Predicting Timber Volume Using Measurement Error Models and Survey Sample Theory

A Thesis presented in partial fulfillment of the requirements for the degree of

Master of Applied Statistics

At Massey University, Palmerston North, New Zealand

Stephen Hitchcock

2001

Acknowledgements

First and foremost I would like that say thank you to my supervisor, Associate Professor Stephen Haslett, who has given me the guidance and continued devotion to help complete this thesis. His knowledge and commitment to statistics amazed me, and I feel very lucky to have had one of the leading people in this area as a supervisor. Due to his stature in the statistics world he was obviously very busy at times, however I appreciated his willingness to help me whenever he could, sometimes dropping other work to fit me in.

I would also like to acknowledge the team at Forest Research who I have had the pleasure of working with for the last few years; Barbara Hock, Andy Gordon, Alan Thorn, John Firth, Rod Brownlie and Chris Goulding. Not only for their continued help, friendship and guidance but for also supplying data, knowledge and giving me the opportunity to apply my statistics skills to a real life situation that is very quickly becoming one of the major industries of New Zealand, forestry.

Thanks to mum and dad obviously, as without their continued support none of this would have happened. There were many times throughout this degree where I lost the focus, but mum and dad (and my sister, Sonya) were always there to help get me back on the right track.

And to the other Statistics staff and postgraduate students at Waikato and Massey Universities, that I spent my time with over the last two and a half years. Without your continued friendship (and the occasional extra long coffee break) my time spent at university would have been much harder. Thank you to you all.

Table of Contents

List of Figures

1.0 Introduction

The possible yield and log type distribution of a stand of trees is a valuable source of information to forest owners. This information allows forest owners to plan logging procedures and forecast potential yields in order to fill national and international orders for now and in the future. Obtaining information of this nature requires an accurate and flexible process that must allow for the changing trends in today's world market.

Flexibility of the process is also required at the inventory[1] level. Ninety five percent of New Zealand's plantation forests consist of the *Pinus Radiata* species. Due to *Pinus Radiata's* competitive tendencies (at the individual tree level) it not only exhibits complex variations over the forest but also between individual trees. Natural tree characteristics, such as malformation, forking and age can also contribute to these variations.

Currently in New Zealand, a forest inventory process known as MARVL (Method for Assessment of Recoverable Volume by Log types) is used to obtain such information. MARVL was developed in the mid 1970's by the New Zealand Forest Research in response to this need for increased detail of timber measures.

We begin by firstly giving the reader a brief introduction to MARVL in order to introduce the different chapters in this report.

1.1 Introduction to MARVL

As stated above, MARVL was developed by Forest Research in the mid 1970's by Deadman and Goulding et al. MARVL is known in by the forestry community as a forest inventory. A forest inventory is a method of obtaining information about forest characteristics (such as timber volume and quality) before the trees are harvested. A forest inventory usually involves a number of different stages. Deadman et al [9] describes MARVL as a three step process,

1. Sampling
2. Cruising
3. Analysis

Statistical sample designs are used to ensure that a representative sample of a stand is obtained. Typically, sampling involves spatially selecting a number of fixed-area bounded plots either randomly or systematically throughout the stand. MARVL supports four specific design options; Simple Cluster Sampling, Stratified Cluster Sampling, Two-phase Sampling with ratio estimation and Stratified two-phase sampling with ratio estimation.

Cruising is a forestry term and refers to visiting the sampling plots in the stand (which is considered the area of interest) and assessing every tree in the sample plot for quality. Other variables such as diameter at breast height[2] (DBH) and tree height are also measured. To assess the quality of a particular tree, field crews are given alphabetic

[1] Forest inventory is defined as the process of obtaining measures on standing timber (such as volumes)
[2] Diameter at breast height (DBH) refers to the diameter of a tree, 1.4 meters from the ground.

codes defined by the user in which to classify the tree (this is called a dictionary). An example of a MARVL dictionary is presented below in table 1.1.

Code	Description
A	Pruned, straight, peeler quality
B	Pruned, straight, NOT peeler quality
C	Pruned, moderate sweep to 25% of diameter, peeler quality
D	Pruned, moderate sweep to 25% of diameter, NOT peeler quality
N	"Default", unpruned sawlog, knots <12cm
P	Pulp
W	Waste

Table 1.1: Example of a MARVL dictionary

To demonstrate how a tree is recorded by the field crew for quality, consider the following example. Suppose the n^{th} tree in a particular plot has a DBH of 401mm, quality A from the bottom 6m. Then it forks into two leaders, both of quality B. One diameter of the fork is 150mm and the other 140mm, measured at 7.4m above the ground. Both leaders are assumed to reach the same height above the ground as if the tree had not forked. Using the alphabetic codes presented in table 1, the field crew would recorded this tree as

n 401 A6 <150B
<140B

These data, in combination with the other recorded variables (DBH and height) can then used to estimate the potential yield (volumes) for the different log types.

The analysis step of the MARVL process begins by defining a "cutting strategy". A cutting strategy is defined by the Forest Research Institute Ltd [6] as a list of log types, each of which is specified by required lengths, minimum and maximum small-end and large-end diameters, and permitted quality codes. Each log type also has a "value" (dollars per unit volume) which is used to indicate log cutting preferences. An example of a cutting strategy is present in table 1.2.

Log Type	Lengths (m)	Min. Small-end Diameter (cm)	Max. Small-end Diameter (cm)	Max. Large-end Diameter (cm)	Value ($/m^3)	Qualities allowed (see table 1)
Pruned Peeler	2.70-5.40	30.0	99.9	99.9	40	A
Pruned Sawlog	2.70-5.40	30.0	99.9	99.9	30	AB
Unpruned Sawlog	3.50-6.00	25.0	99.9	120.0	20	ABCDN
Pulp	3.50-6.00	15.0	99.9	120.0	10	ABCDNP

Table 1.2: Example of a MARVL cutting strategy

Volumes for specified log types for each tree in each plot are then estimated using existing regional volume and taper equations. The analysis step combines the cutting

strategy, DBH, height and estimated volumes to obtain an optimal solution for the population of interest in terms of a monetary value. Estimates for the stand are calculated by scaling the estimates for the plots (that is, estimate for the area of the plots) by the total area of the stand.

This information thus allows forest owners to plan harvesting schedules to fill current orders.

1.2 Structure of the Thesis

This thesis begins by firstly considering the volume equation used to estimate the volume of an individual tree. Currently MARVL uses a non-linear regression equation to obtain this measure. Regression functions are typically used in not only forest surveys but surveys in general. When a parameter of interest (which is usually expensive to measure) is related by one or more variables (which are relatively inexpensive to measure) a regression model can be formulated. Regression functions aim to fit an overall trend to this relationship in such a way as to minimize some error criterion (such as, least squares or maximum likelihood).

In chapter two we investigate the nonlinear regression equation used to estimate the timber volume of an individual tree by considering the predictor variables used in the equation. Regression equations assume that the predictor variables are measured without error. However, in MARVL the two predictor variables used to estimate volume (Diameter at breast height[3] and height of the tree) are never measured free from error. Measurement error model theory allows the predictor variables of the regression equation to contain error.

Chapter 2 begins by deriving the theory of linear measurement error models for a single explanatory variable. Two methods of fitting linear measurement error models with a single explanatory variable are considered namely, least squares and maximum likelihood estimation. This section acts as a natural stepping stone to the theory of linear measurement error models with p explanatory variables, where the theory is also outlined.

The theory of linear measurement error models with p explanatory variables is then applied to the non-linear regression function used in MARVL to estimate the timber volume of an individual tree. Using the logarithmic function allows us to obtain a linear approximation of the nonlinear volume equation, and hence allows us to use the linear measurement error model theory derived in this section. Using the linear measurement model theory, a simulation study is conducted to show the effect varying levels of error in the predictor variables has on the estimate of volume. Section 2.7 summarizes the results of the simulation study.

[3] Diameter of the tree 1.4 meters from the ground

Sections 2.6.1 to 2.6.4 of this chapter contain a discussion on the possible forms and reasons for the error (measurement error) in the predictor variables of the volume equation (namely, DBH and height). The reader is directed to Schreuder [9] chapter 7 for a more comprehensive discussion of these errors.

Chapter 3 considers the statistical sampling theory used by MARVL to obtain estimates of timber volume for a given area of interest. Statistical sampling theory refers to the statistical designs and the appropriate statistical techniques used to analyze them. Sampling theory comprises two general categories: design and estimation. Statistical designs relate to the method of selection for the primary sampling units (individual tree volumes in MARVL's case). Estimation refers to the formulas used to estimate a parameter of interest (for example a mean or a total volume) for a particular design. Care must be taken when selecting the correct formula for a particular design otherwise bias in the estimates can occur.

As was noted above, the variable of interest (namely timber volume of an individual tree) is seldom measured free of error. However to introduce the sampling theory of the designs supported by MARVL to the reader, chapter 3 assumes that the variable of interest is measured free from error, that is, the observed value is assumed to be the actual value.

Due to such reasons as disease, thinning and disasters a sampling frame[4] for an area of interest does not exist. Also, due to the rather strict budget constraints a sampling frame, in the form of a list of trees, is usually much too expensive to obtain. This leads to cluster sampling as the most viable design for MARVL (and also for a number of other large-scale surveys) as cluster sampling is the only design that allows us to conduct a probability sample without a detailed sampling frame.

The theory of simple cluster sampling is discussed first. This acts as the basis for more complex designs such as stratified cluster sampling and two-phase sampling (or double sampling) with ratio estimation, which are also discussed in this chapter. The prime reason for the more complex designs is to reduce the sampling error associated with the estimate of volume and a discussion on the reduction of variation for each complex design is given.

For each of the cluster designs mentioned above a number of important findings are discussed. It is hoped that these findings can be used in future versions of MARVL to obtain estimates of volume with increased efficiency, compared to current estimates.

[4] A list of the population members (individual trees in MARVLs case)

In addition to obtaining an estimate of the parameter of interest (such as a mean or total), surveyors are also interested in obtaining an estimate of the error for this parameter. Statistical error (which is called total error or total survey error in this thesis) consists of two components, namely: sampling error and non-sampling error. Sampling error relates to the error incurred by taking a sample to estimate a variable of interest rather than the entire population. Sampling error is subject to sample to sample variation and is discussed in chapter 3. Non-sampling errors include all other possible errors associated with the sample survey, such as measurement errors (errors incurred when measuring variables) and non-response (missing values). To obtain a measure of the total survey error[5] of an estimate both types of errors need to be used in its calculation.

Chapter 4 combines measurement error models (chapter 2) (a form of non-sampling error) and the sample designs supported by MARVL (chapter 3) (which incur the sampling error), and aims to obtain an estimator of the total error. The technique derived by Sarndal [10] chapter 16, treats the variable of interest as a random variable rather than a fixed variable. This allows standard statistical sampling techniques to be used to evaluate the effect of the non-sampling errors (including measurement errors in MARVL's case).

The chapter begins by deriving the theory outlined in Sarndal [10] chapter 16. Using this theory we are able to obtain a theoretical mean square error. This in turn allows us to demonstrate the effect measurement error has on total survey error. To give the reader an example the theoretical mean square error for a simple cluster design (see section 2.2.1) with measurement error is derived. From the theoretical mean square error we are able to discuss situations where the total error will be significantly different from the sample error, hence the situations where the total error will be underestimated are discussed.

Unfortunately in practice a true estimate of non-sampling error is not attainable, since only a single observation for each variable of interest is available. For this reason we also demonstrate, theoretically, that a negative bias results for measurements of accuracy when just the sampling error is used as an estimate of total sampling error, which again leads to the result discussed above, namely that an underestimate of total sampling error results.

Chapter 4 introduces an important point, that is, total sampling error involves not only sampling error but also non-sampling error. In a large number of surveys today, non-sampling errors are seldom discussed since no complete theory exists on these types of errors. However, due to the increasing importance of survey results to plan such things as, harvest procedures and government policy there is a need for precise estimates to be obtained. As mentioned above, not including non-sampling error into the total error estimate can lead to a bias ie. an underestimate of the total error, which leads to estimates which are more precise than what they really are.

[5] Sarndal et al. [10] defines total survey error as: total survey error = sampling error + non-sampling error

When, literally, millions of dollars are based on these estimates a small difference in the error estimate could lead to a decision that could potentially cost not only the forestry companies but also the country.

Finally the thesis discusses the practical implications of the theory – which errors are most important, which are most easily controlled, and how survey and calibration resources are best allocated to get the best estimates of timber volume for a fixed cost.

2.0 Measurement Error Models

2.1 Introduction to Measurement Error Models

In this chapter we introduce measurement error model theory and apply it to the volume equations used in MARVL. Measurement error models (as discussed in Cheng [7]) are an extension of the usual regression model covered in most undergraduate statistics courses. Regression models attempt to fit linear or non-linear relationships between a predictor variable or variables (which are assumed error free) and a variable of interest. Measurement error models extend this idea to allow predictor variables to contain an error term.

If we consider MARVL for the moment, a volume equation is used to predict volume from such variables as diameter at breast height (DBH)[6] and height of the tree (see section 2.5). Currently, MARVL assumes that these predictor variables are measured free from error, and hence uses non-linear regression theory to obtain an estimate of the tree volume. However, in practice these variables cannot be measured free from error (due to uncontrollable circumstances such as instrumental error[7]), hence measurement error model theory should be used to obtain an estimate of tree volume.

To introduce measurement error models to the reader we firstly derive the theory of measurement error models with a single explanatory variable. Then we generalize to p explanatory variables (more than one explanatory variable) in section 2.2. A brief discussion about non-linear measurement error models is also considered in section 2.3.

Throughout this chapter, two measurement error models are considered; structural measurement error models and functional measurement error models. Cheng [7] defines structural measurement error models as models where the predictor variables are random and functional measurement error models as models where the predictor variables are fixed. Section 2.2.1 gives a more detailed definition of functional and structural measurement error models.

Let us consider the functional measurement error model (as described in section 2.2) for the moment. In MARVL the volume equation (see section 2.5) is calibrated from 233 trees random selected throughout New Zealand forests. Because this volume equation is applied to all forests (based on these sampled plots) the volume equation used in MARVL is actually a structural model rather than a function model, since the predictor variables are random (see the example given in section 2.2). Hence, theory for the functional measurement error model is not derived in this report (however see Cheng [7] for a full derivation of the functional measurement error model theory).

[6] A term given to the diameter of a tree 1.4 meters from the ground

[7] Instrument error is a term given to the error associated with a value obtained from a physical instrument Sarndal [24]

Standard regression theory uses two techniques to fit models, namely least squares and maximum likelihood. In this chapter we begin by considering the least squares approach for a measurement error model with a single explanatory variable. The inclusion of an extra term (the measurement error) leads to a number of problems when using standard least squares and are discussed. Secondly, maximum likelihood estimators are derived for the unknown parameters of the measurement error model with a single explanatory variable. Similar problems as encountered with the least square's approach are discussed.

Deriving the theory of measurement error models with a single explanatory variable acts as a natural 'stepping stone' to measurement error models with vector explanatory variables (that is, more than one explanatory variable) and are also discussed (section 2.3). Similarly, measurement error models with vector explanatory variables acts as the basis to discuss non-linear measurement error models and the theory is also derived (section 2.4).

Using this theory we investigate the volume equation used by MARVL to estimate timber volume of an individual tree. It is shown that using the logarithmic function transforms the non-linear equation into a linear equation. This allows us to construct a simulation study to study the effects on estimated volume for increased levels of measurement error. The results of this study are summarized in section 2.7.

2.2 Linear Measurement Error Model with a Single Explanatory Variable

The classical linear regression model with one independent variable is usually defined as,

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad \ldots\ldots\ldots\ldots\ldots\ldots (2.2.1)$$

Where the e_i independent $N(0, \sigma_{ee})$ random variables

Let us consider the explanatory variable (x_i) in equation 2.2.1 for the moment. In the regression equation above, two possibilities arise. Firstly say for example we wanted to compare the readings made on soil density from fixed locations in three different forests in New Zealand. If inferences to be made are for these locations in these three forests only, then the x_i's in the regression model above are considered fixed. Suppose, on the other hand, there are more than three forests and we are interested in making inferences on all these forests based only on the three sampled, then the x_i's in the regression model need to be considered as random. This can be a very important point, as neglecting this fact can lead to incorrect parameter estimates.

Further reference to this issue for measurement error models will be made later in this section.

Let us now introduce the simple linear measurement error model such as derived by Cheng [7] and Fuller [12]. The classic regression model presented above assumes that the independent variable (x_i) is measured without error or that the model is fitted conditional on the $\{x_i\}$ (that is, the $\{x_i\}$ are assumed to be fixed). This specification of $\{x_i\}$ highlights the significant difference between regression models and measurement error models. Measurement error models instead allow for this independent variable (x_i) to contain error (u_i). That is, instead of one observing x_i directly, one observes the sum,

$$X_i = x_i + u_i \quad \ldots\ldots\ldots\ldots\ldots\ldots (2.2.2)$$

where u_i is the measurement error and is a $(0, \sigma_{uu})$ random variable. Fuller [12] states that the observed X_i variable is sometimes called the *manifest* or the *indicator* variable and the unobservable x_i variable is sometimes called the *latent* variable.

Let us now return to the point made earlier concerning regression equations with fixed or random explanatory variables. Cheng [7] suggests that different names are given to measurement error models with fixed or random explanatory variables. Firstly, measurement error models with fixed x_i are called *functional* models. If the x_i are random the measurement error is called a *structural* model. If, inaddition to the $\{x_i\}$ being random variables they are also be independent and identically distributed and independent of the errors, we have a *structural* model for which,

$$E(x_i) = \mu_x \quad \text{and} \quad Var(x_i) = \sigma_{xx}^{\ 2}$$

Finally, similar to the *structural* model, the *ultra-structural* model again assumes that the $\{x_i\}$ are independent random variables. However they are not identically distributed. Instead they may for example have different means μ_i, and common variance σ_{xx}^2.

Similar to regression theory, there are different methods of analysis for each type of measurement error model.

2.2.1 Least Squares approach

In this section we will illustrate that the inclusion of measurement error results in a number of problems when using the standard regression technique, least squares, to obtain estimates of the unknown measurement error model parameters.

To obtain the least squares estimate of the slope β_1 for a simple linear regression, we must minimize the error sum of squares of the regression equation above (that is, minimize $\sum e_i^2 = \sum (Y_i - \beta_0 - \beta_1 x_i)^2$). Dobson [10] demonstrates that this leads to the normal equations, from which we are able to obtain the least squares estimate of β_1,

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 \hat{\beta}_1 - \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = 0$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

By way of example let us consider the *structural* model as defined in section 2.2. Fuller [12] investigates the effect the measurement error has on the least squares slope coefficient in the simple measurement error model which comes from combining equations (2.2.1) and (2.2.2). Assuming that

$$(x_i, e_i, u_i)' \sim NI\left[(\mu_x, 0, 0)', diag(\sigma_{xx}, \sigma_{ee}, \sigma_{uu})\right],$$

where ~NI is an abbreviation for 'distributed normally and independently". If follows for this structural model that the vector $(Y_i, X_i)'$, where Y_i is defined by (2.2.1) and X_i by (2.2.2), is distributed as a bivariate normal vector with mean vector

$$E\{(Y, X)\} = (\mu_Y, \mu_X) = (\beta_0 + \beta_1 \mu_X, \mu_X)$$

The covariance matrix can be derived from usual expectation algebra, considering each component in turn:

$$Var(Y_i) = \sigma_{YY} = Var(\beta_0 + \beta_1 x_i + e_i)$$

$$\Rightarrow \sigma_{YY} = \beta_1^2 Var(x_i) + Var(e_i)$$

$$\Rightarrow \sigma_{YY} = \beta_1^2 \sigma_{xx} + \sigma_{ee}$$

Similarly,

$$Var(X_i) = \sigma_{XX} = Var(x_i + u_i)$$
$$\Rightarrow \sigma_{XX} = \sigma_{xx} + \sigma_{uu}$$

From the estimate of β_1 we can derive the covariance between Y_i and x_i,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\Rightarrow \hat{\beta}_1 = \sigma_{xx}^{-1}\sigma_{xy}$$
$$\Rightarrow \sigma_{xy} = \beta_1\sigma_{xx}$$

Thus the covariance matrix of the mean vector is,

$$\begin{bmatrix} \sigma_{YY} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{XX} \end{bmatrix} = \begin{bmatrix} \beta_1^2\sigma_{xx} + \sigma_{ee} & \beta_1\sigma_{xx} \\ \beta_1\sigma_{xx} & \sigma_{xx} + \sigma_{uu} \end{bmatrix}$$

We are now at a point where we can derive the expected least square's estimate of the slope coefficient of the structural measurement error model, denoted γ_{1l}. From the normal equations,

$$E(\hat{\gamma}_{1l}) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$
$$\Rightarrow E(\hat{\gamma}_{1l}) = \sigma_{XX}^{-1}\sigma_{XY}$$
$$\Rightarrow E(\hat{\gamma}_{1l}) = (\sigma_{xx} + \sigma_{uu})^{-1}\beta_1\sigma_{xx}$$
$$\Rightarrow E(\hat{\gamma}_{1l}) = \beta_1\kappa_{xx} \quad \text{.....(2.2.3)}$$

Where

$$\kappa_{xx} = (\sigma_{xx} + \sigma_{uu})^{-1}\sigma_{xx} = \sigma_{XX}^{-1}\sigma_{xx}$$

We can see that the expected value of the estimated slope coefficient (using least squares) is biased for the simple linear measurement error model if the predictor variable is measured with error (see equation 2.2.3). The expected value if the predictor variable is measured with error is in fact the product of the true slope β_1 and a ratio, κ_{xx} (which Cheng [7] and Fuller [12] called the *reliability ratio*). κ_{xx} is always less than one (since $\sigma_{uu} > 0$), hence the expected value of the slope coefficient for what should be modelled as a measurement error model, when using simple least squares instead, is biased towards zero.

In addition to the least squares estimate of the slope parameter being biased towards zero, Cheng [7] and Fuller [12] show that the least squares estimator of the slope parameter also gives inconsistent estimates. Casella [6] defines consistent estimates for an infinite population size as follows,

> 'Denote by $\hat{\theta}_n$ an estimator of a parameter, θ, based on a sample of size n, that is $\hat{\theta}_n = \hat{\theta}_n(X_1,....,X_n)$. As the sample size increases, the sequence of estimators, $\{\hat{\theta}_n\}$, is said to be consistent if for any constant, $\varepsilon > 0$
>
> $$P_\theta\left(\left|\hat{\theta}_n - \theta\right| < \varepsilon\right) \to 1 \quad \text{as } n \to \infty \text{ '}$$

Consistency is a large-sample property. This means that it considers the asymptotic properties of an estimator as the sample size increases. It is useful to consider consistency when an estimator has undesirable properties for small n, but may still be a reasonable estimator in certain applications if it has good asymptotic properties as the sample size increases.

