comparative analysis of growth functions based on Mallows’ Cp type criterion, FORMATH 12: 133–147.

Kamo, K. and Yoshimoto, A. (2013b) Comparative analysis on selecting growth function, J. For. Sci. Tech. 9(2): 65–71.
Kamo, K., Tonda, T., Satoh, K.

Keyes, K.M., Utz, R.L., Robinson, W., and Li, G. (2010) What is a cohort effect? Comparison of these statistical methods for modeling cohort effects in obesity prevalence in the United States, 1971-2006, Soc. Sci. Med., 1100-1108.

King, D.A. (1986) Tree form, height growth, and susceptibility to wind damage in acer saccharum, Ecology 67(4): 980-990.

Kleinbaum, D.G. and Klein, M. (2010) Logistic regression: A self-learning text (Statistics for biology and health), Springer-Verlag, New York.

Kraft, N.J., Metz, M.R., Condit, R.S., and Chave, J. (2010) The relationship between wood density and mortality in a global tropical forest data set, New Phytol. 188(4): 1124-1136.

Maeda, Y., Yoshida, S., Nagashima, K., and Murakami, T. (2004) Evaluation of forest type classification method and species diversity in plantations by using the national forest inventory data, Kyushu J. For. Res. 57: 203-206 (in Japanese).

Nelder, J.A. (1962) New Kinds of systematic designs for spacing experiments, Biometrics, 18: 283-307.

Richards, F.J. (1958) A flexible growth function to empirical use, J. Exp. Bot. 10: 290-300.

Ryan, M.G. and Yoder, B.J. (1997) Hydraulic limits to tree height and tree growth, BioScience, 47: 235-242.

Sands, R. and Mulligan, R.S. (1990) Water and nutrient dynamics and tree growth, For. Ecol. Manage. 30: 91-111.

Schott, J.R. (2005) Matrix Analysis for Statistics, 2nd Edition, John Wiley & Sons, Inc., New York.

Tonda, T., Satoh, K., and Kamo, K. (2015) Detecting a local cohort effect for cancer mortality data using a varying coefficient model, J. Epidemiol. 25: 639-646.

Yin, C., Duan, B., Wang, X., and Li, C. (2004) Morphological and Physiological responses of two contrasting Popular species to drought stress and exogenous abscisic acid application, Plant. Sci. 167: 1091-1097.

Yoshimoto, A., Yanagihara, H., and Ninomiya, Y. (2005) Finding Factors Affecting a Forest Stand Growth through Multivariate Linear Modeling, J. Jpn. For. Soc. 87: 504-512 (in Japanese).

Yoshimoto, A., Kamo, K., and Yamagihara, H. (2012) Environmental data analysis by R, Asakura Syoten (in Japanese).

Zeide, B. (1993) Analysis of growth equation, For. Sci. 39: 594-616.

Appendix 1 (The derivation for $\hat{\mu}$ and its covariance)

Suppose that the matrix $A$ is blocked as $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{pmatrix}$. If $A_{11}$ is regular, then $A^{-1}$ is expressed as

$$A^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{12}^{-1}A_{22}^{-1}A_{12} & A_{12}^{-1}A_{22}^{-1} \\ -A_{22}^{-1}A_{12}^{-1}A_{11} & A_{22}^{-1} \end{pmatrix}$$

where $A_{22}^{-1} = A_{22} - A_{12}A_{11}^{-1}A_{12}$ (e.g. Schott (2005)). By applying the formula [15] to [8], $\hat{\mu}$ is obtained as

$$\hat{\mu} = -\{V'(I - PU)V\}^{-1}V'U(U'U)^{-1}U'y + \{V'(I - PU)V\}^{-1}V'y$$

$$= -\{V'(I - PU)V\}^{-1}V'PUy + \{V'(I - PU)V\}^{-1}V'y$$

$$= \{V'(I - PU)V\}^{-1}V'(I - PU)y$$
The covariance of $\hat{\mu}$ is also obtained as

$$\text{cov} [\hat{\mu}] = \{\mathbf{V}'(\mathbf{I} - \mathbf{P}_U)\mathbf{V}\}^{-1}\mathbf{V}'(\mathbf{I} - \mathbf{P}_U)\text{cov}[\mathbf{y}]\mathbf{V}(\mathbf{I} - \mathbf{P}_U)\{\mathbf{V}'(\mathbf{I} - \mathbf{P}_U)\mathbf{V}\}^{-1}$$

$$= \sigma^2\{\mathbf{V}'(\mathbf{I} - \mathbf{P}_U)\mathbf{V}\}^{-1}$$

**Appendix 2 (Unbiasedness of MLEs)**

All of the MLEs $\hat{\mu}$, $\hat{\lambda}$ and $\hat{\sigma}^2$ obtained by [9], [11] and [12] are unbiased. This is checked by calculating the expected value of each MLE. At first, the expected value for $\hat{\mu}$ is obtained as

$$E[\hat{\mu}] = \{\mathbf{X}'(\mathbf{I} - \mathbf{P}_U)\mathbf{X}\}^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{P}_U)E[\mathbf{y}]$$

$$= \{\mathbf{X}'(\mathbf{I} - \mathbf{P}_U)\mathbf{X}\}^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{P}_U)(\mathbf{U}\lambda + \mathbf{X}\mu)$$

$$= \mu$$

Hence $\hat{\mu}$ is the unbiased estimator. Applying the similar way to $\hat{\lambda}$ and $\hat{\sigma}^2$, we obtain the unbiasedness of them as

$$E[\hat{\lambda}] = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'(E[\mathbf{y}] - \mathbf{X}\mu)$$

$$= (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'(\mathbf{U}\lambda + \mathbf{X}\mu - \mathbf{X}\mu)$$

$$= \lambda$$

and

$$E[\hat{\epsilon}'] = \sigma^2\text{tr}((\mathbf{I} - \mathbf{P}_U - \mathbf{P}(\mathbf{I} - \mathbf{P}_U)\mathbf{X})$$

$$= \sigma^2(mn - k - m)$$

respectively. Thus, $E[\hat{\sigma}^2] = \sigma^2$. 

21