Informally this definition means that as the sample size tends to infinity, the probability that the estimator gets closer and closer to the true value of the parameter tends toward one. We can see from the definition above that the least squares estimate of the slope parameter in the linear measurement error model is not consistent and hence does not yield reasonable estimates as the sample size increases (since $E(\hat{\gamma}_{1l}) = \beta_1 \kappa_{xx}$).

Another problem that is encounter when using the simple least squares approach for obtaining estimates of parameters from a measurement error model is identification. Casella [6] defines identifiability as,

> 'A parameter θ for a family of distributions $\{f(x \mid \theta) : \theta \in \Theta\}$ is *identifiable* if distinct values of θ correspond to distinct probability density functions. That is, if $\theta \neq \theta'$, then $f(x \mid \theta)$ is not the same function of x as $f(x \mid \theta')$.'

Informally, this means that, for a parameter to be *identifiable*, different sets of parameters don't lead to the same joint distribution of x and y (that is, there must be a unique parameter vector that leads to the joint distribution of x and y). It should be noted that *identifiability* of a parameter doesn't guarantee the existence of a *consistent* estimator of the parameter. However identifiabity is sometimes redefined so that the parameter is *identifiable* if it is *consistent*. Identifiability often arises in the ANOVA, when we have more unknowns than equations. Dobson [10] suggests that when this situation arises, extra constraints such as sum-to-zero or corner-point constraints are commonly used to ensure identifiability.

If we compare the linear regression model, equation (2.2.1) (which is identifiable) and the normal measurement error model, which comes from combining equations (2.2.1) and (2.2.2) (which isn't identifiable), we can see that the measurement error model has an additional parameter (namely, σ_{uu}). This additional 'degree of freedom' allows more than one model to yield the same joint distribution of x and y. It should be noted that this problem is similar to the ANOVA problem stated earlier, and as we will see extra constraints are required to yield an identifiable estimator.

There are a number of techniques to deal with the problems outlined above when using standard least squares theory to obtain parameter estimates of the linear measurement error model. Cheng [7] presents six side assumptions, which can make the linear measurement error model identifiable:

1. The ratio of the error variances, $\lambda = \sigma_{ee} / \sigma_{uu}$, is known
2. The reliability ratio, κ_{xx}, is known
3. σ_{uu} is known
4. σ_{ee} is known
5. Both the error variances, σ_{ee} and σ_{uu} are known
6. The intercept, β_0, is known and $E(x) \neq 0$.

The second assumption is the most popular, namely the reliability ratio is known. Fuller [12] states that there are a number of situations, particularly in psychology, sociology and survey sampling, where information about the reliability ratio is known. From above we know that,

$$E(\hat{\gamma}_{1l}) = \beta_1 \kappa_{xx}$$

$$\Rightarrow \hat{\beta}_1 = \hat{\gamma}_1 \kappa_{xx}^{-1}$$

Where

$$\hat{\gamma}_1 = \left[\sum_{i=1}^{n} (X_i - \bar{X})^2 \right]^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

This coefficient, $\hat{\beta}_1$ above, is sometimes called the regression coefficient corrected for attenuation. We can see that if the reliability ratio is known we can obtain an unbiased estimate of the slope parameter, β_1.

Fuller [12] derives the conditional variance of this estimator $\{\beta_1\}$ for n>3 given X,

$$V\{\hat{\beta}_1 \mid \mathbf{X}\} = \kappa_{xx}^{-2} \left[\sum_{i=1}^{n} (X_i - \bar{X})^2 \right]^{-1} s_l^2$$

Where

$$s_l^2 = (n-2)^{-1} \sum_{i=1}^{n} [Y_i - \bar{Y} - \hat{\gamma}_{1l}(X_i - \bar{X})]^2$$

It should further be noted that under the assumption that the ratio of the errors, λ, is known and the data is scaled so that $\lambda = 1$, the maximum likelihood solution of the normal measurement error regression is orthogonal regression[8]. Even in the unscaled case, the maximum likelihood solution is orthogonal regression under an appropriate metric . Cheng [7] pages 9 to 11 derives the theory. Comparison of the three regression distances are illustrated in figure 2.2.1 below,

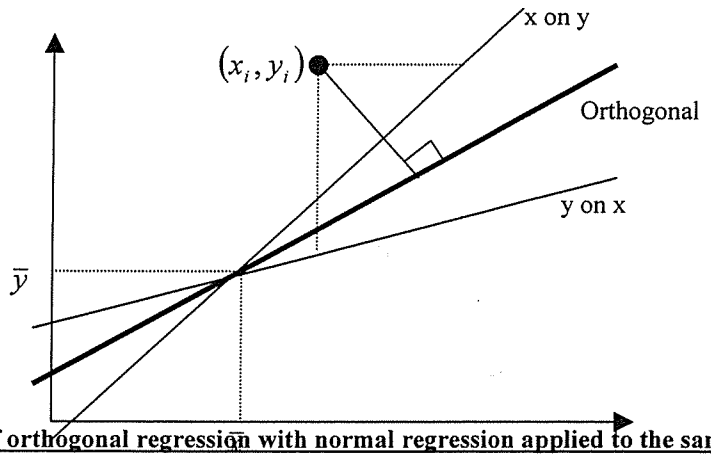

Figure 2.2.1: Comparison of orthogonal regression with normal regression applied to the same dataset

2.2.2 Maximum Likelihood Approach

A second approach to finding the parameters of the linear measurement error model is the method of Maximum Likelihood. The method of Maximum likelihood involves obtaining the first derivatives of the likelihood function[9] and setting this equal to zero and solving for the desired parameter. Alternatively, one could make use of the functional invariance property of the maximum likelihood estimates to derive estimates from means, variances and covariances. Bain [2] defines the invariance property of the maximum likelihood estimate as,

> 'If $\hat{\theta}$ is the maximum likelihood estimate of θ and if $u(\theta)$ is a function of θ, then $u(\hat{\theta})$ is the maximum likelihood estimate of $u(\theta)$.'

Using the invariance property we can make use of the usual, well known maximum likelihood estimates for the parameters of a bivariate normal distribution given by Cheng [7], these are,

---

[8] Orthogonal regression minimizes the sum of squares of the orthogonal distances from the data point to the regression line.

[9] The likelihood function is defined by Bain [2] as the joint density function of n random variables $Z_1,..., Z_n$ evaluated at $z_1,..., z_n$, say $f(z_1,..., z_n; \theta)$, is referred to as the likelihood function.

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum x_i$$

$$\hat{\mu}_y = \bar{y} = \frac{1}{n}\sum y_i$$

$$\hat{\sigma}_{XX} = s_{XX} = \frac{1}{n}\sum \left(X_i - \bar{X}\right)^2$$

$$\hat{\sigma}_{YY} = s_{YY} = \frac{1}{n}\sum \left(Y_i - \bar{Y}\right)^2$$

$$\hat{\sigma}_{YX} = s_{YX} = \frac{1}{n}\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

Using these maximum likelihood estimates we wish to solve the following set of five equations for $\hat{\mu}, \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_{xx}, \hat{\sigma}_{ee}$, and $\hat{\sigma}_{uu}$ to obtain valid estimates[10],

1. $\bar{X} = \hat{\mu}$
2. $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1\hat{\mu}$
3. $s_{XX} = \hat{\sigma}_{xx} + \hat{\sigma}_{uu}$
4. $s_{YY} = \hat{\beta}_1^2\hat{\sigma}_{xx} + \hat{\sigma}_{ee}$
5. $s_{XY} = \hat{\beta}_1\hat{\sigma}_{xx}$

It should be noted that these estimates are exactly the same as the least squares estimates obtained in section 2.2.1

To obtain 'valid estimates', for the measurement error model, that is estimates with nonnegative variances, Cheng [7] (equation 1.27, page 15) derives the following restrictions from the likelihood estimates given above,

1. $s_{XX} \geq s_{XY} / \hat{\beta}_1$
2. $s_{YY} \geq \hat{\beta}_1 s_{XY}$
3. $s_{XX} \geq \hat{\sigma}_{uu}$
4. $s_{YY} \geq \hat{\sigma}_{ee}$
5. $sign(s_{XY}) = sign(\hat{\beta}_1)$

Unfortunately we have six unknowns, $\hat{\mu}, \hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_{xx}, \hat{\sigma}_{ee}$, and $\hat{\sigma}_{uu}$, but only five equations. Similar to the least squares approach, we have a model unidentifiability problem again. However, by using one of the six side assumptions mentioned in section 2.1.2 we are able to obtain unique maximum likelihood estimates for each of these parameters.

[10] By 'Valid Estimates' one means that the solutions for the variances, $\hat{\sigma}_{xx}, \hat{\sigma}_{uu}, \hat{\sigma}_{ee}$ must be nonnegative.

Let us now derive estimates for all the unknown parameters, namely $\beta_0, \beta_1, \sigma_{xx}, \sigma_{uu}$ and σ_{ee}, for the structural model using the method of maximum likelihood, assuming the reliability ratio is known. Since the reliability ratio is assumed known, we know that,

$$\kappa_{xx} = \frac{\sigma_{xx}}{\sigma_{XX}}$$
$$= \frac{\sigma_{xx}}{\sigma_{xx} + \sigma_{uu}}$$

Rearranging this equation we obtain,

$$\kappa_{xx} = (\sigma_{xx} + \sigma_{uu})^{-1} \sigma_{xx}$$
$$\Rightarrow \sigma_{xx} = \kappa_{xx}(\sigma_{xx} + \sigma_{uu})$$

From solution three obtained from the invariance property of the maximum likelihood estimate, presented above, we also know that

$$s_{XX} = \hat{\sigma}_{xx} + \hat{\sigma}_{uu}$$

Substituting this equation into the equation above it is possible to obtain an estimate for σ_{xx},

$$\sigma_{xx} = \kappa_{xx}(\sigma_{xx} + \sigma_{uu})$$
$$\Rightarrow \hat{\sigma}_{xx} = \kappa_{xx} s_{XX}$$

Similarly, substituting this estimate into the following equation allows us to obtain an estimate for σ_{uu},

$$s_{XX} = \hat{\sigma}_{xx} + \hat{\sigma}_{uu}$$
$$\Rightarrow \hat{\sigma}_{uu} = s_{XX} - \hat{\sigma}_{xx}$$
$$\Rightarrow \hat{\sigma}_{uu} = s_{XX} - \kappa_{xx} s_{XX}$$
$$\Rightarrow \hat{\sigma}_{uu} = s_{XX}(1 - \kappa_{xx})$$

From solution five obtained from the invariance property of the maximum likelihood estimate and the derived estimates above it is possible to obtain the maximum likelihood estimate of the slope parameter, β_1,

$$s_{XY} = \hat{\beta}_1 \hat{\sigma}_{xx}$$
$$\Rightarrow \hat{\beta}_1 = s_{XY} \hat{\sigma}_{xx}^{-1}$$
$$\Rightarrow \hat{\beta}_1 = s_{XY} \left(\kappa_{xx} s_{XX}\right)^{-1} \qquad (\text{since } \hat{\sigma}_{xx} = \kappa_{xx} s_{XX})$$
$$\Rightarrow \hat{\beta}_1 = \kappa_{xx}^{-1} \frac{s_{XY}}{s_{XX}}$$
$$\Rightarrow \hat{\beta}_1 = \kappa_{xx}^{-1} \hat{\gamma}_{1l} \qquad (\text{since } \hat{\gamma}_{1l} = \frac{s_{XY}}{s_{XX}})$$

It should be noted that the maximum likelihood estimator of the slope coefficient is equivalent to the least squares estimator under the measurement error model when the reliability ratio is known.

From solution four obtained from the invariance property of the maximum likelihood estimate we can obtain an estimate for σ_{ee},

$$s_{YY} = \beta_1^2 \hat{\sigma}_{xx} + \hat{\sigma}_{ee}$$
$$\Rightarrow \hat{\sigma}_{ee} = s_{YY} - \beta_1^2 \hat{\sigma}_{xx}$$
$$\Rightarrow \hat{\sigma}_{ee} = s_{YY} - \beta_1 \left(\beta_1 \hat{\sigma}_{xx}\right)$$
$$\Rightarrow \hat{\sigma}_{ee} = s_{YY} - \beta_1 s_{XY}$$
$$\Rightarrow \hat{\sigma}_{ee} = s_{YY} - \frac{s_{XY}^2}{\kappa_{xx} s_{XX}}$$

Finally, β_0 is estimated using the following equation,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Note that similar derivation of parameters for the MLE can be achieved when another one of the six side assumptions are assumed.

2.2.3 The Linear Measurement Error Model with correlated errors

Previously we have assumed that the measurement errors σ_{uu} and σ_{ee} have been independent of each other (that is, $\sigma_{ue} = 0$). In this section we explore the effect of these two errors being correlated (that is, $\sigma_{ue} \neq 0$) in the context of MARVL.

MARVL fits a non-linear equation to tree measurements to predict the volume of timber. As will be shown later, this is essentially a linear model for the logarithm of volume, with log diameter at breast height and other logarithm based variables as predictors.

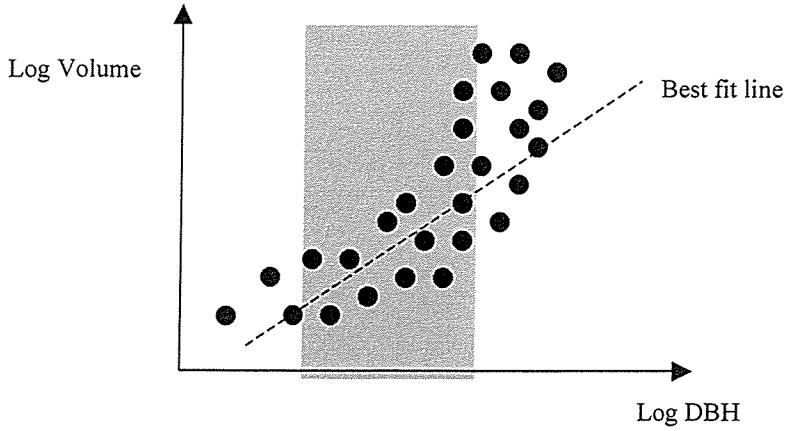

Figure 2.2.2: Demonstration of correlated measurement and equation error

Say we only sample trees with a diameter at breast height shown in the shaded area. Hence we fit a line of best fit to this data. However, say that the actual relationship (for all log diameter at breast height) was as shown in figure 2.2.2. In this case the measurement errors (σ_{uu} and σ_{ee}) would no longer be independent. For small and large values of diameter at breast height the line of best fit will consistently overestimate or underestimate the actual volume (hence $\sigma_{ue} \neq 0$).

However, in MARVL as we will later show this situation seldom arises, that is as the logarithm of diameter at breast height increases so to will the logarithm of volume (linearly). Hence in MARVL the equation used to model the relationship between the predictor and independent variables holds, and hence we may assume that the measurement errors (σ_{uu} and σ_{ee}) are in fact independent.

2.2.4 Summary of the Linear Measurement Error Model with a Single Explanatory Variable

The linear measurement error model with a single explanatory variable (equations 2.2.1 and 2.2.2) differs from the simple linear regression model since the explanatory variable (equation 2.2.1), as well as the response variable, contains an error term (called measurement error).

The extra parameter (the measurement error) causes the linear measurement error model to become unidentifiable. This leads to a number of problems when obtaining estimates for the parameters of the measurement error model. In addition to this finding the estimates of the parameters of interest also become 'inconsistent', that is as the sample size tends to infinity the probability that the estimator gets closer and closer to the true value of the estimator is not equal to one.

Due to the consistency and identifiability problems faced with measurement error models, Fuller [12] and Cheng [7] suggest that one of the six side assumptions outlined in section 2.2.1 must be used in addition to the usual parameter estimates to obtain consistent and identifiable parameter estimates.

It should also be noted from sections 2.2.1 and 2.2.2, that the least squares and the maximum likelihood estimates of the measurement error models yield the same parameter estimates. This observation is much like standard regression theory (Dobson [10]). The measurement error estimate of slope is not however equal to the standard regression theory estimate.

2.3 Linear Measurement Error Models with Vector Explanatory Variables

Fuller [12] defines the linear measurement error model (for a single dependent variable, y) with vector explanatory variables as follows,

$$y_i = \mathbf{x}_i\boldsymbol{\beta}, \quad \text{.....(2.3.1)}$$

$$(Y_i, \mathbf{X}_i) = (y_i, \mathbf{x}_i) + (e_i, \mathbf{u}_i) \quad \text{.....(2.3.2)}$$

Where

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$

$\boldsymbol{x}_i$ = 1 × p matrix of unobservable explanatory variables

$\boldsymbol{X}_i$ = 1 × p matrix of observable explanatory variables

$\boldsymbol{u}_i$ = p × 1 matrix of measurement errors

$(e_i, \boldsymbol{u}_i') = \boldsymbol{\varepsilon}_i$ (A vector of measurement errors)

y_i and e_i = scalars

It should also be noted that similar to the models described in section 2.1, the measurement errors, $(\boldsymbol{u}_i', e_i) = \boldsymbol{\varepsilon}_i$, are independent and identically distributed random variables, which are independent of the true values $\boldsymbol{x}_i$.

As we can see from above, linear measurement error models with vector explanatory variables (equations 2.3.1 and 2.3,2) are an extension to the linear measurement error model with a single explanatory variable. Most of the theory in this section is an extension of the theory derived in section 2.1.

In addition to this, many of the problems incurred using a linear measurement error model with a single explanatory variable (such as inconsistency and identifiability) occur with linear measurement error models with vector explanatory variables. This section highlights these problems and gives examples as to how to deal with them.

Identifiability of linear measurement error model is one of the first problems we will discuss. Similar to the linear measurement error model with a single explanatory variable, the addition of a measurement error matrix to the usual regression problem, causes the model to become unidentifiable. In the same way as we dealt with the problem of identifiability of the linear measurement error model with a single explanatory variable additional assumptions are required to make the measurement error model with more than one explanatory variable identifiable.

Cheng [7] and Fuller [12] derive the maximum likelihood estimates for the linear measurement error model with vector explanatory variables when the measurement errors $(u_1,...,u_p)$ and the model error (e) are independent of each other. The usual identifiability assumption (that is, the additional assumption required to make the measurement error model identifiable) is that the error covariance matrix is known up to a proportionality factor (that is, $\Sigma_{\varepsilon\varepsilon} = \Psi_{\varepsilon\varepsilon}\sigma^2$, where $\Psi_{\varepsilon\varepsilon}$ is the proportionality factor). It should be noted that this assumption is an extension to the multivariate model of the ratio $\lambda = \sigma_{ee}/\sigma_{uu}$ assumed known for the univariate model.

Cheng [7] derives the maximum likelihood estimate for the parameter matrix using the method of Lagrange multipliers to obtain maximum likelihood estimates of the parameters. The theory behind the Lagrange multipliers is not derived here, the reader is referred to Stein [26] for a detailed derivation. Also the theory of obtaining the maximum likelihood estimates is not derived here, readers are referred to Cheng [7] for a detailed derivation. However, an example is given in Cheng [7] which involves two explanatory variables (X_1 and X_2). Parameters are derived for both the measurement error model (using the theory derived in Cheng [7]) and for the regression model. The results are illustrated in figure 2.3.1 below (note the measurement error plane is the semi-transparent one, that is the plane with the steepest slope),

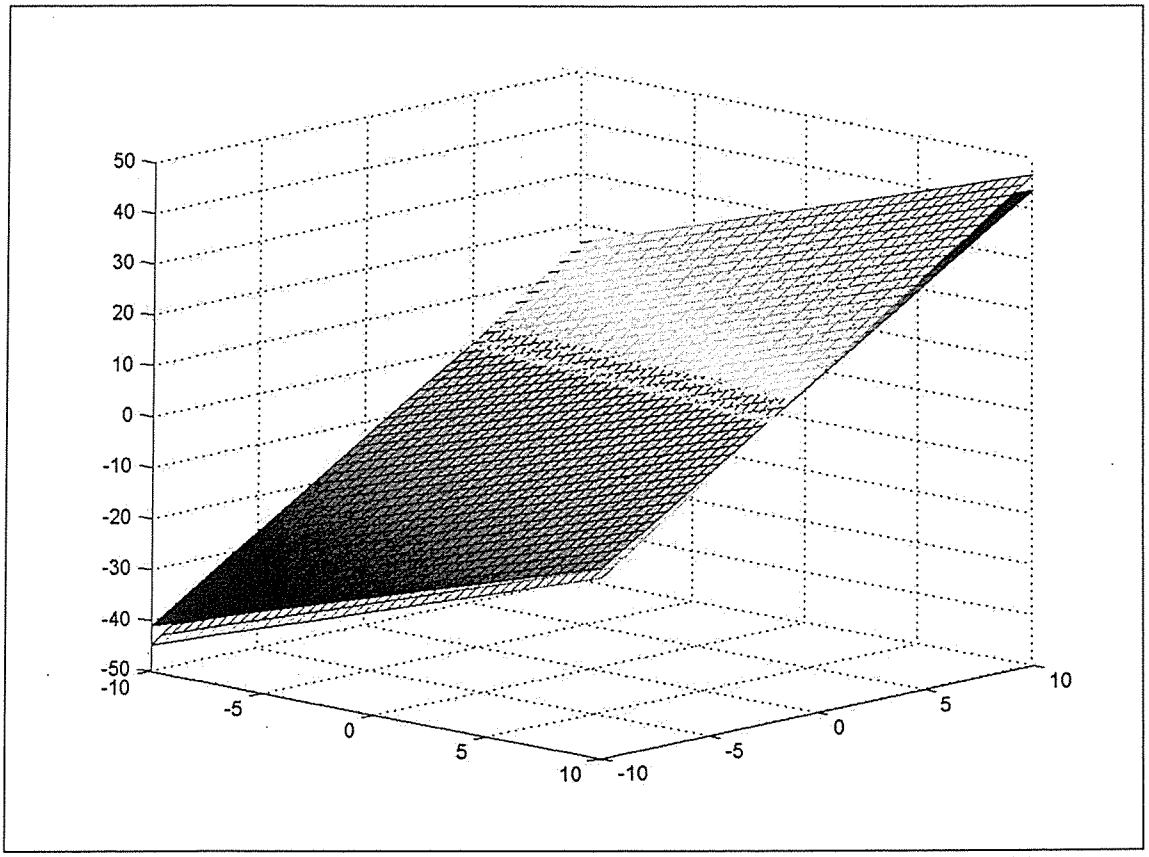

Figure 2.3.1: Measurement error model (transparent) and regression model (solid) with two explanatory variables

We can see from figure 2.3.1 that the two planes (measurement error and regression planes) intersect the mean $(\overline{Y}, \overline{X}_1, \overline{X}_2)$. In addition to this, it is obvious that estimates around this point will be similar for the measurement error model estimate and the regression estimate. However, estimates about the extremities of the planes will be much different (relatively speaking) and hence incur larger errors.

2.4 Non-Linear Measurement Error Models with Vector Explanatory Variables

Fuller [12] derives the theory of non-linear measurement error models in detail. In this section we attempt to give a brief overview of nonlinear measurement error models and point out a few interesting results.

Fuller [12] defines the nonlinear measurement error model as follows,

$$y_i = g(\mathbf{x}_i;\beta), \quad \text{.....(2.4.1)}$$

$$(Y_i, \mathbf{X}_i) = (y_i, \mathbf{x}_i) + (e_i, \mathbf{u}_i) \quad \text{.....(2.4.2)}$$

where $g(\boldsymbol{x}_i;\boldsymbol{\beta})$ is a real valued continuous function[11], $\{\boldsymbol{x}_i\}$ is a sequence of fixed p-dimensional row vectors, $\boldsymbol{\beta}$ is a k-dimensional column vector, and $\boldsymbol{\varepsilon}_i = (e_i, \boldsymbol{u}_i)$ is the vector of measurement errors. It is also assumed that $\boldsymbol{\varepsilon}_i$ is distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$, where at least one element of $\boldsymbol{u}_i$ has positive variance.

Typically in regression, models that are nonlinear in the $\boldsymbol{x}$ are generally called curvilinear rather than nonlinear. Models where $g(\boldsymbol{x}_i;\boldsymbol{\beta})$ are nonlinear in the parameters, $\boldsymbol{\beta}$ are usually called nonlinear regression models. In this section we will concentrate on the latter as this is the form of the volume equations used in MARVL.

When the measurement error models are nonlinear in $\boldsymbol{\beta}$ but linear in $\boldsymbol{x}$, Fuller [12] redefines equations 2.4.1 and 2.4.2 as,

$$Y_i = \sum_{j=1}^{k} g_j(\boldsymbol{\beta}) x_{ij} + e_i,$$

$$X_{ij} = x_{ij} + u_{ij}$$

where $g_j(\boldsymbol{\beta})$ are nonlinear functions of $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}_i = (e_i, \boldsymbol{u}_i) \sim NI(\boldsymbol{0}, \boldsymbol{\Sigma}_{\varepsilon\varepsilon})$, and $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ is known. Fuller [12] states that if $\sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{x}_i$ is nonsingular (that is the determinant is nonzero), estimates of $g_j(\boldsymbol{\beta})$, ignoring the restrictions imposed by $\boldsymbol{\beta}$, can be obtained by the usual linear measurement error models. Using these estimators of $g_j(\boldsymbol{\beta})$, improved estimators of $\boldsymbol{\beta}$ can be obtained by the adaptation of nonlinear procedures.

It should also be noted that some non-linear models can be transformed into linear models[12] and the normal linear regression (see section 2.3) can be applied. The measurement error model used in MARVL to estimate the volume of individual trees in an example of this, and is discussed in the following section.

---

[11] Fuller [12] defines this model to be nonlinear if $g(\boldsymbol{x};\boldsymbol{\beta})$ is nonlinear in $\boldsymbol{x}$ when $\boldsymbol{\beta}$ is fixed or if $g(\boldsymbol{x};\boldsymbol{\beta})$ is nonlinear in $\boldsymbol{\beta}$ when $\boldsymbol{x}$ is fixed.

[12] Draper [11] defines these models as intrinsically linear as these models can be transformed into a linear form, as opposed to intrinsically nonlinear which cannot be transformed into a linear form.

2.5 Measurement Error Models in MARVL

The volume of standing wood product within an area of interest is one of the most important measures in a forest inventory. As Schrueder [9] states, when foresters speak of volume, they refer to the amount of wood in a tree, stand of trees, or other specified area according to some unit of measurement during forest inventory.

The methods available for obtaining estimates for volumes of wood can be grouped into two categories: direct and indirect. Direct methods related to physical contact with the tree through measurement (usually a felled tree, as this makes measurement more accurate), as opposed to indirect methods which involves no physical contact with the tree.

An example of a direct method is using a special piece of equipment known as a xylometer. To determine the volume of a tree a stem is placed in a xylometer and a measure of water displacement is recorded. From this water displacement reading an accurate estimate of volume can be obtained. Though direct methods give more reliable estimates of stem volume than indirect methods, they are more hazardous, more time consuming and hence more costly to apply. For these reasons, direct methods are seldom used in a forest inventory context as time and costs are major considerations.

Indirect methods are most commonly used in forest inventories (such as MARVL) since they are efficient and less costly than direct methods. The most common form of indirect method for obtaining an estimate of volume is the use of volume equations. It should be noted that MARVL uses this method. Volume equations are much like regression equations in the sense that a number of independent variables are used to predict volume. A typical volume equation used by MARVL is presented below,

$$Volume\_under\_Bark = e^{\beta_1} d^{\beta_2} \left[h^2 / (h - 1.4)\right]^{\beta_3} + \varepsilon$$

Where

d = Diameter at breast height (Over Bark)

h = Height of the tree

It should be noted that this volume equation can be linearized (in terms of the parameters) in the sense described below. This would allow the use of multiple linear regression and general linear model (GLM) theory, except that it is necessary to note that the equation for 'Volume under bark' is in fact a measurement error model.

$$\log(VUB) = \log\left(e^{\beta_1} d^{\beta_2} V_h{}^{\beta_3} + \varepsilon\right)$$
$$\Rightarrow \log(VUB) = \log\left(e^{\beta_1} d^{\beta_2} V_h{}^{\beta_3} (1+\varepsilon')\right)$$
$$\Rightarrow \log(VUB) = \log\left(e^{\beta_1} d^{\beta_2} V_h{}^{\beta_3}\right) + \log(1+\varepsilon')$$
$$\Rightarrow \log(VUB) = \beta_1 + \beta_2 \log(d) + \beta_3 \log(V_h) + e$$

where

VUB = Volume under Bark

d = diameter at breast height (Over Bark)

h = Height of the tree

$$V_h = \frac{h^2}{(h-1.4)}$$

$$\varepsilon' = \frac{\varepsilon}{e^{\beta_1} d^{\beta_2} V_h{}^{\beta_3}}$$

$$e = \left[\varepsilon' + \frac{\varepsilon'^2}{2!} + \frac{\varepsilon'^3}{3!} + \ldots\right] = \log(1+\varepsilon')$$

Note that the variance-covariance structure of this model changes at each iteration via ε' and the Taylor series expansion of the error term.

Depending on the site, age and condition of the trees within an area of interest, suitable volume equations must be developed for the different conditions. In New Zealand, volume equations relating to different sites, ages and conditions for the *Pinus Radiata* species, have been developed and are used by MARVL.

Schreuder [25] states that most volume equations typically consist of the following independent variables,

- *Diameter at Breast Height (DBH)*, which is defined to be the diameter of a tree 1.4 meters from the ground[13] and is denoted by *d*.
- *Total Height*, which is defined as the vertical distance from ground level to its uppermost point and is denoted by *h*.
- *Bark Thickness*, defined as the distance between the cambium and region of convex closure outside the bark.
- *Taper*, defined as the way the tree decreases in diameter for base to tip.

[13] This is true for USA, New Zealand, Burma, India, Malaysia and South Africa. However, for countries such as Continental Europe, United Kingdom, Australia and Canada, DBH is measured at 1.3 meters from the ground.

Similar to obtaining the volume of a tree or stand of trees, there are two methods that can be used to obtain the four measures presented above, namely direct and indirect methods. MARVL uses indirect methods to obtain estimates of each of the variables above. For this reason the following section discusses the possible errors and magnitude of these errors for obtaining variables outlined above using indirect methods only.

2.6 Sources of error for the independent variables of the volume equation

2.6.1 Sources of Error in DBH

Firstly we consider diameter at Breast Height (DBH). Errors in the estimation of DBH arise from three main sources: the tree, the instrument[14], and the operator.

Schreuder [25] discusses the effect of each of these sources of error. For a perfectly circular tree no error occurs with either instrument (tape or calipers), however trees are seldom perfectly circular. For trees that are noncircular, the tape always obtains estimates of volume that are positively biased, that is the estimate is an overestimate (note: the bias increases as the tree shape becomes less circular), however Schreuder [25] states that this bias is generally small, nonetheless an error still exists. The error obtain by using calipers to measure DBH on a noncircular tree maybe larger or smaller than that of the tape and positive or negative, depending on where the caliper arms are placed.

The error associated with the instrument used in obtaining DBH is the second source of error. This error is not as severe as the source of error associated with the tree (discussed above), nevertheless it still must be discussed. The error is negligible for correctly graduated steel tapes but can occur with cloth and fiberglass tapes due to stretching with continued use. Calipers are more subject to instrumental error than tapes because they are more easily damaged. The extent of this error varies considerably and hence Schreuder [25] does not discuss the magnitude of this error.

Schreuder [25] identifies 3 main sources of operator error: incorrect location of the point of measurement on the tree, incorrect tension (tape) and pressure on the arms (caliper) at the time of measurement and incorrect reading of the scale of the instrument. Most of these errors can be minimized and even avoided entirely by thorough training of the measuring crew.

2.6.2 Sources of Error in Measuring Height

The height of the trees is a variable of fundamental importance in forest measurement because it is used to describe the quantity and quality of tree and forest growth. As we can see in the volume equation presented above, height is one of the independent variables, so it is important to consider the possible forms of error associated with it. To obtain an estimate of height of a standing tree is a difficult task since many variables must be considered. Trigonometric methods are generally used to obtain an estimate of height. Figure 2.6.1 below demonstrates how trigonometry is used,

[14] The instruments generally used to obtain a measure of DBH are tape measures or calipers. Note: MARVL uses tape measures in the inventory process.

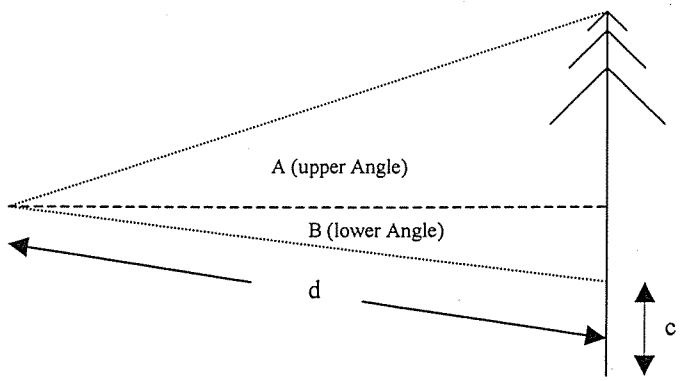

Figure 2.6.1: Demonstration of Trigonometric methods used to estimate Height

Using figure 2.6.1, we can obtain an estimate of height use the following formula,

$$Height = d\left[\frac{\sin(A-B)}{A}\right]+c$$

Unfortunately, trees seldom grow perfectly horizontal. In fact it has been shown that on sloping ground, trees lean down hill. This point leads to a major source of error in obtaining a height estimate. Schrueder [9] illustrates that a lean of 10 degrees can result in an error in the estimate of height of anywhere between –16.1% and 19.2%, depending on the direction from which the tree is observed. Schrueder [9] also shows that by measuring the height perpendicular to the plane of lean this error can be minimized. Employing this method for a tree with a 10 degree lean results in an error of –3.6%, which is a highly significant improvement.

2.6.3 Sources of Error in Measuring Bark Thickness

In most countries, timber is sold on an inside bark basis. To obtain such a result a measure known as diameter inside bark or diameter under bark (DIB or DUB respectively) must be obtained. DUB is the name given to diameter at breast height minus the thickness of bark at this point. MARVL requires indirect methods for obtaining such a measure since MARVL is a pre-harvest inventory system and hence does not physically harm the trees.

Similar to volume estimates, functions must be developed which generally uses DBH and height as explanatory variables to predict DUB. However, due to the high variability of bark thickness within a tree, estimates of volume under bark can vary substantially and hence can incur significant errors. Schreuder [25] discusses a bark thickness model developed by Johnson and Wood (1987) which was developed for radiata pine plantations in the Australian Capital Territory of Australia. The model is independent of site and age and predicts back thickness (RB_I) at any point along the tree, relative to bark thickness at breast height. The model is as follow,

$RB_i = b_0 + b_1 RD_i^4$

Where

RB_i = Bark thickness at any point, i, along the tree, relative to bark thickness at breast height

RD_i = Ratio of diameter over bark at i to that at breast height

b_0 and b_1 = simple linear regression coefficients

Using the model presented above, Johnson and Wood estimated the under bark volume of 109 trees with an average arithmetic error of –0.5% assuming a linear regression rather than a measurement error model. We can see that the error associated with bark thickness can be very small, however it must be considered when obtaining an estimate of total volume.

2.6.4 Sources of Error in Measuring Taper

Again indirect methods for obtaining taper estimates are performed by MARVL, namely taper equations. Goulding and Murray (1976) and Gordon (1983) developed polynomial volume-compatible taper equations to predict diameter under bark at any point along the tree stem of radiata pine in New Zealand. The general form of the polynomial equations used by MARVL are presented as follows,

$$d(h) = \left[(v_0 / KH)(b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4 + b_5 x^5 + b_6 x^6)\right]^{0.5}$$

$$v_t = v_0 (b_7 x^2 + b_8 x^3 + b_9 x^4 \ldots + b_{l+5} x^l)$$

Where

$d(h)$ = DUB at any point along the stem (cm),

v_0 = total volume under bark (m^3),

K = a constant (in metric : $0.0000785398\ m^2 / cm^2$),

H = Total tree height (m),

$x = (H - h)/H$ where h is the height (m) to d(h),

v_t = Volume from tip to d(h), that is, volume of the section (H - h) (m^3),

$b_1, b_2, \ldots, b_{12}$ = coefficents, some which may be zero,

l = an integer, $4 < l < 40$.

The first of these equations are nonlinear in both the parameters and x, however the second equation is linear in the parameters but non-linear in x. Again they would be better fitted using measurement error models rather than regression, although to date this has not been done, at least in New Zealand.

In addition to taper equations being used to predict volume, taper is also an important variable in some of the sampling designs used in today's forest inventories, namely importance sampling and critical height sampling. Because taper equations yield a measure of taper indirectly, an error is incurred which needs to be included in the total error calculation.

These errors associated with each variable (DBH, Height, Bark thickness and taper) are important when considering the volume equation. Because there are obvious errors associated with obtaining an estimate for each of the variables the general conclusion is that the volume equation used by MARVL is not a regression model but a measurement error model and must be analyzed accordingly.

2.7 Simulation Study

2.7.1 Setting up the simulation

First, let us consider the volume equation (a nonlinear regression function) currently used by MARVL to estimate the volume of an individual tree,

$$Volume\_under\_Bark = e^{\beta_1} d^{\beta_2} \left[h^2/(h-1.4)\right]^{\beta_3} + \varepsilon \quad \text{.....(2.7.1)}$$

Where

d = Diameter at breast height (Over Bark) = DBH

h = Height of the tree (in meters)

ε = Error

In section 2.5 it was noted that this equation could be linearized by taking logs, hence obtaining

$$\log(VUB) = \beta_1 + \beta_2 \log(DBH) + \beta_3 \log(V_h) + e \quad \text{.....(2.7.2)}$$

where

VUB = Volume under Bark

DBH = diameter at breast height (Over Bark)

h = Height of the tree

$$V_h = \frac{h^2}{(h-1.4)}$$

$$\varepsilon' = \frac{\varepsilon}{e^{\beta_1} d^{\beta_2} V_h^{\beta_3}}$$

$$e = \left[\varepsilon' + \frac{\varepsilon'^2}{2!} + \frac{\varepsilon'^3}{3!} + \ldots\right] = \log(1+\varepsilon')$$

Note that the solution to this equation for log volume under bark requires iteration since the covariance for the error e, depends on $\{\beta_1, \beta_2, \beta_3\}$ and that this can be achieved via generalized linear models.

Currently MARVL assumes that the predictor variables in the volume equation presented above are measured free from error, however in practice this is seldom the case. In this simulation study we wish to investigate the effect that incorrectly measuring the predictor variables has on total volume. The theory outlined in this chapter allows us to investigate this and is used in the following simulation study.

The first task of this simulation study is to identify the relevant factors needed. If we recall from the theory presented above, to obtain parameter estimates for the measurement error model an additional piece of information is required. Let us assume for this simulation study that the Σ_{uu} matrix (that is, the covariance matrix of the measurement errors) is known. We can see from the volume equation above that this is a 2 by 2 matrix and is given as

$$\Sigma_{uu} = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

Where

$$a = \sqrt{Var(\log(DBH))}$$

$$b = \sqrt{Var(\log(Vh))}$$

This covariance matrix indicates that there are three factors involved namely the variance of the log of DBH, the variance of the log of V_h and the correlation between these. Having these three factors allows us to select different levels of error for the two predictors and also different levels of the correlation between these two errors.

Because we are using the log transformed volume equation we must take care when selecting the levels of the variance of the log of DBH and variance of the log of V_h. Errors made on measurements on DBH and V_h would be approximately 4% of the true value. This corresponds to an error in DBH of 2cm for a tree of diameter 50cm, and an error of about 1.5m in a tree of 35m. If this is interpreted as providing the range in which 95% of measurements would fall, this leads to an approximate standard error on the log scale of 0.02.

To illustrate how this standard error is calculated consider the following argument. Let us consider DBH. The range of DBH is given by the interval,

$$DBH = 50 \pm 2cm = 50cm \pm 4\%$$

However, we are using the log scale, hence the interval for the log of DBH is,

$$\log(DBH) = \log(50 \pm 2) = \log[50(1 \pm 0.04)]$$

Thus,

$$\log[50(1-0.04)] \le \log(DBH) \le \log[50(1+0.04)] \quad \text{(with 95\% probability)}$$

$$\Rightarrow \log(50) - \log(1.04) \le \log(DBH) \le \log(50) + \log(1.04) \quad \left(\text{since } 0.96 \approx \frac{1}{1.04}\right)$$

Now, $\log(1.04) \approx 0.04$, so

$$\log(50) - 0.04 \le \log(DBH) \le \log(50) + 0.04 \quad \text{(with 95\% probability)},$$

This can be interpreted as $\log(DBH)$ having a standard error of $\frac{0.04}{2} = 0.02$, hence the variance of the log of DBH is $0.02^2 = 0.004$. Note a similar argument yields the same standard error and variance for the log of V_h as for the log of DBH.

2.7.2 Simulation

Based on the information above the simulation was structured so that the levels of the variance of the log of DBH and variance of the log of V_h should be 0.001, 0.003, 0.005 and 0.007. And the levels of ρ, that is the correlation between variance of the log DBH and variance of the log of V_h are 0 (that is, no correlation), 0.2, 0.4, 0.6 and 0.8.

The simulation was based on a MARVL data set of 233 trees for which DBH, V_h and volume under bark was known. The data set was prepared for Fletcher Challenge Forests by NZ Forest Research Institute Ltd and contains trees for different compartments throughout the Kaingaroa Forest which is situated in the central North Island of New Zealand. Boxplots for volume under bark, DBH and Vh for these 233 trees are presented below in figures 2.7.1-2.7.3.

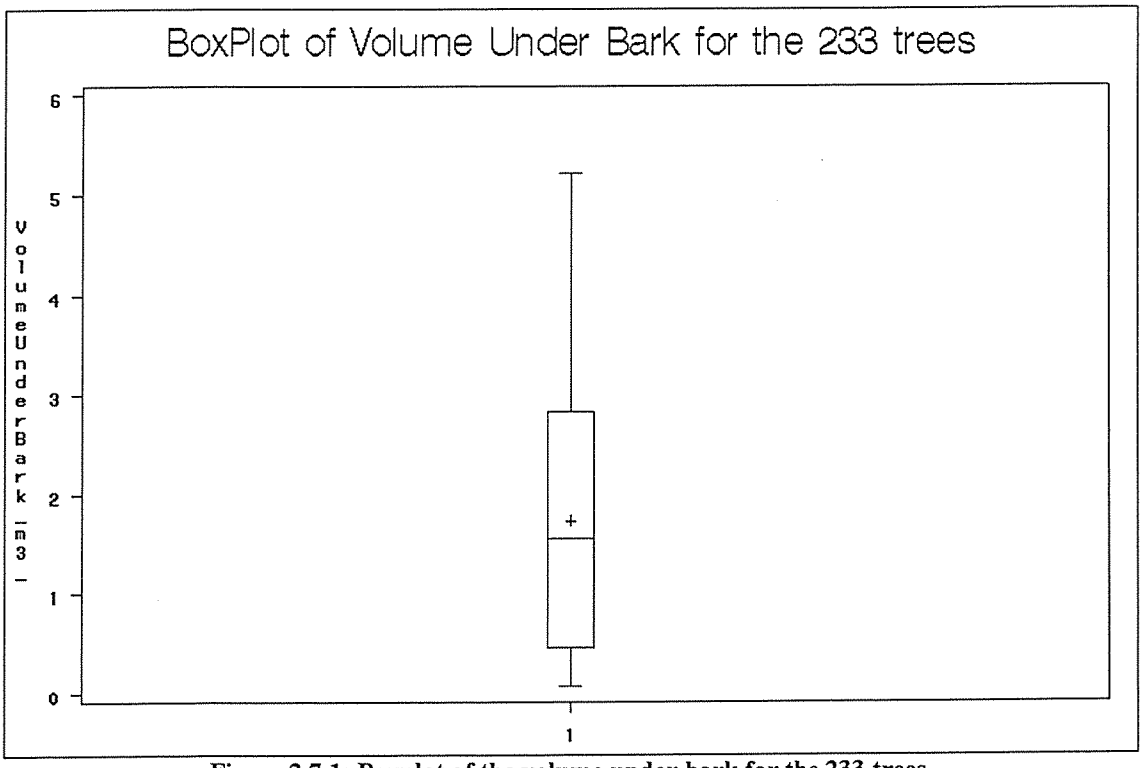

Figure 2.7.1: Boxplot of the volume under bark for the 233 trees

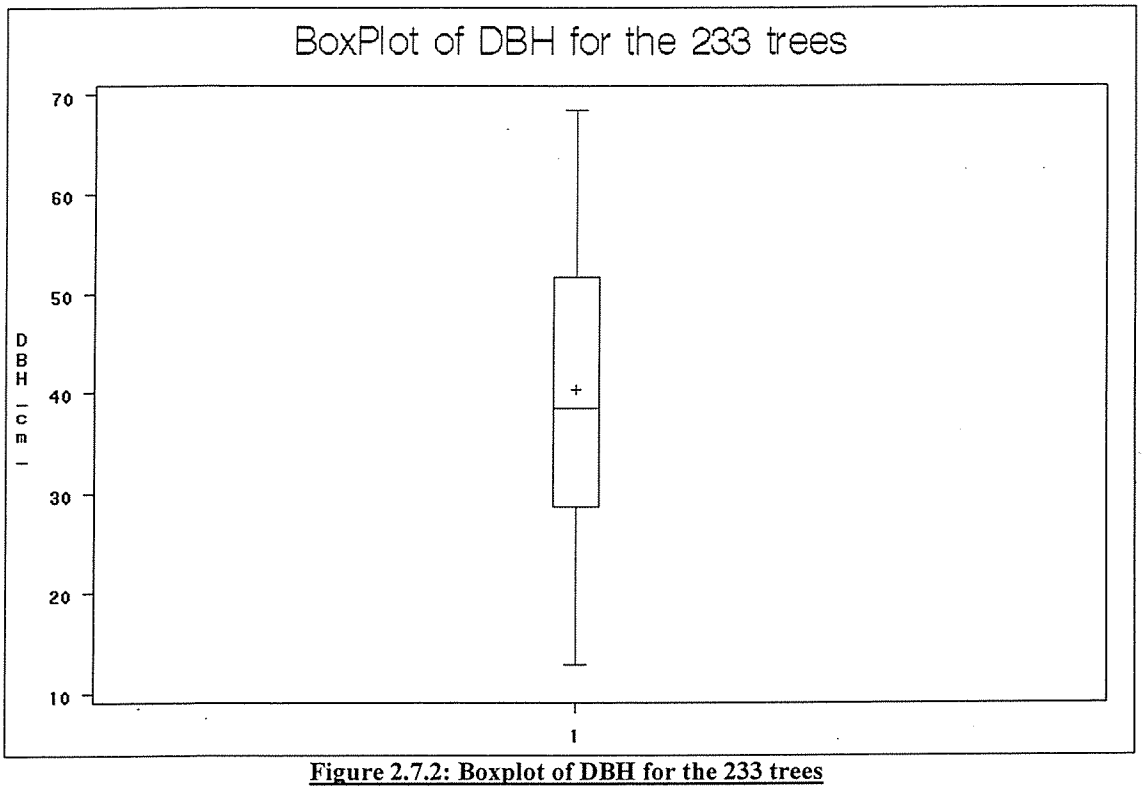

Figure 2.7.2: Boxplot of DBH for the 233 trees

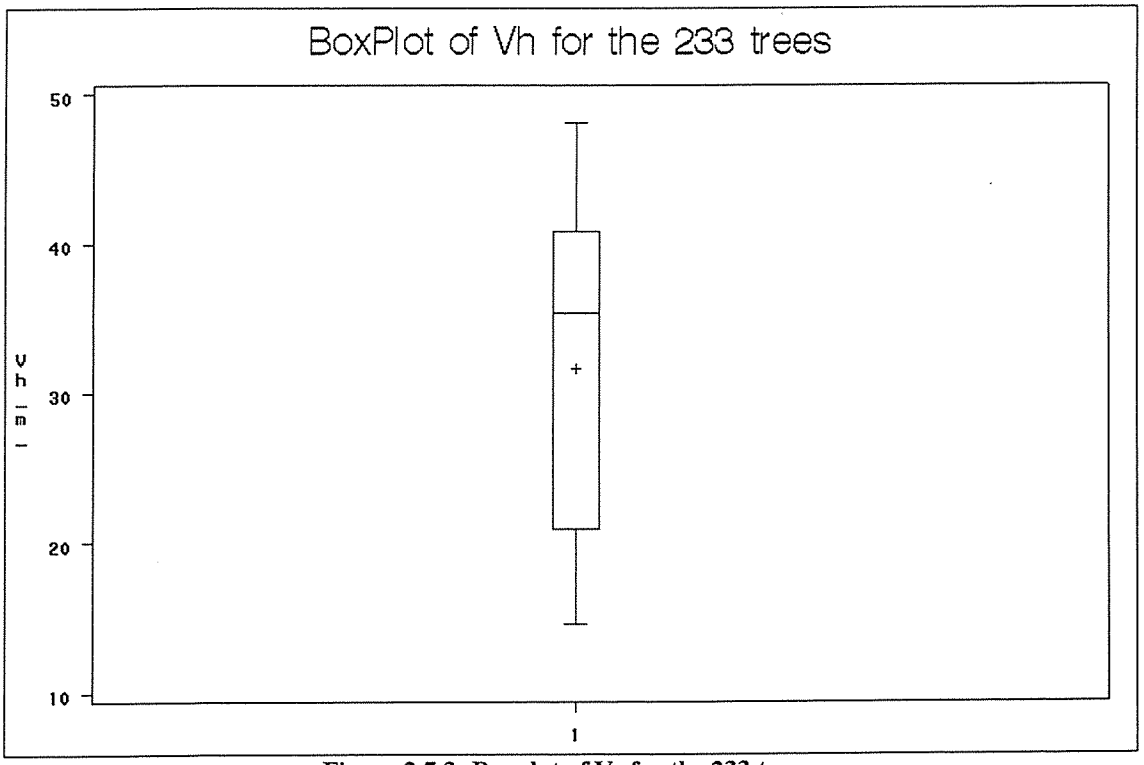

Figure 2.7.3: Boxplot of V_h for the 233 trees

In addition to figures 2.7.1 – 2.7.3 a number of summary statistics were obtained for volume under bark, DBH and V_h and are present below in table 2.7.1.

Variable	Mean	Standard Deviation	Standard Error
Volume under Bark (m^3)	1.7254	1.3423	0.0879
DBH (cm)	40.269	14.061	0.921
V_h (m)	31.641	10.098	0.662

Table 2.7.1: Summary statistics for the 233 trees

To the DBH and V_h variables were added measurement errors determined by the simulation for each value of Var(log(DBH)), Var(log(V_h)) and ρ. Ten replicates were formed for each combination of these factors. Volumes were then estimated using the MARVL equation and the simulated DBH and V_h which contain measurement error. The experimental design was a repeated design with four repeated measures, namely Var(log(DBH)), Var(log(V_h)), replicate and ρ. For each simulated combination of Var(log(DBH)), Var(log(V_h)) and ρ, estimated volume was determined from the MARVL equation parameters from the original data set.

Note that the way the simulation is structured does not imply that either DBH or V_h is consistently over (or under) estimated, only that the percentage error of measurement changes, and that volume under bark is estimated for varying values of the correlation between these variables.

However although the errors are close to symmetric about zero for DBH and for V_h, the error for volume under bark need not have this property. To see this, consider for example the situation where DBH and V_h (or their logarithms) are highly correlated. If DBH and V_h were each overestimated by say 10%, and volume under bark were proportional to DBH squared times V_h then volume under bark would be overestimated by 33.1%., but if each were underestimated by 10%, volume under bark would only be underestimated by 27.1%. Although the model used is not this simple, the same general argument applies. Increasing even symmetrical measurement error in DBH and V_h can lead to upward biases in volume under bark estimates.

2.7.3 Simulation Results

The parameter estimates for the linear simulation model are not given explicitly below, but estimable combinations of parameters in the form of average estimated volumes are available in figures 2.7.4 – 2.7.14

Each plotted point in these figures are both the best linear unbiased estimate under the repeated measures linear model, and the subgroup average. The fact that many of the plots are approximately linear is not because linear regression type effects were fitted in the model; they were not. In fact lines in these figures are not exactly straight, which reflects the fact that var(log(DBH)), var(log(Vh)) and ρ have been divided into categories rather than treated as regression effects. The regularity of the estimates and the stability (and similarity) of standard errors in each plot is then not because a linear regression model was fitted. The tight structure is a real effect not an artifact of the analysis, and instead reflects the fact that the effects of var(log(DBH)), var(log(V_h)) and ρ are very regular.

Standard errors are also given in the figures. These reflect the number of replicates used in the simulation (in this case, 10). Increasing the number of replicates would improve the significance levels (since this would decrease the standard errors) but would not alter the ordering of importance of the effects.

The results for all the main effects and interactions obtained from SAS, are summarized in table 2.7.2 below, it should be noted that the P-Value presented in table 2.7.2 are adjusted Greenhouse-Geisser P-Values. The Greenhouse-Geisser adjusted F-value adjusts the error degrees of freedom to compensate for the correlation between the repeated measures.

Effect	F-value	Df (num,den)	Adj P-Value
Var(log(DBH))	17.97	(3,696)	<0.0001
Var(log(V_h))	3.58	(3,696)	0.0596
ρ	38.86	(4,928)	<0.0001
Var(log(DBH))* Var(log(V_h))	7.02	(9,2088)	<0.0001
Var(log(DBH))* ρ	7.99	(12,2784)	0.0014
Var(log(V_h))* ρ	48.07	(12,2784)	<0.0001
Var(log(DBH))*Var(log(V_h))* ρ	3.35	(36,8352)	<0.0001

Table 2.7.2: Summary of the Simulation Statistical Tests

In order of priority the statistically significant effects are Var(log(V_h))* ρ, ρ, Var(log(DBH)), Var(log(DBH))* Var(log(V_h)), Var(log(DBH))* ρ, Var(log(DBH))*Var(log(V_h))* ρ and Var(log(V_h)). Note some interactions are more important than the main effects. This is discussed in greater detail below.

It should be noted that the estimate of log volume was back-transformed to volume (using the exponent function) before the ANOVA model was fitted. The volume estimates still conform to equations (1) and (2). Back-transforming aids intuitive understanding of the results because tree volume is the variable of interest, and the focus is on the effect of the levels of the predictor variables; var(log(DBH)), var(log(V_h)) and ρ on tree volume estimates in the presence of measurement error under model (1). Hence all tests and profile plots use volume (not log volume) as the dependent (y) variable, and the graphs below use tree volume as the vertical axis.

Firstly, let us consider the three main effects. We can see from table 2.7.2 that there is a significant effect for variance of the log diameter at breast height (Var(log(DBH))). This means that incorrectly obtaining the measure of diameter at breast height leads to a significant difference in the estimate of mean volume. We can see from the profile plot below (figure 2.7.4) that the larger the error in the log of DBH the greater the volume, as the mean volume is increasing for each level of the variance of log of DBH.

In addition to this the profile plot (figure 2.7.4) also shows the standard error for each of the four points. We can see that the standard error regions of the first point and the last two points do not overlap, which demonstrates the significant Var(log(DBH)) effect.

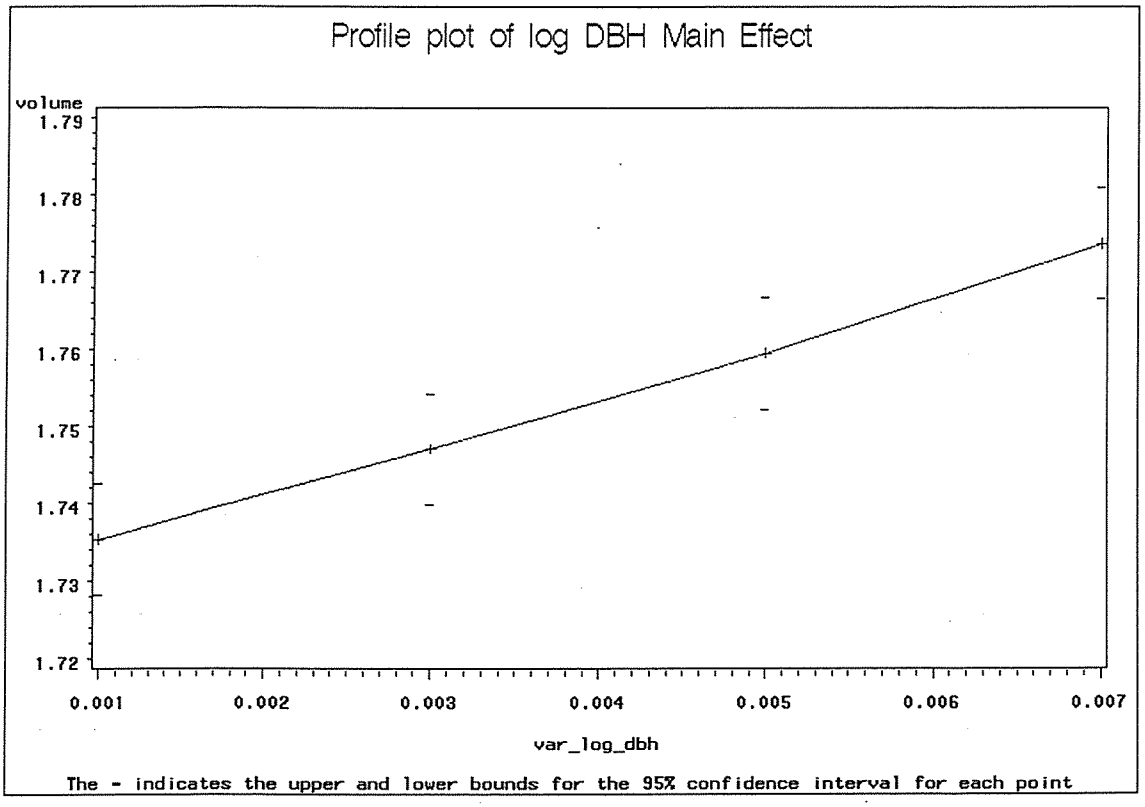

Figure 2.7.4: Profile plot of the Var(log(DBH)) main effect

We can see from table 2.7.2 that incorrectly measuring the height of the tree in itself is the smallest effect (via V_h) on the estimate of mean volume (P-Value=0.0596, based on the ten replicates used in this simulation). This result is encouraging, in terms of MARVL, as obtaining a measure of height for an individual *Pinus Radiata* can sometimes be difficult, due to intertwining branches from neighboring trees at the crown canopy. Again let us consider the profile plot of this main effect (see figure 2.7.5 below) we can see that this plot is very similar to the variance of the log of DBH plot above, however we must consider the y axis scale and the standard errors of each point to give a reason for insignificance of this main effect.

When considering main effects only, the effect of errors in DBH is more important than those in V_h for estimating volume under bark.

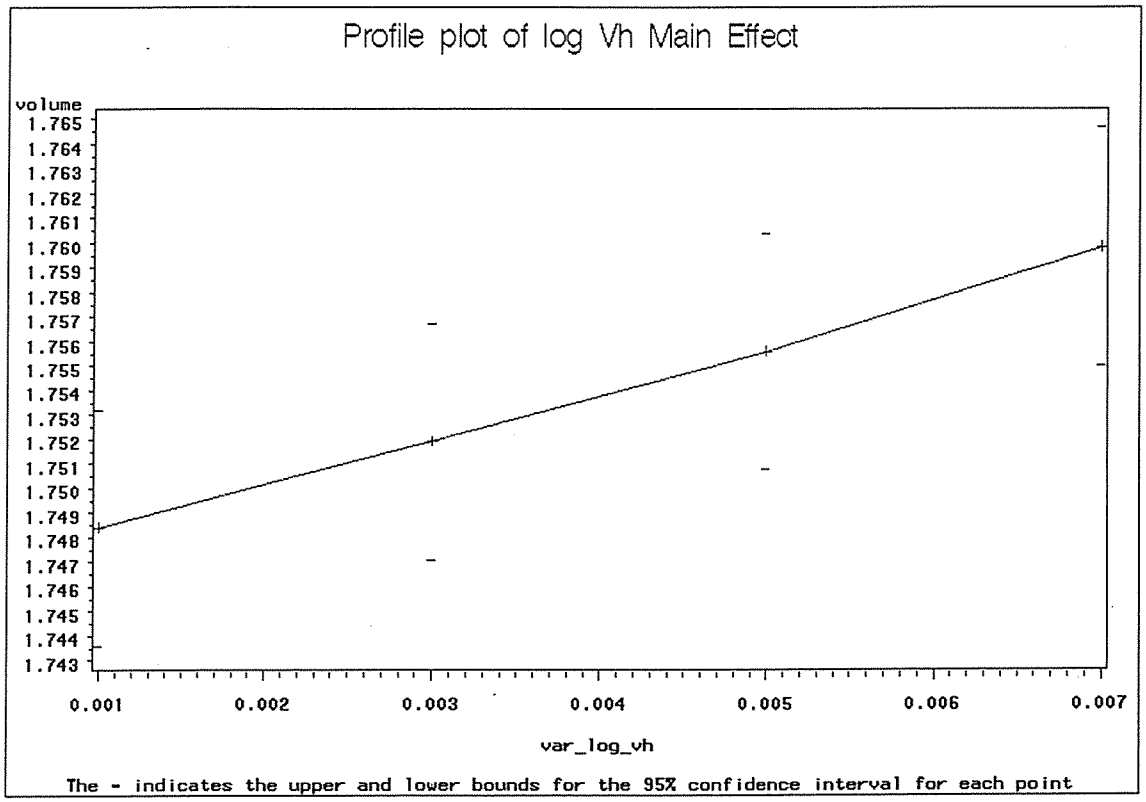

Figure 2.7.5: Profile plot of the Var(log(V_h)) main effect

We can see that the range of volume over the levels of V_h is approximately between 1.748 and 1.760, which is a difference of 0.012. The profile plot of the variance of the log of DBH on the other hand has a difference of 0.04, which is obviously much larger than the difference for the variance of the log of V_h, and consequently it is this second test that is not significant. Also consider the standard errors for each point (see figure 2.7.5), we can see that these are similar for Var(log(V_h)) and Var(log(DBH)). The smaller difference in mean volume over the levels of Var(log(V_h)) is the reason for the marginally significant result for Var(log(V_h)).

Let us now consider the main effect of ρ. From table 2.7.2 we can see that the correlation between the variance of the log of DBH and the variance of the log of V_h has a significant effect on the estimated mean volume. If we consider the following profile plot (figure 2.7.6), we can in fact see that as the level of ρ increases, so to does the mean volume. That is, the more correlated the measurement errors are between DBH and V_h the larger the estimate of mean volume. This is not surprising as high correlation corresponds either to overestimation (or underestimation) of both DBH and V_h simultaneously for each tree. We can see from the profile plot of this effect below, that as the correlation between the two effects increase so too does the estimated volume, which is shown by the increase in volume from left to right. Also the small standard errors of each point leads to the significant result.

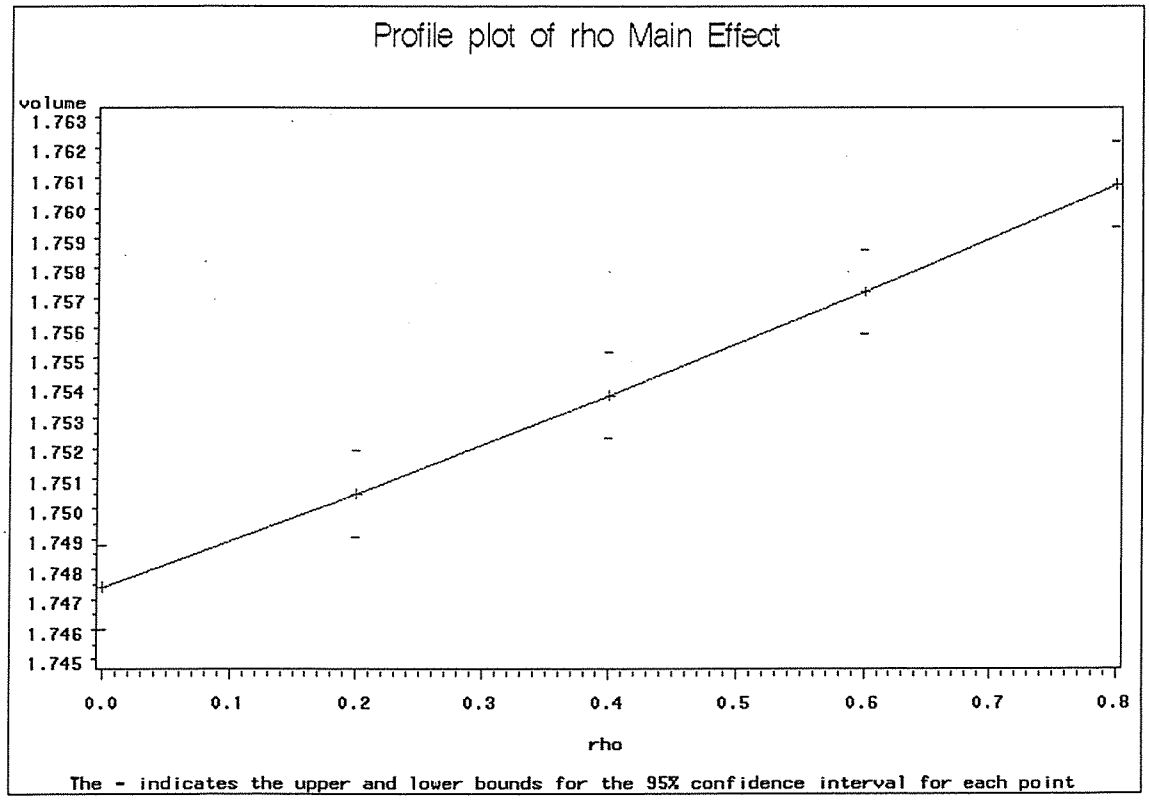

Figure 2.7.6: Profile plot of the rho main effect

Now lets consider the two-way interaction effects. Firstly from table 2.7.2 the P-Value for the variance of the log of DBH and the variance of the log of V_h interaction suggests that there is a significant interaction effect between these two factors (since the p-value = <0.0001). This suggests that as the levels of one factor increases, the effect of the other factors depends on the particular level of the first (factor). Let us consider the profile plot of the variance of the log of DBH and the variance of the log of V_h presented below (figure 2.7.7),

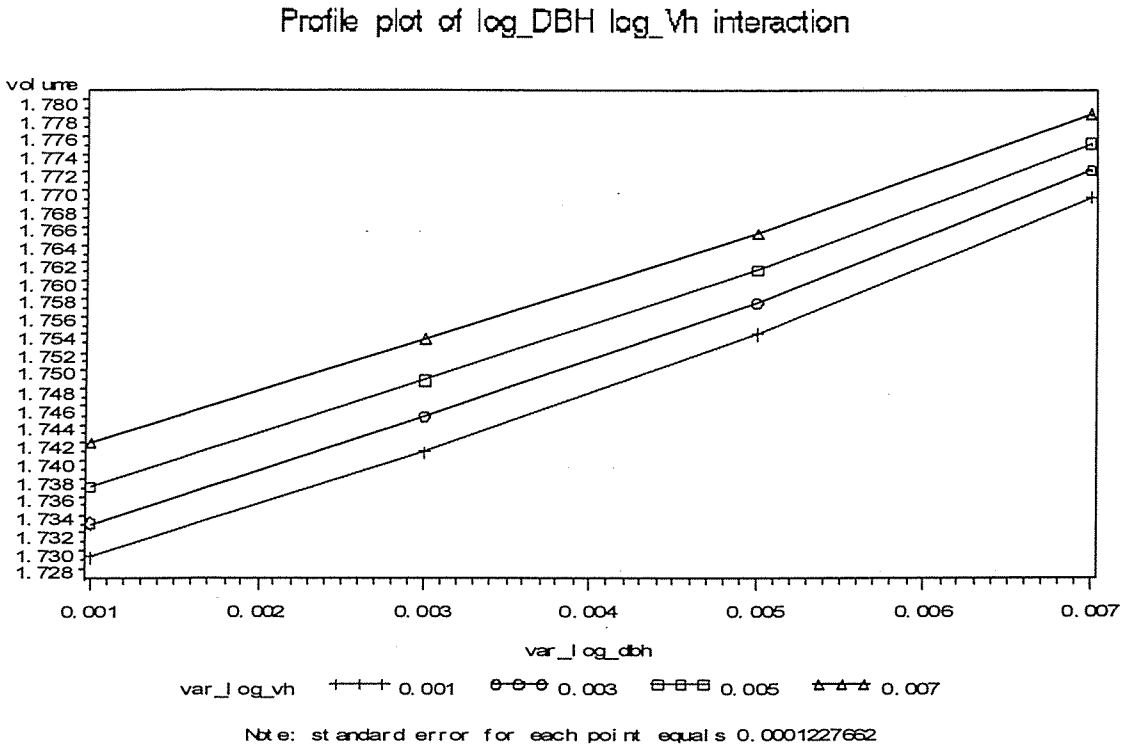

Figure 2.7.7:Profile plot of the Var(log(DBH)), Var(log(V_h)) interaction

We can see from figure 2.7.7 that the four lines (which indicate the four levels of the variance of log of V_h) are actually converging.

To help explain why these lines are converging let us consider the volume equation, or more importantly the log of the volume equation,

$$\log(VUB) = \beta_1 + \beta_2 \log(DBH) + \beta_3 \log(V_h)$$

$$\Rightarrow \text{var}(\log(VUB)) \approx {\beta_2}^2 \text{var}(\log(DBH)) + {\beta_3}^2 \text{var}(\log(V_h)) \text{ if } \rho = 0$$

$$\Rightarrow SE(\log(VUB)) \approx \sqrt{{\beta_2}^2 \text{var}(\log(DBH)) + {\beta_3}^2 \text{var}(\log(V_h))} \text{ if } \rho = 0$$

$$= \sqrt{a^2 + b^2} \text{ using the notation earlier in this section.}$$

We can see from the standard error of the log of volume under bark that the diameter at breast height variable drives the standard error (since $\beta_2 >> \beta_3$), hence high variance of DBH (or log DBH) will increase the standard error more than high variance for V_h (or log V_h). When a^2 is relatively small (that is, low values of Var(log(DBH))), then b^2 (which is always relatively small for values of Var(log(Vh)) used in the simulation) has a more marked effect on the standard error of the log of VUB, than when a^2 is large. This is why the lines in the plot converge as Var(log(DBH)) increases. This observation is especially true for small values of ρ. For higher values of ρ,
$Var(\log(VUB)) \approx \beta_2^2 Var(\log(DBH)) + \beta_3^2 Var(\log(V_h) + \beta_2 \beta_3 Cov(\log(DBH), \log(V_h))$
and since the last term in this equation is positive, the convergence effect is less pronounced.

The second two way interaction in table 2.7.2 (variance of the log of DBH and ρ) yields a highly significant result (p-value<0.0001). This indicates a strong interaction between these two factors (variance of the log of DBH and ρ). Let us consider the profile plot below (figure 2.7.8),

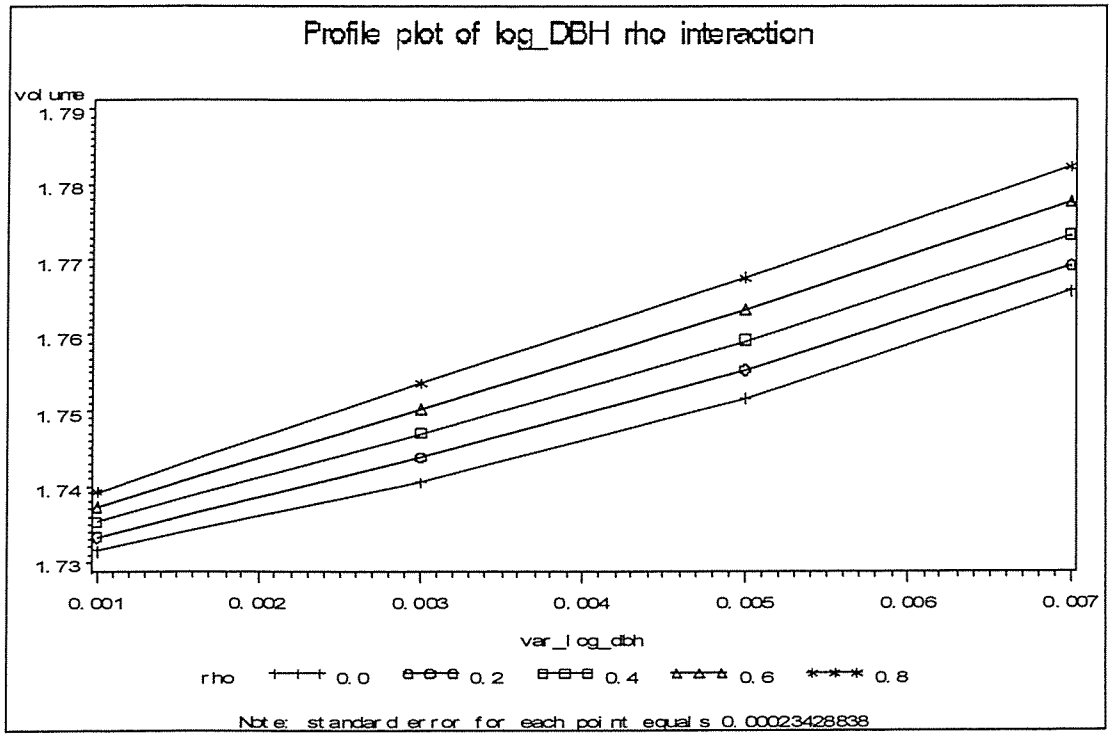

Figure 2.7.8: Profile plot of the Var(log(DBH)), rho interation

From figure 2.7.8 we can see that low levels of the variance of the log of DBH are much less varied over the levels of ρ, than higher levels of the variance of the log of DBH. This means that with increased variance of DBH the mean volume of the tree tends to become more varied over the different levels of ρ (hence the spanning effect of the lines).

This indicates that if the DBH of an individual tree is measured with error, and if this measurement is highly and positively correlated to the error of measuring the height of the tree, (for example, under estimating DBH and under estimating height or over estimating DBH and over estimating height) this will lead to decreased or increased volume estimates and the change in estimated volume will be statistically significant. This result is quite intuitive, for example if one under estimates the DBH and subsequently under estimates the height of the tree, then this will result in a more marked under estimate of the timber volume for this tree than if undermeasurement of DBH is not related to an underestimate of V_h.

It is also interesting to note that for a similar reason for each level of the variance of the log of DBH, the mean volume is largest for high correlation between the variance of the log of the DBH and height factors and smallest for low correlation. This can be seen in the profile plot above.

As we can see from the profile plot below (figure 2.7.9), similar conclusions can be obtained for the very strong V_h, ρ two-way interaction. That is, for low levels of error in measuring height, the variance over fixed values of ρ is much smaller than for high levels of error in measuring height. This interaction is particularly important given its statistical significance, because although the main effect of V_h is marginally significant only, in the simulation, if V_h and DBH are highly correlated then errors in V_h are nevertheless important in estimating volume under bark.

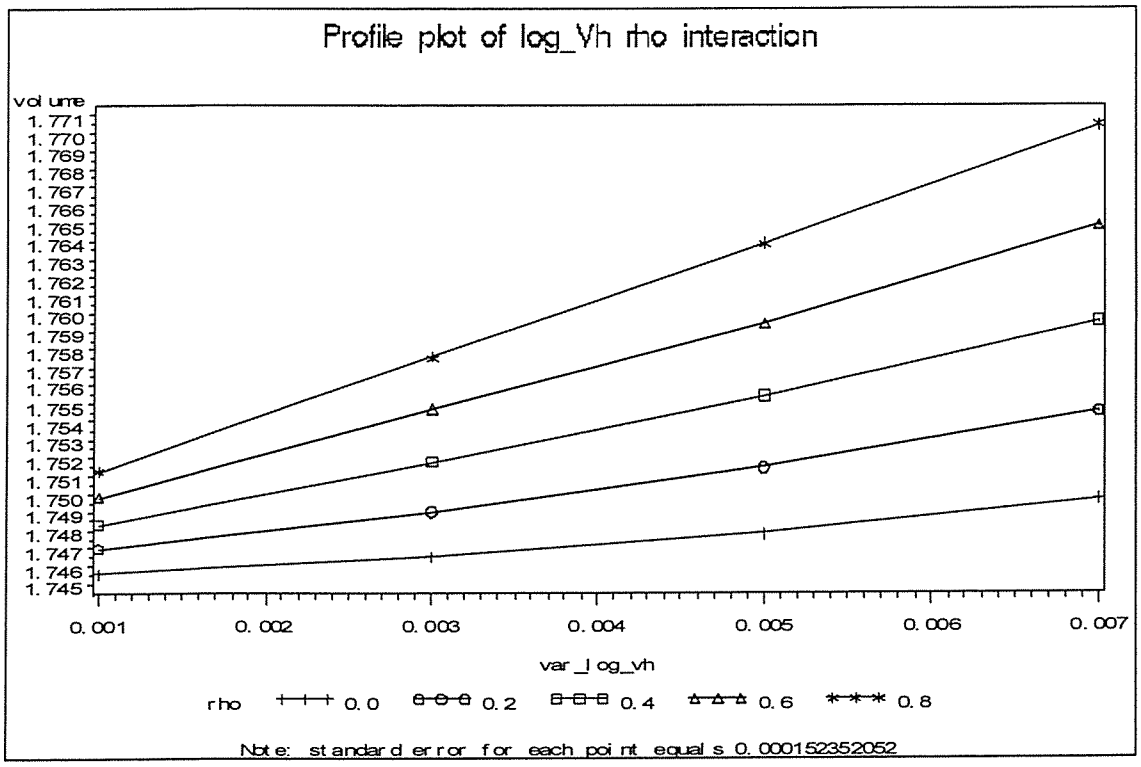

Figure 2.7.9: Profile Plot of the Var(log(V_h)), rho interaction

Finally let is consider the three way interaction between the variance of the log of DBH, the variance of the log of V_h and ρ. From table 2.7.2 we can see that this interaction is highly significant (p-value<0.0001). To explain the significance of the three way interaction the following profile plots (figures 2.7.10 – 2.7.14) were constructed,

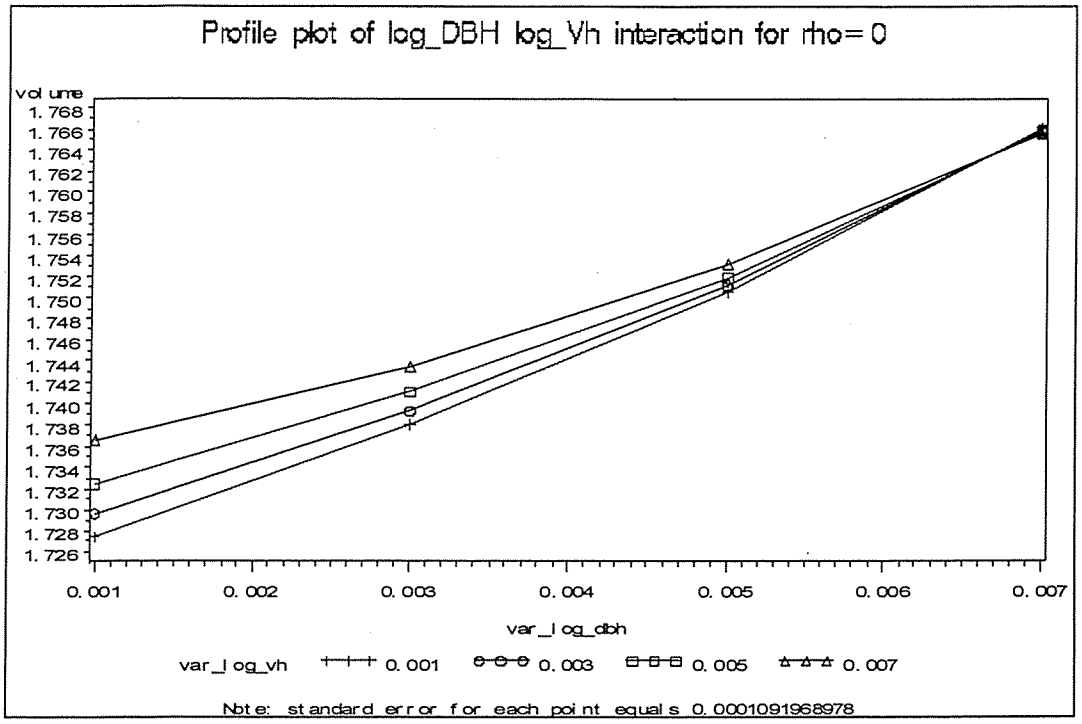

Figure 2.7.10: Profile plot of the Var(log(DBH)), Var(log(V_h)) interaction when rho=0

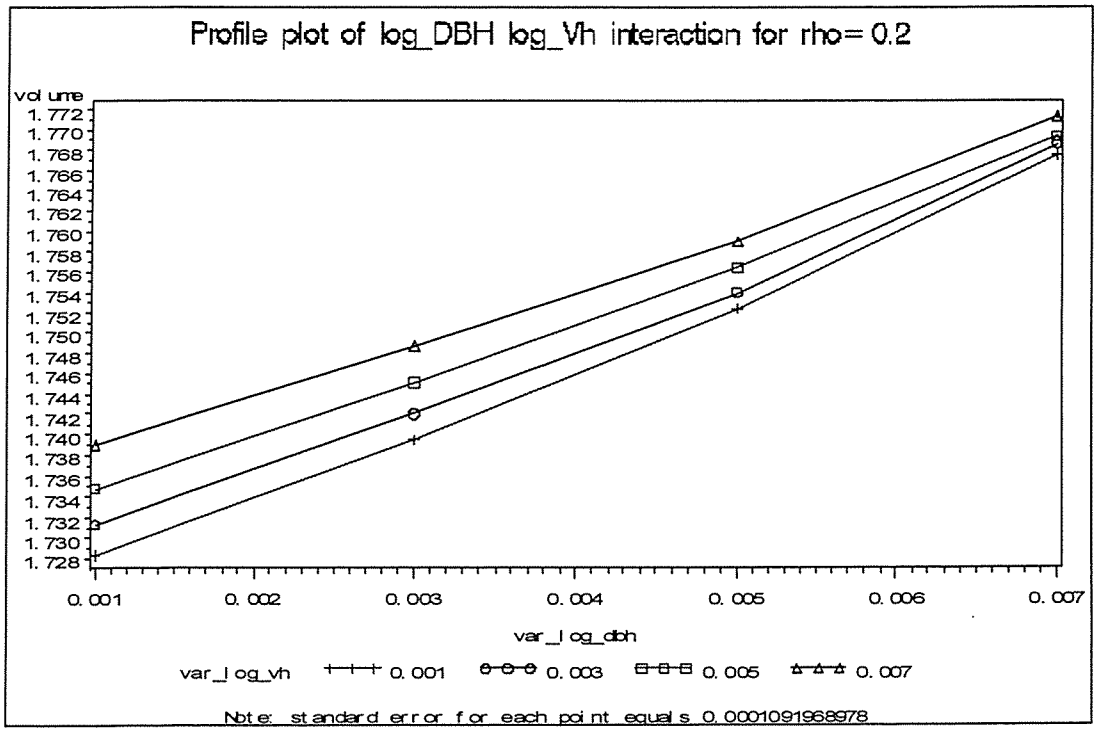

Figure 2.7.11: Profile plot of the Var(log(DBH)), Var(log(V_h)) interaction when rho=0.2

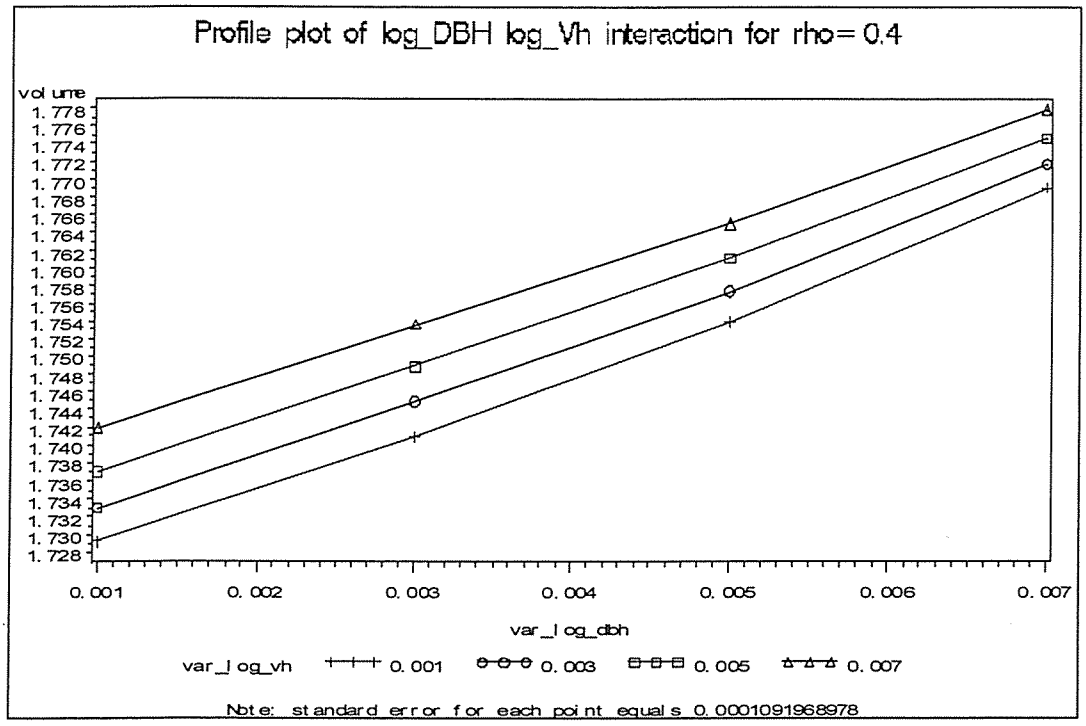

Figure 2.7.12: Profile plot of the Var(log(DBH)), Var(log(V_h)) interaction when rho=0.4

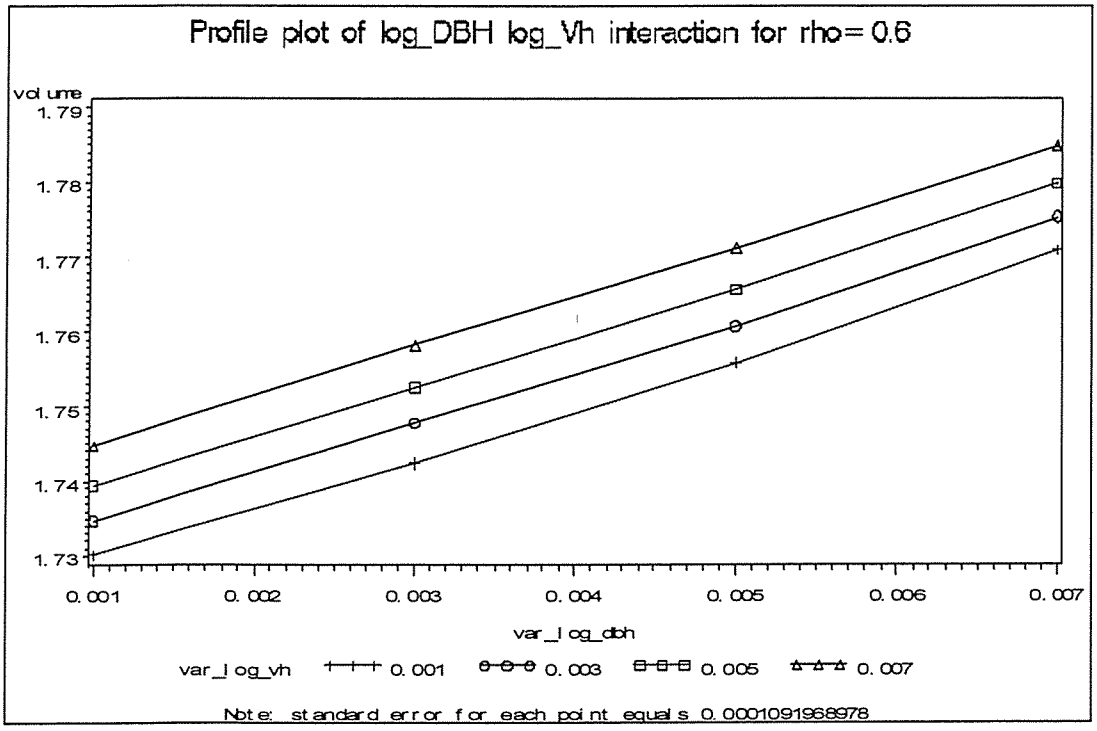

Figure 2.7.13: Profile plot of the Var(log(DBH)), Var(log(V_h)) interaction when rho=0.6

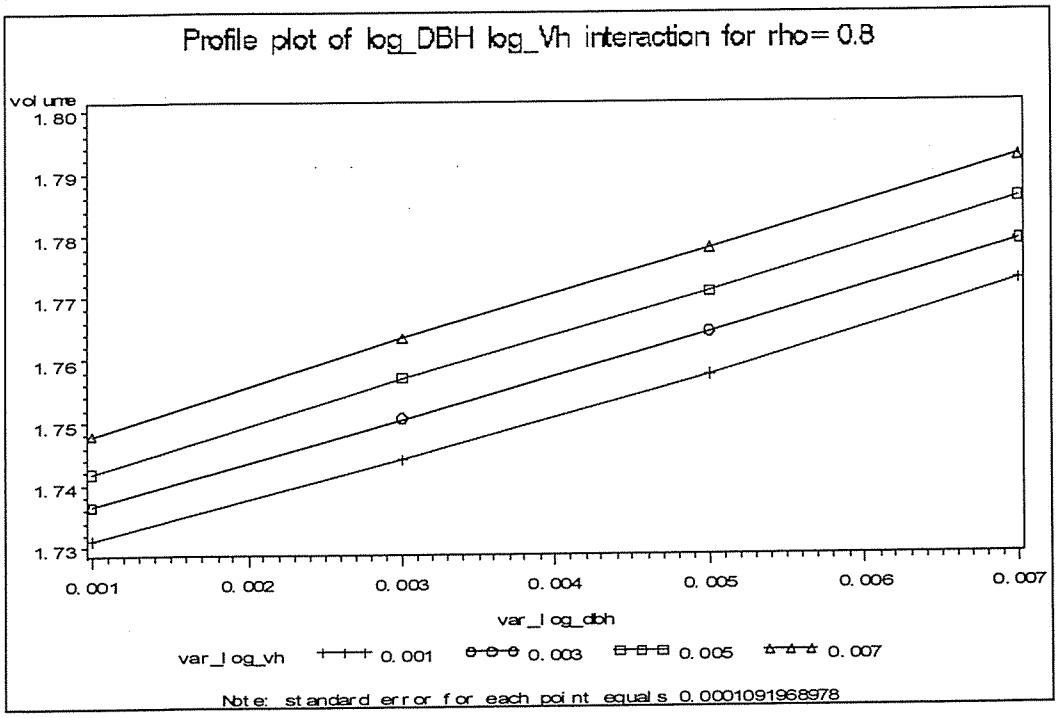

Figure 2.7.14: Profile plot of the Var(log(DBH)), Var(log(V_h)) interaction when rho=0.8

We can see from figures 2.7.10 – 2.7.14, that as the level of ρ increases the lines are becoming more parallel. This effect is precisely what was expected from the discussion earlier in this subsection.

3.0 Sampling Designs Supported by MARVL

3.1 Introduction to Sample Designs Supported by MARVL

This chapter discusses the sample designs and theory employed by MARVL to obtain estimates of timber volume for an area of interest. It considers the situation where there is no error associated with the volume (that is, the observed response variable has no measurement error and no equation error) as a preliminary to discussing the situation that applies to MARVL. The reader is directed to chapter 4 for a discussion concerning error associated with the volume and the effect this has on the analysis for a given design (that is, combining sampling design and measurement errors).

MARVL is based on a sampling method known as *area sampling*[15]. Area sampling is a special case of a single-stage cluster sampling design, and is employed when elementary units[16] of the population might be very costly to measure. In area sampling, instead of individual elements, such as trees, being studied clusters of individual elements are studied and a mean or total for a fixed area (spatially speaking) is used as the primary sampling unit. Thus the sample design can be considered a sample of fixed area's (or clusters) rather than individual elements.

Let us consider MARVL for the moment. In MARVL we are interested in individual trees, however obtaining the necessary information for a sample of individual trees can be very costly, in terms of time and money. For example, New Zealand forests are seldom free of under growth (such as black berries and gorse), which makes finding sample trees, at times, extremely difficult.

Another point to consider is how estimates of totals or means for a particular area of interest are calculated. If individual trees are used as the secondary sampling unit, estimates of total volume for an area of interest require the total number of trees be known in that area. However, this is seldom known in current forest inventories, due to such issues as thinning and disease. With these points in mind, area sampling (or single stage cluster sampling) allows the forest surveyor to obtain reliable estimates at a substantially smaller cost.

For example, let us consider the point made about the total number of trees needs to be known to obtain a measure of the total volume for an area of interest (given the primary sampling units are the individual trees). If area sampling is employed, the total area (instead of the total number of trees) must be known in order to obtain a measure of the total volume. Aerial photography in combination with photogrammetry allows the surveyor to obtain a relatively inexpensive estimate (in terms of time and money) of the total area of interest. This is the method currently employed by the MARVL inventory.

---

[15] Hansen [15] defines "area sampling" as the entire area in which the population is located is subdivided into smaller areas, and each elementary unit (individual trees in MARVLs case) is associated with one and only one such area. If we consider MARVL, then each tree can only belong to one fixed area plot (which is the case).

[16] The individuals whose characteristics are to be measured in the analysis (Hansen [15]), ie trees in MARVL's case.

In MARVL the elementary sampling unit is the individual tree. In survey sampling texts such as Sarndal [24], the type of design is based on the way the elementary sampling units are selected. If we consider the designs supported by MARVL, in terms of these criteria, then four designs are supported (Sarndal [24]),

1. Simple Cluster Sampling
2. Stratified Cluster Sampling,
3. Two-Phase Cluster Sampling with Ratio Estimation and,
4. Stratified Two-Phase Cluster Sampling with Ratio Estimation.

The following sections (3.2.1-3.2.3) discuss these four sampling designs above, and derive the formulas for the estimate of the mean or total and their respective variance.

3.2 Sampling Designs Supported by MARVL

This section outlines the four sample designs currently used by MARVL to obtain estimates of timber volume for a particular area of interest. It should be noted that the primary sampling unit (PSU) in all the following designs is a fixed area based plot. However, the elementary sampling units are the individual trees (that is, the unit we are interested in obtaining an estimate of volume for). Since the usual convention for defining a sample design is in terms of the elementary sampling unit, the following sections follow this convention.

3.2.1 Simple Cluster Sampling

Typically cluster sampling is employed when one of the two situations arise,

1. There exists no sampling frame that identifies each and every elementary unit and the cost of producing such a frame is too expensive due to budget constraints.
2. The population elementary units are scattered over a wide area. In this situation direct sampling of elementary units would result in a widely scattered sample. Hence the cost of obtaining such data would be very costly.

As we will see when these situations arise, simple cluster sampling allows one to obtain a hypothetical sampling frame of clusters, from which one can obtain estimates of means or totals and their respective variances perhaps at a smaller cost than other sampling designs.

Let us begin by firstly defining some notation. In this chapter we follow the notation outlined in Sarndal [24], chapter 4.2.

In single stage cluster sampling, the finite population $U = \{1,...,k,...,N\}$ is partitioned into N_I subpopulations, called clusters and are denoted by, $U_1,...,U_i,...,U_{N_I}$ and are symbolically represented as $U_I = \{1,...,i,...,N_I\}$.

U_I represents a population of clusters from which a sample of clusters s_I is selected. The clusters are the primary sampling unit. The number of population elements in the ith cluster U_i is denoted M_i.

This notation now allows us to define a single-stage cluster sample as follows (Sarndal [24]):

1. A probability sample s_I of clusters is drawn from U_I according to the probability design $p_I(\cdot)$. The size of s_I is denoted by n_I, for a fixed size design, or by n_{s_I}, for a variable size design (note, the number of sample elements in the ith sampled cluster s_i, is denoted m_i).
2. Every population element in the selected clusters is measured so that $m_i = M_i$ for all sampled clusters

Note that there is a generalisation of simple cluster sampling, called single-stage cluster sampling. A probability sample of clusters is taken (which for a simple cluster sample is a simple random sample of clusters). All elementary units (within each selected cluster) are measured for attributes of interest. Sometimes simple cluster sampling is called simple random cluster sampling (for example, Sarndal [24]).

Let us now direct our attention towards the simple cluster sample design employed by MARVL to obtain estimates of timber volume for a given area of interest. In MARVL the elementary units are the individual trees. However, no sampling frame exists due to such issues as thinning, disease and costs. Instead, fixed area based plots (or clusters) are used as the primary sampling unit for which we have a sampling frame (since the total area of interest is always known or can be estimated using remote sensing techniques).

The following gives an illustration of the technique employed by MARVL to conduct a simple cluster sample. Firstly, a random positioning of fixed-area plots are selected for a stand[17] of interest (which technically constitutes the survey population) . The usual method for selecting plots is to drop a 100cm by 100cm grid onto a scaled map of the stand of interest. Before the grid is dropped onto the map, random points are drawn onto the grid. The points on the grid then determine the random positioning of the plots to be sampled throughout the stand.

An illustration of a basic MARVL simple random cluster sample design is presented below in figure 3.2.1.

[17] A stand is defined as a group of trees (usually greater than a hectare) which was established at the same time. Usually the area of interest for a MARVL sample is the stand.

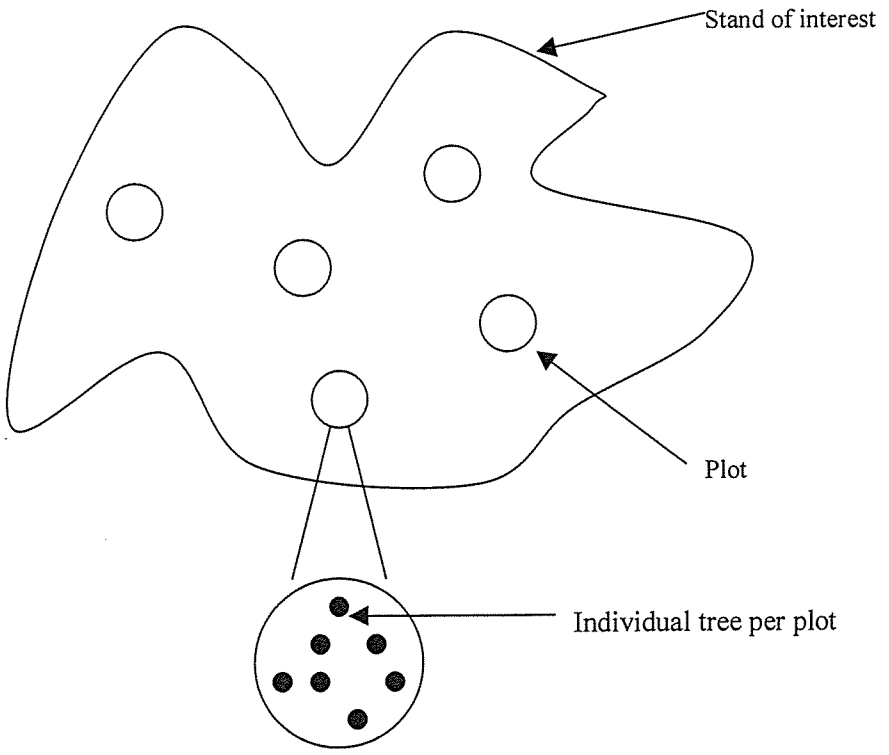

Figure 3.2.1: Example of a basic MARVL simple random cluster sample design

Let us now derive the formulas for the total and its variance of a simple random cluster sample by looking at the more general single stage cluster sample . Sarndal [24] derives this by firstly deriving the π estimator (also known as the Horwitz-Thompson estimator). Sarndal [24] defines the π estimator as follows,

$$\hat{t}_{\pi} = \sum_{S} \frac{y_i}{\pi_i} \quad \text{.....(3.2.1)}$$

Where

y_i = y-value for the i^{th} cluster

π_i = The probability of including the i^{th} cluster (also known as the inclusion probability)

With variance,

$$\hat{V}\left(\hat{t}_{\pi}\right) = \sum\sum_{S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \quad \text{.....(3.2.3)}$$

Where

π_{ij} = The probability that both the i^{th} and j^{th} cluster are included in the sample

The π estimator (equation 3.2.1) was developed by Horwitz and Thompson in 1952. The reason for defining the π estimator is that it can be used for any non-informative probability based statistical survey design for which a mean or total is to be estimated. For a definition of non-informative see Sarndal [24]. In addition to this, Thompson [28] shows that the π estimator will also yield unbiased estimates of the total and its variance (equation 3.2.2), no matter what survey design is used. Defining the π estimator forms the basis to derive the formulas of the total and its variance for any design.

Using the equations 3.2.1 and 3.2.2 for the π estimator total and its variance, we are now in a position to derive the simple cluster sampling total and its variance. As suggested above, this is how MARVL currently calculates its total and variance if this design is employed. Let us begin by considering the inclusion probabilities if a simple random sample is used to select the clusters (the first stage). In this case we select n_I clusters from a possible N_I clusters, hence the probability that the ith cluster will be selected in our sample, π_i, is

$$\pi_i = \frac{n_I}{N_I}$$

Hence the formula for the total is (given by Sarndal [24]),

$$\begin{aligned}\hat{t}_{\pi} &= \sum_{s_I} \frac{t_i}{\pi_i} \\ &= \sum_{s_I} \frac{N_I t_i}{n_I} \\ &= \frac{N_I}{n_I} \sum_{s_I} t_i \qquad \text{.....(3.2.3)}\end{aligned}$$

If we consider how equation 3.2.3 applies to MARVL, then we can see that the total timber volume for an area of interest is the total timber volume averaged over the sampled clusters $(\bar{t}_{s_I})$ scaled up to the number of clusters in the population of interest.

The variance estimate of this total is also given by Sarndal [24],

$$\hat{V}\left(\hat{t}_\pi\right) = N_I^2 \frac{1-f_I}{n_I} S_{ts_I}^2 \quad \text{.....(3.2.4)}$$

Where

$$S_{ts_I}^2 = \frac{1}{n_I - 1} \sum_{s_I} \left(t_i - \bar{t}_{s_I}\right)^2$$

$$\bar{t}_{s_I} = \frac{1}{n_I} \sum_{s_I} t_i$$

$f_I = \frac{n_I}{N_I}$ (The cluster sampling fraction)

Equation 3.2.4 is not unexpected, since it is the usual simple random sampling (without replacement) estimate of the variance, and the simple cluster sample is a simple random sample of clusters.

3.2.1.1 Efficiency of a Single-Stage Cluster Sample

Let us now return to the variance of a single-stage cluster sample (equation 3.2.4). Note that if the design is a fixed-sized design (as is the case with a simple random cluster sample), that is, if the probabilistic sample of clusters is a fixed-sized design then the estimated variance $\hat{V}\left(\hat{t}_\pi\right)$ can be rewritten as,

$$\hat{V}\left(\hat{t}_\pi\right) = -\frac{1}{2} \sum\sum_{s_I} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{t_i}{\pi_i} - \frac{t_j}{\pi_j}\right)^2$$

See for example, Sarndal [24] page 128.

This result leads to a number of important points about the efficiency of a single-stage cluster sample which Sarndal [24] discusses on page 128. Assuming that the π estimator (equation 3.2.1) is used and that the probability design is a fixed size design (such as in MARVLs case) then,

- From the alternative form of the variance (shown above) if all t_i/π_i are equal, then $V\left(t_\pi\right) = 0$. So therefore, if we can choose π_i proportional to the cluster totals, t_i, then cluster sampling will be highly efficient (that is yield an estimator with a small variance).
- An equal probability cluster sampling design (that is, one where all π_i are equal) is often a poor choice when the clusters are of different sizes (which is the case in MARVL). For such a design to be highly efficient, one must have $\bar{y}_{s_i}$ roughly proportional to n_i^{-1}.

3.2.2 Stratified Cluster Sampling.

Stratified random sampling is employed when homogeneous portions of the area of interest can be divided into non-overlapping groups. See figure 3.2.2 below

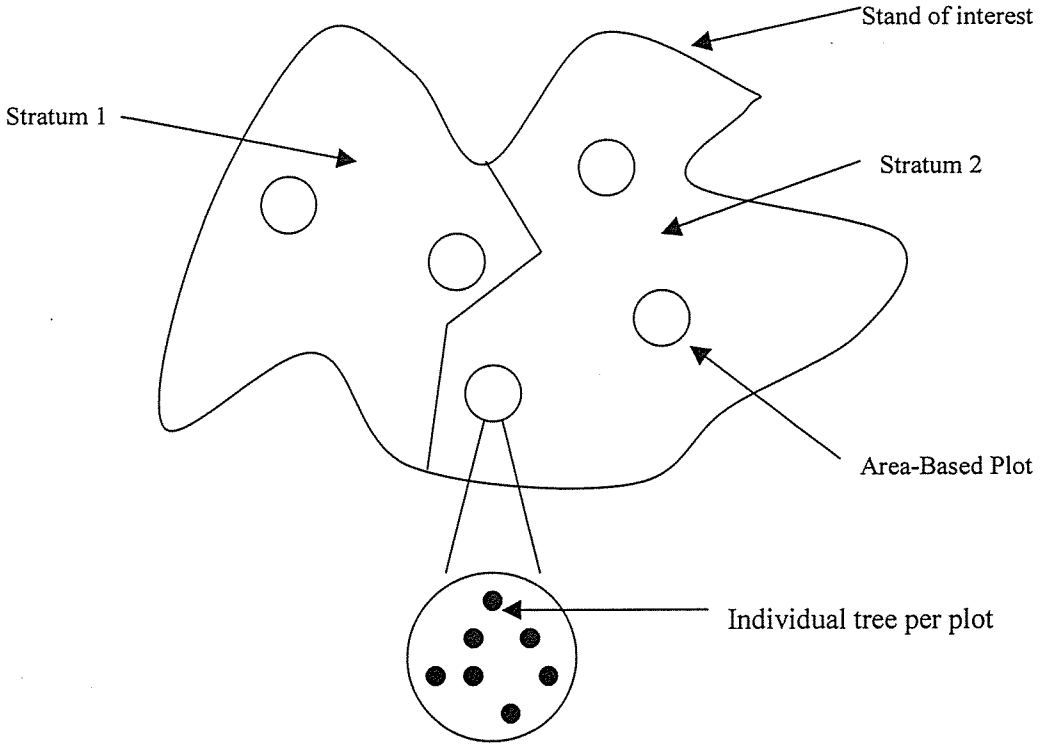

Figure 3.2.2: Example of a basic MARVL stratified cluster sample design

We can see in figure 3.2.2 that the stand, which is the population of interest, has been divided into two strata (stratum 1 and stratum 2). As stated above in stratification the population is divided into non-overlapping homogeneous sub-populations. When we consider the definition of homogeneous as it relates to MARVL, we mean areas of the forest or trees of interest that are similar in terms of size, age and quality etc. Natural stratification usually occurs, in terms of spatial distance from a specified point, that is trees that are close together are more similar than trees found further away, both in aspect of the ground and date of planting.

Within each stratum a simple cluster sample of fixed area plots are taken, hence in the diagram above we have two strata so we would take a simple cluster sample of fixed area-based plots from each of these strata (as described in section 3.2.1).

The reason why stratification is used is to reduce the variance in the estimators. As Thompson [28] states, because each of the selections in different strata are made independently, the variances of estimators for individual strata can be added together to obtain variances of the estimators for the whole population. Since the *within-stratum* (and not the *between-stratum*) variances enter into the variances of the estimators, the principal of stratification is to partition the population in such a way that the units within a stratum are as similar as possible. Further discussion about this fact will be presented later when investigating the formula for the variance of an estimator under stratified random sampling.

The total volume of timber for a particular area of interest is the major interest in MARVL. Under stratified random sampling Sarndal [24] derives the total and its variance of the π estimator (see equation 3.2.1),

$$\hat{t}_{\pi st} = \sum_{h=1}^{H} \hat{t}_{h\pi}$$

Where

$\hat{t}_{\pi st}$ = The π estimator of the total using a stratified sampling design

$\hat{t}_{h\pi}$ = The π estimator of the total in the hth stratum

H = The total number of strata

We can see from above that the total of the stratified π estimator is the sum of the individual totals from each stratum. This observation reflects the independence of the strata. The variance of this estimator is also of major importance in MARVL, Sarndal [24] defines the variance of this estimate as follows,

$$\hat{V}_{\pi st}\left(\hat{t}_{\pi st}\right) = \sum_{h=1}^{H} \hat{V}_h\left(\hat{t}_{h\pi}\right)$$

Where

$\hat{V}_{\pi st}\left(\hat{t}_{\pi st}\right)$ = The estimated variance of the stratified π estimator total

$\hat{V}_h\left(\hat{t}_{h\pi}\right)$ = The estimated variance of the π estimator total in the hth stratum

Again we can see that the variance is basically the sum of the variance for each strata, demonstrating the independence of the individual strata. Let us now consider the specific design employed by MARVL, namely that a simple cluster sample (see section 3.2.1) is taken within each of the strata. This leads to the following formula used to calculate total volume of timber under a stratified cluster design,

$$\hat{t}_{\pi st} = \sum_{h=1}^{H}\left(\frac{N_{hI}}{n_{hI}}\sum\nolimits_{s_{hI}} t_{hi}\right)$$

Where

N_{hI} = Total number of clusters in the hth strata

n_{hI} = Total number of sampled clusters in the hth strata

t_{hi} = Total of the ith cluster in the hth strata

s_{hI} = The sampled clusters in the hth strata

With estimated variance given by,

$$V_{\pi st}\left(\hat{t}_{\pi st}\right) = \sum_{h=1}^{H} N_{hI}^{2}\frac{1-f_{hI}}{n_{hI}} s_{hts_{hI}}^{2}$$

Where

$$s_{hts_{hI}}^{2} = \frac{1}{n_h - 1}\sum\nolimits_{s_{hI}}\left(t_{hi} - \bar{t}_{s_{hI}}\right)^2$$

$f_{hI} = \frac{n_{hI}}{N_{hI}}$ (the cluster sampling fraction for the hth startum)

t_{hi} = The total of the ith cluster in the hth stratum

$\bar{t}_{s_{hI}}$ = The mean total over all sampled clusters in the hth stratum

Let us now discuss the different components of variation and show how this sampling design yields estimators with smaller variances than a simple cluster sample. With a simple cluster sample we take a sample of clusters from the entire population, hence every cluster in the population has an equal opportunity of selection into the sample and hence the variance can be much larger than a stratified cluster sample. Consider the following diagrams (figure 3.2.3) to illustrate this point,

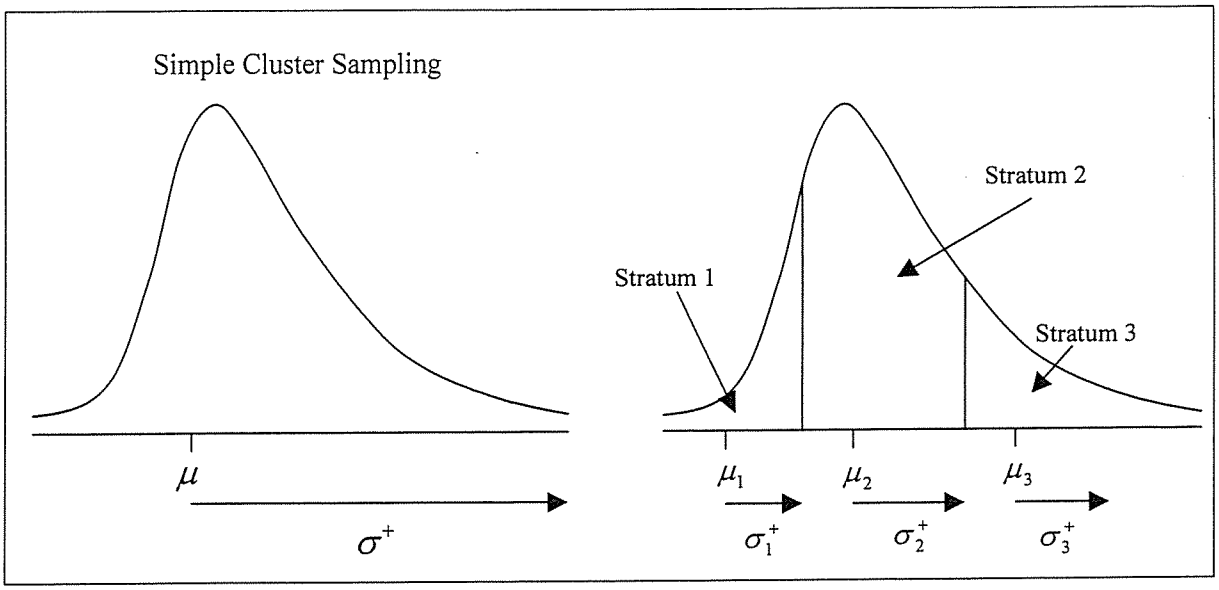

Figure 3.2.3: Graphical demonstration of the reduction in variance when a stratified cluster sample is employed

Figure 3.2.3 illustrates the possible reduction in variance using a stratified cluster design over a simple cluster sample. Consider the left hand diagram, if we take simple random samples of clusters from this distribution we obtain a mean μ and a standard deviation of σ. How ever, if we now consider the right hand diagram we see that if we can partition the 'assumed' distribution into three non-overlapping strata and take a simple cluster sample from each we obtain three different means and three corresponding standard deviations.

Due to the independent nature of the stratified cluster sample we can now simply add the standard deviations together to possibly obtain a smaller variance than the simple cluster sampling design. To give a theoretical argument to this diagram let us consider the design effect statistic, Lehtonen [17] for example defines the design effect as,

$$deff_{p(s)}(\hat{t}) = \frac{\hat{v}_{p(s)}(\hat{t})}{\hat{v}_{SRS}(\hat{t})}$$

The design effect is simply a ratio of the variance from a given complex design to that of a simple random sample. We can see that if the design effect is less than one then the variance of the complex design is smaller than that of a simple random sample. Thompson [28] states that if the *principle of stratification*[18] is used to partition the population then,

$$deff_{stratified\_random\_sample}(\bar{y}) \leq 1$$

[18] Principle of stratification states that the population is partitioned in such a way that the units within a stratum are as similar as possible (Thompson [28])

Hence the variance of the stratified cluster sample is smaller than that of a simple cluster sample due to the reduction in variation.

When ever possible a stratified cluster sample should be used instead of a simple cluster sample. In MARVL stratification is often employed with the aid of remote sensing. Where aerial photography in combination with photogrammetry are used to obtain the different strata.

3.2.3 Two-Phase Cluster Sampling with Ratio Estimation

Typically, double sampling is employed by MARVL to obtain estimates of total volume. Double sampling is a special form of a ratio estimator where auxiliary information (linearly related to the variable of interest) is used to obtain estimates with increased efficiency and precision.

The double sampling procedure for MARVL involves two stages. Firstly "secondary" plots or clusters are randomly selected (using the method outlined in section 3.2.1). Measurement of diameter at breast height of all trees in all secondary clusters are obtained (this is considered the auxiliary information). Secondly "primary" plots or clusters are randomly selected. Primary clusters are a random sub-sample of the secondary clusters, where all trees in these plots are fully cruised[19].

An illustration of a basic MARVL double sampling design discussed above is presented below in figure 3.2.4.

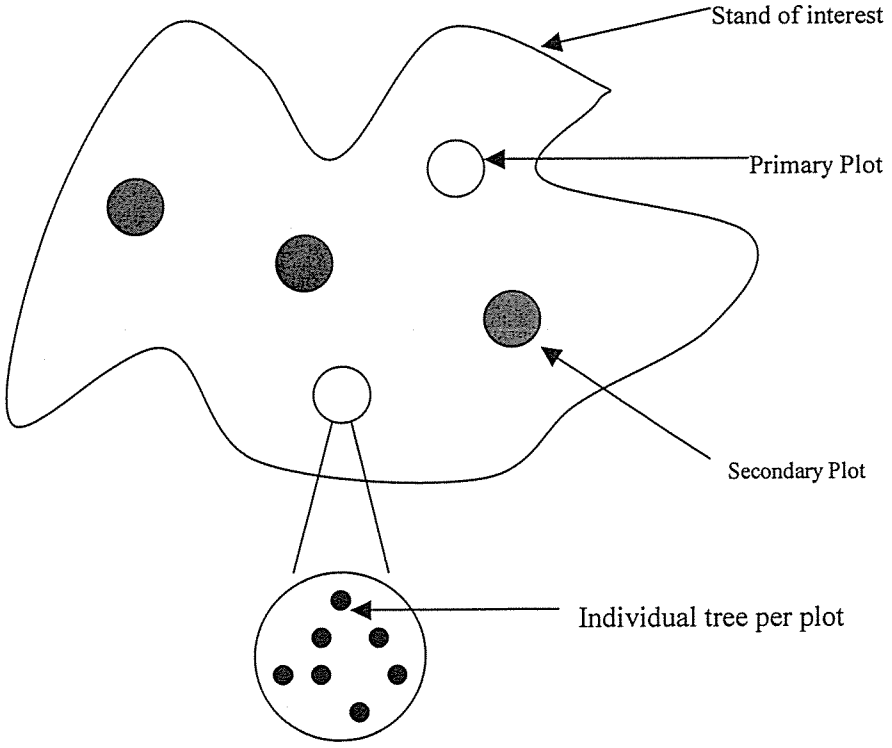

Figure 3.2.4: Example of a basic MARVL double sampling design

Volume equations are used to obtain volumes for individual trees in the primary plots. In this section, we assume that there is no error in the estimates obtained from these volume equations, the reader is referred to chapter 4 for a full discussion on the effect of these errors on the total sample survey error associated with MARVL.

[19] Term given to obtaining measures such as height and quality of an individual tree.

These volume estimates are then used in combination with the diameter at breast height information recorded from the secondary plots to calculate an estimate of the mean volume, denoted $(\bar{y}_{Double\_Sampling})$,

$$\bar{y}_{Double\_Sampling} = \frac{\bar{y}}{\bar{x}}\bar{x}' = R\bar{x}'$$

Where

$\bar{y}$ = The mean volume for the primary plots

$\bar{x}$ = The mean diameter at breast height for the primary plots

$\bar{x}'$ = The mean diameter at breast height for the secondary plots

As mentioned earlier, double sampling is a special case of a ratio estimator. Similarly, a ratio estimator is a special case of a regression estimator (that is, a ratio estimator is a regression estimator with no intercept). Hence we can re-write the double sampling estimator $(\bar{y}_{Double\_Sampling})$, in the form of a weighted linear model (where the weights are the inverse of the selection probabilities) and use standard regression theory (Draper [11]) to obtain estimates of means and their variances,

$$y = X\hat{\beta}$$

Where

y = column vector of observations for the variable of interest

X = The design matrix (note this is equivalent to $\bar{x}'$ in the formula above)

$\hat{\beta} = \left(X'\Pi^{-1}X\right)^{-1}X'\Pi^{-1}y$ (note this is equivalent to $\hat{R}$ in the formula above)

$\Pi^{-1} = diag\left(\pi_1^{-1}, \pi_2^{-1}, ..., \pi_{n'}^{-1}\right)$

π_i = the selection probability of the ith cluster ($\frac{1}{X_i}$ in the equation above)

Let us now consider the general theory of two-phase sampling design[20] with regression estimation. Sarndal [24] chapter 9, defines two-phase sampling (or double sampling) as follows,

a) In the first phase, select a rather large sample of elements (denoted s_a) by a probability based sampling design $p_a(\cdot)$ ($p_a(\cdot)$, at the tree level, is a simple cluster sample in MARVL's case). For the elements in s_a gather inexpensive information on one or more auxiliary variables.
b) With the aid of the auxiliary information collected in the first phase, select a second-phase sample s from s_a by the design $p(\cdot \mid s_a)$ ($p(\cdot \mid s_a)$, at the tree level (in MARVL's case) is a simple random sample of the clusters from s_a). This s is a sub-sample. The study variable y (volume of the individual trees in MARVL's case) is then observed for all the trees in each selected second phase sample.

[20] Also known as Double Sampling

The first task is to find an unbiased estimator of the population total, $t = \sum_U y_k$ (that is the sum of the y_k's over the population U). Sardnal [10] page 345, suggests the π estimator[21] (equation 3.2.1) as a possibility,

$$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$$

Where

$\hat{t}_\pi$ = The π estimator (or Horvitz - Thompson estimator)

y_k = The kth observation of the variable of interest

$\pi_k = \Pr(k \in s)$ (the inclusion probability of the kth element)

However, this estimator requires that the π_k can be calculated for all k, but this is not always possible in practice since,

$$\pi_k = \sum_{k \in s} \sum_{s \subset s_a} p_a(s_a) p(s \mid s_a)$$

$$\pi_k = \sum_{k \in s_a} p_a(s_a) \pi_{k|s_a} \quad \text{Sarndal [10]}$$

Where

$p_a(s_a)$ = The probability of a sample being selected in the first phase

$p(s \mid s_a)$ = The probability of a sample being selected in the second phase given it was selected in the first phase

$\pi_{k|s_a}$ = Selection probability of the kth element given it was selected in the first phase

From the decomposition of the selection probability of the kth element, π_k above, we can see that we must know $p_a(s_a)$ (which we ordinarily do in two phase sampling as this is simply the probability of a cluster being selected in the first phase in MARVL's case) and $\pi_{k|s_a}$ for all s_a (which we ordinarily do not).

As we have seen from above, Sarndal [24] states that in practice the Horvitz-Thompson estimator cannot be calculated (since $\pi_{k|s_a}$ is not always possible to calculate in practice). Instead Sarndal [24] page 347, forms a new estimator, called the π^* estimator, which can be calculated in practice when a two-phase sample design is employed. Sarndal [24] page 347 defines the π^* estimator as follows,

[21] Also known as the Horvitz-Thompson Estimator, [28]

$$\hat{t}_{\pi^*} = \sum_s \frac{y_k}{\pi_k^*} \quad \text{.....(3.2.5)}$$

Where

$\hat{t}_{\pi^*}$ = The π^* estimator of the total

$\pi_k^* = \pi_{ak}\pi_{k|s_a}$ (note that $\pi_{k|s_a}$ is known for all k in a particular sample s_a)

An unbiased estimator of the variance of the π^* estimator is derived by Sarndal [24] page 348 and is presented below,

$$\hat{V}\left(\hat{t}_{\pi^*}\right) = \sum\sum_s \frac{\Delta_{akl}}{\pi_{kl}^*}\frac{y_k}{\pi_{k|s_a}}\frac{y_l}{\pi_{l|s_a}} + \sum\sum_s \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}}\frac{y_k}{\pi_k^*}\frac{y_l}{\pi_l^*} \quad \text{.....(3.2.6)}$$

Where

$\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$

$\Delta_{kl|s_a} = \pi_{kl|s_a} - \pi_{k|s_a}\pi_{l|s_a}$

$\pi_{kl}^* = \pi_{akl} - \pi_{kl|s_a}$

The discussion thus far has only considered measurement of the variable of interest, y_k. However, in MARVL we have auxiliary information, that is linearly related to the variable of interest, obtained from extra sample plots (from the first phase), the following discussion will demonstrate the reduction of variance using this extra information through the use of regression estimators.

Firstly we need to define a number of vectors,

a) Let $\boldsymbol{x}_k$ be a vector of J auxiliary values available for all $k \in s_a$

b) Let $\boldsymbol{x}_{1k}$ be a vector of J_1 auxiliary values available for all k in the population U.

Furthermore, we assume that $\boldsymbol{x}_k$ contains variable values known beforehand for all of U as well as variable values known for $k \in s_a$. Hence we can re-write $\boldsymbol{x}_k$ as $\boldsymbol{x}_k = \left(\boldsymbol{x}_{1k}', \boldsymbol{x}_{2k}'\right)'$, where $\boldsymbol{x}_{1k}$ is the vector of J_1 values known for all of U, and $\boldsymbol{x}_{2k}$ is the vector of $J_2 = J - J_1$ values recorded by observation of elements k in the first phase of the sample design.

Sarndal [24] demonstrates that using the vectors defined above through linear regression models it is possible to obtain an estimator of t_{π^*} with a smaller variance. To show this, Sarndal [24] page 356 firstly derives the theory of the difference estimator. The difference estimator acts as a natural stepping stone to two-phase regression estimators.

Firstly, lets suppose the following relationships hold,

$$y_k \approx \boldsymbol{x}_k' \boldsymbol{A} = y_k^0$$

and

$$y_{1k} \approx \boldsymbol{x}_{1k}' \boldsymbol{A}_1 = y_{1k}^0$$

where $\boldsymbol{A}$ and $\boldsymbol{A}_1$ are known vectors. From these equations we can form the following differences,

$$D_k = y_k - y_k^0$$

and

$$D_{1k} = y_k - y_{1k}^0$$

Sarndal [24] page 357 illustrates how the difference estimator of the total and its variance are derived, here we just state the total and variance of the difference estimator,

$$\hat{t}_{dif} = \sum_U y_{1k}^0 + \sum_{s_a} \frac{y_k^0 - y_{1k}^0}{\pi_{ak}} + \sum_s \frac{y_k - y_k^0}{\pi_k^*}$$

The unbiased estimate of the variance of the difference estimator is presented below (Sarndal [24]),

$$\hat{V}\left(\hat{t}_{dif}\right) = \sum_s \sum \frac{\Delta_{akl}}{\pi_{kl}^*} \frac{D_{1k}}{\pi_{ak}} \frac{D_{1l}}{\pi_{al}} + \sum_s \sum \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}} \frac{D_k}{\pi_k^*} \frac{D_l}{\pi_l^*}$$

If we consider the auxiliary value vector of MARVL for the moment, we see that $\boldsymbol{x}_{1k}$ is not known, since in MARVL a double sampling design is employed. Instead $\boldsymbol{x}_{2k}$ is known which means that $\boldsymbol{x}_k$ becomes $\boldsymbol{x}_{2k}$. Sarndal [24] page 358, shows how this leads to a simplification of the difference estimator presented above,

$$\hat{t}_{dif2} = \sum_{s_a} \frac{y_k^0}{\pi_{ak}} + \sum_s \frac{y_k - y_k^0}{\pi_k^*}$$

with estimated variance,

$$\hat{V}\left(\hat{t}_{dif2}\right) = \sum_s \sum \frac{\Delta_{akl}}{\pi_{kl}^*} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \sum_s \sum \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}} \frac{D_k}{\pi_k^*} \frac{D_l}{\pi_l^*}$$

As mentioned before, the difference estimator acts as a natural stepping stone to regression estimators for two-phase sampling. Firstly, the difference estimator $\hat{t}_{dif}$ shown above suggests the regression estimator,

$$\hat{t}_r = \sum_U \hat{y}_{1k} + \sum_{s_a} \frac{\hat{y}_k - \hat{y}_{1k}}{\pi_{ak}} + \sum_s \frac{y_k - \hat{y}_k}{\pi_k^*}$$

where $\hat{y}_k$ and $\hat{y}_{1k}$ are predicted values obtained from appropriate regression fits. Sarndal [24] page 362 derives the regression models and shows the standard least squares fits for $\hat{y}_k$ and $\hat{y}_{1k}$. The theory will not be shown here as most standard general linear model books (see Dobson [10]) cover these results. Instead we will investigate the regression estimator when $\boldsymbol{x}_{2k}$ is known and $\boldsymbol{x}_{1k}$ is unknown which means that $\boldsymbol{x}_k$ becomes $\boldsymbol{x}_{2k}$ (this is what happens in MARVLs case since measurements on predictor variables (namely, DBH) are only obtained from the secondary clusters). The total of the regression estimator is presented below (Sarndal [24], page 364),

$$\hat{t}_{r2} = \sum_{s_a} \frac{\hat{y}_k}{\pi_{ak}} + \sum_s \frac{y_k - \hat{y}_k}{\pi_k^*}$$

Where

$\hat{t}_{r2}$ = The regression estimate of the total for a two phase design

$\hat{y}_k = \boldsymbol{x}_{2k}\hat{\boldsymbol{\beta}}_s$ (that is the fit from the linear regression model)

The variance of the regression estimator of the total for a two phase sampling design is,

$$\hat{V}(\hat{t}_{r2}) = \sum_S \sum \frac{\Delta_{akl}}{\pi_{kl}^*} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \sum_s \sum \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}} \frac{e_{ks}}{\pi_k^*} \frac{e_{ls}}{\pi_l^*} g_{ks} g_{ls}$$

Where

$$e_{ks} = y_k - \hat{y}_k$$

$$g_{ks} = 1 + \left(\sum_{s_a} \frac{\boldsymbol{x}_k}{\pi_{ak}} - \sum_s \frac{\boldsymbol{x}_k}{\pi_k^*} \right)' \left(\sum_s \frac{\boldsymbol{x}_k \boldsymbol{x}_k'}{\sigma_k^2 \pi_k^*} \right)^{-1} \frac{\boldsymbol{x}_k}{\sigma_k^2}$$

The theory developed above is very general. Let us consider the method used in MARVL to obtain estimates of volume. Firstly, a ratio estimator is used, this simplifies the above equations considerably, secondly only a single predictor variable (namely, diameter at breast height) is used to form the regression equation, with these points in mind, Sarndal [24] page 365, shows the simplification of the above formulas. Firstly the regression model becomes,

$$E_\xi(y_k) = \beta x_k$$

$$V_\xi(y_k) = \sigma^2 x_k$$

where,

$$\hat{y}_k = \hat{\beta}_s x_k$$

and,

$$\hat{\beta}_s = \frac{\sum_s \frac{y_k}{\pi_k^*}}{\sum_s \frac{x_k}{\pi_k^*}}$$

The residuals of this model are given by,

$$e_{ks} = y_k - \hat{\beta}_s x_k$$

and the g weights simplify to,

$$g_{ks} = \frac{\sum_{s_a} \frac{x_k}{\pi_{ak}}}{\sum_s \frac{x_k}{\pi_k^*}}$$

Hence the estimator of the total, $\hat{t}_{r2}$ simplifies to,

$$\hat{t}_{r2} = \left(\sum_{s_a} \frac{x_k}{\pi_{ak}}\right) \frac{\sum_s \frac{y_k}{\pi_k^*}}{\sum_s \frac{x_k}{\pi_k^*}} = \left(\sum_{s_a} \frac{x_k}{\pi_{ak}}\right) \hat{\beta}_s$$

and the variance of this estimator is the same as for the un-simplified variance, however, with g weights and residuals as defined above.

Sarndal [24] page 365 shows how the theory outlined above leads to and estimate of the total and its variance when both phases are a simple cluster sample. The total is,

$$\hat{t} = N\bar{x}_{s_a}\left(\bar{y}_s / \bar{x}_s\right)$$

and the variance estimator of this total is,

$$\hat{V}\left(\hat{t}\right) = N^2 \frac{s_{ys}^2}{n_a} + N^2 \left(\frac{\bar{x}_{s_a}}{\bar{x}_s}\right)^2 \frac{s_{es}^2}{n}$$

Where

$$s_{ys}^2 = \frac{1}{n-1} \sum_s \left(y_k - \bar{y}_s\right) \text{(Between sample cluster variance)}$$

$$s_{es}^2 = \frac{1}{n-1} \sum_s \left(y_k - \frac{\bar{y}_s}{\bar{x}_s} x_k\right)^2 \text{ (Within sample cluster variance)}$$

It should be noted that similar formulas can be derived for stratified two-phase sampling with ratio estimation.

3.2.3.1 Comparing the current MARVL Design to this two-phase design.

The current design used by MARVL treats the clusters as the elementary sampling unit and uses standard simple random sampling theory (see Cochran [8]) to derive estimates of volume from these plots by finding the total of the individual sample plots and using this as the variable of interest.

As we can see from the estimate of the variance for a two-phase design with ratio estimation derived by Sarndal [24], the formula involves two parts, namely the variation between clusters and the variation within clusters. It should be noted that the variation between clusters drives the total estimated variance, however if we consider the within cluster variance formula (s_{es}^2), the information required to calculate this statistic is available to us in a MARVL inventory, and hence should be used to obtain correct estimates of variance. The current MARVL procedure does not include the second term in its variance estimates and consequently underestimates variance for the two phase design (that is for double sampling).

3.3 General comment on sample design issues

Proper sample design for forest inventory is critical. It is not simply a question of unbiasedness. Inefficient designs cost no less, but lead to what may be considerably more inaccurate estimates of timber volume. It is also imperative that the correct variance and estimated variance formulae are used, else designs and estimates of timber volume may be perceived as being more accurate than they actually are.

4.0 Inventory design (MARVL), incorporating measurement error and sampling design issues.

4.1 Introduction

In chapter 2 we discussed measurement error models and the effect these have on obtaining estimates of parameters compared to standard regression techniques. Chapter 3 discussed sample designs and relevant issues relating to MARVL assuming that there is no error in the variable of interest (in MARVL's case, this is the volume of the individual trees). However, MARVL currently uses non-linear regression techniques to obtain estimates for the volumes of individual trees. This suggests that the volumes used in the formulas in chapter 3, are not fixed but actually contain an associated measurement error. Not allowing for this error can not only lead to incorrect estimates of total volume but also to errors in their respective variances.

Furthermore, chapter 2 illustrated that the volume equations used in MARVL are in fact not regression equations, but instead require measurement error models. This is because the predictor variables used in the equations also contain errors. Again not allowing for this additional error leads to incorrect estimates of volume and their respective variances.

In this chapter we attempt to incorporate the ideas of the previous two chapters and discuss the effect these have on obtaining estimates of volume both in a statistical and a practical sense. The chapter follows the ideas detailed in chapter 16 (measurement errors) of Sarndal [24].

The basic idea behind Sarndals' [24] theory is that the measurements (for the variable of interest – volume in MARVLs case) are modelled as random variables, instead of a fixed variable (which MARVL currently assumes). This allows the use of standard statistical tools to be used to evaluate the effect of the measurement errors on total survey error.

4.2 Measurement Model Theory

In order to combine the measurement error model estimate of y_k (since the observed variable of interest contains an error) and the sample design we must formulate a statistical model for the measurement errors in the context of a sample from a finite population. As stated earlier, the basic idea is that the measurements and subsequently the measurement errors are modeled as random variables. This allows the use of standard statistical tools to be used to evaluate the effect of measurement error on the accuracy of the usual estimators.

As discussed in section 2.2 the measurement model (or in MARVL's case a measurement error model) is given by,

$$Volume\_under\_Bark = e^{\beta_1} d^{\beta_2} \left[h^2/(h-1.4)\right]^{\beta_3} + \varepsilon$$

Where

d = Diameter at breast height (over bark)

h = Height of the tree (in meters)

It should be noted that this measurement model is non-linear in the parameters. This means that non-linear measurement error model theory (see section 2.3) needs to be used to obtain estimates of y_k (volume under bark). However, it was shown in section 2.2, that by using the logarithmic function on the above model, a linear approximation could be achieved,

$$\log(VUB) = \log\left(e^{\beta_1} d^{\beta_2} V_h^{\beta_3} + \varepsilon\right)$$

$$\Rightarrow \log(VUB) = \log\left(e^{\beta_1} d^{\beta_2} V_h^{\beta_3} (1+\varepsilon')\right)$$

$$\Rightarrow \log(VUB) = \log\left(e^{\beta_1} d^{\beta_2} V_h^{\beta_3}\right) + \log(1+\varepsilon')$$

$$\Rightarrow \log(VUB) = \beta_1 + \beta_2 \log(d) + \beta_3 \log(V_h) + e$$

where

VUB = Volume under Bark

d = diameter at breast height (Over Bark)

h = Height of the tree

$$V_h = \frac{h^2}{(h-1.4)}$$

$$\varepsilon' = \frac{\varepsilon}{e^{\beta_1} d^{\beta_2} V_h^{\beta_3}}$$

$$e = \left[\varepsilon' + \frac{\varepsilon'^2}{2!} + \frac{\varepsilon'^3}{3!} + \ldots\right] = \log(1+\varepsilon')$$

The model above allows us to obtain estimates of y_k, that vary from the true value θ_k where y_k is the term given to the estimated volume (from the volume equation presented above) of the kth tree and θ_k is the term given to the (conceptual) actual volume of the kth tree. Sarndal [24] shows the effect measurement error has on the standard estimators of totals and more importantly, on the standard design variances (see chapter 3).

Sarndal [24] page 606, views the survey with measurement error as a two-stage process, with each stage contributing randomness,

1. The sample selection, which results in a selected sample s. Stochastic structure is given by the sampling design, $p(\cdot)$ (which leads to the error incurred by the sampling design).
2. The measurement procedure, which generates an observed value, y_k for each $k \in s$. Stochastic structure is given by the measurement model, *m.* (which leads to the error incurred from estimating the volume, y_k).

Let us now define a few terms that will be used in this section (Sarndal [24], page 606). Let,

$$E_{pm}(\cdot) = E_p[E_m(\cdot \mid s)]$$

Where

$E_{pm}(\cdot)$ = The expection with respect to the sample design and measurement model jointly

$E_m(\cdot \mid s)$ = The conditional expectation with respect to the measurement model, for a given sample, *s*.

$E_p(\cdot)$ = The expectation with respect to the sample design $p(\cdot)$.

and,

$$V_{pm}(\cdot) = E_p[V_m(\cdot \mid s)] + V_p[E_m(\cdot \mid s)]$$

Where

$V_{pm}(\cdot)$ = The variance with respect to $p(\cdot)$ and m jointly

$V_m(\cdot \mid s)$ = The conditional variance with respect to the model m, and the sample s

$V_p(\cdot)$ = The variance with respect to the sample design, $p(\cdot)$.

Note that both the above definitions of $E_{pm}(\cdot)$ and $V_{pm}(\cdot)$ can be justified by the usual theorems on conditional expectations (see Bain [2]).

Further, let us define the measurement model (which Sarndal [24] page 605, calls the simple measurement model), *m*, moments for elements *k* and *l* belonging to the same sample *s* as,

$$\mu_k = E_m(y_k \mid s)$$
$$\sigma_k^2 = V_m(y_k \mid s)$$
$$\sigma_{kl} = C_m(y_k, y_l \mid s)$$

The model above states that, for any given sample *s*, the measurement y_k on population element *k* has the mean μ_k and the variance σ_k^2, and between y_k and y_l there is the covariance σ_{kl}.

Sarndal [24] page 607, defines the simple measurement model *m*, (in terms of long run frequency) as follows; There is a given probability sample *s* and a given measurement procedure that generates an observed value for each element $k \in s$. Suppose measurements could be independently repeated many times on the same sample, *s*, thus generating a long series of measurements on each element $k \in s$. The observed y-values for a particular element *k* would not necessarily be the same in all repetitions, but would vary in a random fashion, around a "long run" mean value μ_k and a "long run" variance σ_k^2.

4.3 Impact of Measurement Error on Total Survey Error

Let us now consider the mean square error of the π estimator (equation 3.2.1) and examine what effect measurement error has on its accuracy. Sarndal [24] page 608, gives the mean square error of $\hat{t}_\pi$ under the sample design and the measurement model as,

$$MSE_{pm}(\hat{t}_\pi) = V_{pm}(\hat{t}_\pi) + [B_{pm}(\hat{t}_\pi)]^2$$

Where

$$V_{pm}(\hat{t}_\pi) = E_{pm}\left\{[\hat{t}_\pi - E_{pm}(\hat{t}_\pi)]^2\right\}$$

$$B_{pm}(\hat{t}_\pi) = E_{pm}(\hat{t}_\pi) - t_\theta$$

To give a simpler formula for the mean square error, Sarndal [24] page 609 decomposes the mean square estimator to give the following results,

$$MSE_{pm}(\hat{t}_\pi) = V_{pm}(\hat{t}_\pi) + B^2$$

Where

$$B = \sum_U (\mu_k - \theta_k) \text{ (The measurement bias)}$$

$$V_{pm}(\hat{t}_\pi) = V_1 + V_2 = V_{11} + V_{12} + V_2$$

Where,

$$V_{11} = \sum_U \sigma_k^2 / \pi_k ,$$

is the simple measurement variance;

$$V_{12} = \sum_{k \neq l} \sum_U (\pi_{kl} / \pi_k \pi_l) \sigma_{kl} ,$$

is the correlated measurement variance; and

$$V_2 = \sum \sum_U \Delta_{kl} \frac{\mu_k}{\pi_k} \frac{\mu_l}{\pi_l},$$

is the sampling variance (note: this is the usual variance formula for the Horvitz-Thompson estimator). Note that π_k, π_l and Δ_{kl} have been defined in section 3.2.1.

Currently, both parts $(V_{11} \text{ and } V_{12})$ of V_1, depend on the measurement model (through σ_k^2 and σ_{kl}) and on the sampling design (through the inclusion probabilities). With this in mind, it is also possible to decompose V_1 in terms of two components, one which does not depend on the sampling design (V_{1cen}), and one which does (V_{1sam}). Sarndal [24] page 610 gives the second decomposition of V_1,

$$V_1 = V_{1cen} + V_{1sam}$$

Where

$$V_{1cen} = \sum \sum_U \sigma_{kl}$$

$$V_{1sam} = \sum \sum_U \Delta_{kl} \sigma_{kl} / (\pi_k \pi_l)$$

If we consider this decomposition in terms of the average model moments of the simple measurement model (Sarndal [24]) we can define,

$$\mu_m = \sum_U \frac{\mu_k}{N},$$

$$\sigma_m{}^2 = \sum_U \frac{\sigma_k^2}{N},$$

$$\rho_m = \sum_{k \neq l} \sum_U \sigma_{kl} / \left[N(N-1)\sigma_m{}^2 \right]$$ (that is, the correlation between elements under the model m)

This leads to the following decomposition of the measurement variance under the sampling design $p(\cdot)$ and the simple measurement model m,

$$V_1 = V_{1cen} + V_{1sam}$$

Where

$$V_{1cen} = N\left[1 + (N-1)\rho_m\right]\sigma_m{}^2$$

$$V_{1sam} = \sum \sum_U \Delta_{kl}\sigma_{kl} / (\pi_k \pi_l)$$

Let us now consider the formulas for the measurement variance in the decomposed form above (that is, $V_1 = V_{1cen} + V_{1sam}$). We can see that when measurements for all pairs of elements are correlated this will lead to an increased measurement variance (since $\rho_m > 0$).

Let us consider MARVL for the moment. Clusters are used as the primary sampling unit. These clusters contain trees which are similar (in terms of timber volume), since trees close together (spatially speaking) having been planted at the same time exhibit similar volume estimates. If measurement error depends on the size of the tree, or each field crew member measures the trees sampled continuously over or under then $\rho_m > 0$, which means that the total sampling error will be increased. The more correlated the measurements are, the larger this increase.

4.3.1 An Example: Simple Cluster Sampling with Measurement Error

To illustrate the results above and hence the effect measurement error has in obtaining the y_k, let us consider a simple cluster sample (note: the simple cluster sample design is outlined in section 3.2.1, however now we derive the variance formula when measurement error occurs). With μ_m, $\sigma_m{}^2$ and ρ_m as defined above in terms of clusters, the simple measurement variance is,

$$\hat{V}_{11} = \frac{N^2 {s_{mI}}^2}{n_I} \frac{(N-1)}{N}$$

Where

n_I = The number of sampled clusters

N = The total number of clusters in the population

$$s_{mI}^2 = \sum_{s_I} \frac{s_k^2}{n}$$

and the correlated measurement variance is

$$\hat{V}_{12} = \frac{N^2 (n_I - 1)\rho_{mI} {s_{mI}}^2}{n_I} \frac{(N-1)}{N}$$

Where

$$\rho_{mI} = \sum_{k=l} \sum_{s_I} s_{kl} / \left[n(n-1) s_{mI}^2 \right]$$

for a total measurement variance of

$$\hat{V}_1 = \hat{V}_{11} + \hat{V}_{12} = \frac{N^2 [1 + (n_I - 1)\rho_{mI}] {s_{mI}}^2}{n_I} \frac{(N-1)}{N}$$

Let us consider the simple measurement variance as the sample size n_I, increases (and ${\sigma_m}^2$ and ρ_m remain constant). As n_I increases, the simple measurement variance, $\hat{V}_{11}$ gets smaller, however reducing the sample size has no effect on the correlated measurement variance, $\hat{V}_{12}$. As Sarndal [24] page 612, points out, as we can see correlated measurements can do great harm to the precision of the survey estimates since the increase due to correlated measurements is measured by the factor $(n_I - 1)\rho_{mI}$.

Since every unit in each sampled cluster is sampled, the sampling variance component for a simple cluster sample (see section 3.1.1) is simply the sampling variance formula,

$$\hat{V}_2 = N^2 \frac{1 - f_I}{n_I} S_{ts_I}^2$$

Where

$$S_{ts_I}^2 = \frac{1}{n_I - 1} \sum_{s_I} \left(t_i - \bar{t}_{s_I} \right)^2$$

$$\bar{t}_{s_I} = \frac{1}{n_I} \sum_{s_I} t_i$$

$$f_I = \frac{n_I}{N} \text{ (The cluster sampling fraction)}$$

Another important observation noted by Sarndal [24] is that $\hat{V}_1$ and $\hat{V}_2$ will in fact decrease as n_1 increases (assuming that μ_m, ${\sigma_m}^2$ and ρ_m remain constant). However, when the sample size is increased it could be argued that the estimate of ${\sigma_m}^2$ might in fact increase (due to taking a larger sample we have more chance of obtaining extreme values), leading to a possible increase in the total error ($V_{pm}(\hat{t}_\pi)$).

With these results in hand let us consider MARVL. These results suggest that increasing the number of sample plots will reduce the sample variance, however it will not necessarily reduce the measurement variance (V_1). This is an important point as currently MARVL assumes that the volume obtained from the volume equation does not contain any error. However, not only is the volume equation a measurement error model but in the simulation study conducted in chapter 2, it was shown that incorrectly measuring DBH and tree height can lead to significant differences in the estimate of volume, suggesting that a significant measurement error may exist. This measurement error must be included in the total sample error calculation for volume in order to obtain a true estimate of its total sample error.

4.4 Risk of Underestimating the Total Variance in Surveys with Measurement Error

Estimating the joint variance under the sample design and simple measurement model proves to be difficult in practice since only a single observation for each observed value (y_k) is obtained. Hence one cannot hope to calculate σ_k^2 or σ_{kl}. However, Sarndal [24] page 613 derives a number of interesting theoretical results concerning the bias of the standard variance of the π estimator. To show this Sarndal [24] begins by giving the standard variance estimate of the π estimator assuming that there is no measurement error (equation 3.2.2),

$$\hat{V}(\hat{\pi}) = \sum\sum_S \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

Where

π_{kl} = The probability that both the kth and lth cluster are included in the sample

However, as we have seen in section 4.3 this will in fact be biased when measurement error exists, since

$$E_{pm}\left(\hat{V}(\hat{\pi})\right) = V_{total} - V_{1cen} = V_{1sam} + V_2$$

Hence the bias of $\hat{V}(\hat{\pi})$ as an estimator of V_{total} is,

$$E_{pm}\left(\hat{V}(\hat{\pi})\right) - V_{total} = -V_{1cen} = -\frac{N^2\left[1 + (N-1)\rho_m\right]\sigma_m^2}{N}$$

Where

ρ_m and σ_m^2 are defined in section 4.3

This formulation of the bias for the standard variance estimate of the π estimator leads to a number of interesting conclusions which are summarized in Sarndal [24] page 613,

1. Usually the total variance V_{total} will be underestimated if the standard variance estimate of the π estimator V_{stand} is used as the estimate of total variance, since the bias is negative if,
   $$\rho_m > -\frac{1}{N-1},$$
   which is likely to be the case in practice, such as in a MARVL inventory.
2. In practice ρ_m is likely to be positive (that is positive measurement covariance), which leads to a bigger increase in the bias, which leads to underestimation of total variance becoming a more important issue.

The practical implications of these findings, relating to MARVL, are discussed in chapter 5.

5.0 Conclusions

This chapter summarizes the solutions found in chapters 2, 3 and 4. This chapter discusses how the mathematical discussions of the three preceding chapters relate to the practical situation of a MARVL inventory.

From these three chapters there are three main conclusions. Firstly, chapter 2 demonstrates that the volume equation currently used by MARVL to estimate the timber volume of an individual tree should not be a regression equation but a measurement error model. What this means is that in a practical situation the predictor variables of the volume equation (namely, diameter at breast height and height of the tree) are measured with error (details have been given in sections 2.6.1 and 2.6.2).

This additional error means that the error of the timber volume estimate for an individual tree is actually greater than what is currently being reported in MARVL output. Currently, MARVL assumes that this volume estimate is free from error. To obtain a true estimate of the error for the estimate of timber volume in a particular area of interest, this error must be included in the total error calculation.

There is also the issue of biased volume estimates because increased measurement error gives upwardly biased volume under bark estimates.

This leads to the second main conclusion. It was shown that increasing the sample size decreases the sampling error, however increasing the sample size does not necessarily mean that the measurement error is reduced. This means that increasing the sample size may not reduce the total error, due to the measurement error. If the measurement error is thought to be significant then an equilibrium between sample size and resources used to decrease measurement error should be found in order to find an estimate of timber volume with the smallest total error for a given financial cost.

Further it was concluded in the simulation study conducted in section 2.7, that correctly measuring the diameter at breast height variable was much more important (in terms of a significant difference in the estimate of timber volume for an individual tree) than measuring the height of the tree (at least if measurement errors in measuring DBH and height of a tree are not correlated). This is essentially because equal percentage errors in these two variables have different effects because the DBH is approximately squared and the V_h is not when estimating volume under bark. This means that more resources should generally be put into obtaining measurements of diameter at breast height with the smallest possible error rather than height. This problem could be considered an important practical point in terms of the total error in MARVL estimates as estimating the height of an individual tree can sometimes be difficult due to the severe competition of *Pinus Radiata* at the canopy level making the top of the tree (and hence the height) difficult to determine. That is, what matters most is getting the DBH measurement accurate, and this is easier than for height because it is measured much nearer to ground level. Also the best timber is nearer the ground level.

However, from the simulation study it was found that the two-way interaction between the error in Height and the correlation between the error in Height and the error in DBH, results in the most significant difference in the estimate of timber volume. What this means, in terms of MARVL, is that if a field crew member over estimates the diameter at breast height and also over estimates the height, then obviously this has much more effect on estimating the timber volume of the tree than if just one of these variables were over estimated. Hence necessary resources should be used in order to educate the field crews in order to allow them to obtain measurements of DBH and Height that have the smallest possible error, thus reducing the total error of the timber volume estimate.

Let us now discuss the effect this measurement error has on the estimate of total timber volume for a particular area of interest. Firstly, the simulation study showed that a measurement error in DBH leads to a significant difference for the estimated volume of an individual tree. More importantly it showed that increasing the error in DBH leads on average to an increase in the estimate of timber volume. What this means in practical terms is that if a field crew member is measuring DBH with greater error, then on average the estimate of timber volume will be over estimated.

This observation becomes more important if we consider the two-way interaction between the error in Height and the correlation between this (via V_h) and the error in DBH. From the simulation study conducted in section 2.7, it was shown that this effect was the most significant. Let us consider this effect in practical terms. What this effect means is that if a field crew member measures the Height of the tree with error (which is usually the case) and also measures DBH with error (but in the same direction as Height, that is if height is overestimated, then DBH is also usually overestimated), then on average the estimate of timber volume will be over estimated. From the output in section 2.7, the overestimate of volume is approximately 4% (at worst).

We must also take into consideration the sample design used to derive the estimate of total volume. It was shown in chapter 3 that the estimate of the mean timber volume is basically the mean of the cluster volume averages. However, if the field crew member is consistently measuring DBH and height incorrectly for each tree in each cluster then by taking averages the average timber volume for a particular cluster will also be overestimated.

The question that should be asked from this simulation study is weather or not a error in timber volume of 3-4% for an individual tree, effects the estimate of timber volume for a particular area of interest enough to cause the estimate to be biased in terms of harvest planning and other uses of MARVL. Essentially increased volume estimates due to measurement error for each tree leads to the same level of measurement error for aggregations of tree, for example clusters, blocks etc, so that overestimates of up to 4% for timber volume from aggregated tree volume in forest inventories using MARVL are possible for feasible levels of measurement error.

The other form of error that is discussed in the preceding chapters is the sampling error. Sampling error relates to the error incurred by taking a sample to estimate a variable of interest rather than the entire population. Sampling error is subject to sample to sample variation. The formulas MARVL currently uses to obtain an estimate of timber volume for an area of interest assume that the primary sampling unit is the average of the individual sample plot. This leads to the assumption that the variation in timber volume within a sample plot is zero. Thus, the design (and subsequent formulas) used to calculate an estimate of timber volume for an area of interest is a simple random sample of cluster averages.

However, the assumption that the variation of the ratio of volume to DBH within a sample plot is zero is not correct since no two trees have identical shapes. Standard survey theory bases sample designs on the elementary unit (in MARVLs case the individual trees). The estimation formulae that should be used are for a two-phase cluster sample.

It should also be noted that the current data collection method used in a MARVL inventory includes the necessary information to derive an estimate of the within plot variation, hence no additional field work needs to be done to obtain the two-phase cluster sample total and its variance. This would increase the sampling error estimates although the effect is difficult to quantify in general in an exact way because it depends on the cluster sizes and within cluster variation of the ratio of volume to DBH for individual trees. Effects of not allowing for this variation often lead to underestimates of variance of about 10% but not to biases in (what in this case are volumes) estimates (which is the more important issue).

It was also concluded in section 3.2.1.1 (the efficiency of single-stage cluster sample), that an equal probability cluster design (that is, a design where all the selection probabilities are equal) is often a poor choice when the clusters are of different sizes. This is the case in MARVL, since a fixed area based plot is used as the cluster, however since this is based on a fixed size, then number of trees with in a cluster may vary. It was shown that if the mean of the cluster was roughly proportional to one over the size of the cluster (that is, one over the number of elements in that cluster), then this would yield an efficient design.

What is necessary then in terms of sample design is to ensure that cluster samples are designed efficiently and that the variance of the estimates of total volume is not underestimated, or the estimates of volume themselves are biased.

The combination of high measurement errors and any incorrect use of sampling variance formulae will compound biases in timber volume estimation and also lead to underestimates of their variance or mean square error, as well as overestimates of volume.

In conclusion, feasible levels of measurement error can lead to upward biases in estimates of timber volume of up to 4% when using MARVL, and this measurement error bias problem is not solved simple by increasing the number of trees sampled. Sample designs themselves need to be efficient, and to use the correct formulae for estimation of sampling error, else volume estimates will be thought to be more accurate then they are, even putting aside the measurement error issue. Finally even if correct formulae are used for estimating sampling error, measurement error needs to be incorporated into the estimates of accuracy, as well as the estimate of volume themselves, if volume of timber is not to be overestimated, and total variance and total mean square error for that volume are not to be serious underestimated.

6.0 References

[1] Anton, H. (1991), "Elementary Linear Algebra", Sixth Edition, John Wiley and Sons Inc.

[2] Bain, L.J. and Engelhardt M, (1992), "Introduction to Probability and Mathematical Statistics", Second Edition, PWS-KENT Publishing Company.

[3] Biemer P.P, Groves R.M, Lyberg L.E, Mathiowetz N.A., Sudman S. (1991) "Measurement Errors in Surveys", John Wiley & Sons Incorporated.

[4] Breidt F. Jay & Fuller Wayne A., (1999), "Design of Supplemented Panel Surveys With Application to the National Resources Inventory", Journal of Agricultural, Biological, and Environmental Statistics, Volume 4, Number 4, Pages 391-403.

[5] Carroll, R.J., Ruppert, D., Stefanski L.A. (1995) "Measurement Error in Nonlinear Models", Chapman and Hall.

[6] Casella, G. & Berger R.L. (1990), "Statistical Inference", Duxbury Press.

[7] Cheng C.L, Van Ness J.W. (1999) "Statistical Regression with Measurement Error", Arnold.

[8] Cochran W. G., (1963), "Sampling Techniques – Second edition", John Wiley & Sons, Inc.

[9] Deadman M.W. & Goulding C.J., (30 May 1979), "A Method For Assessment of Recoverable Volume by Log Types", New Zealand Journal of Forestry Science, Volume 9, Pages 225-239.

[10] Dobson A.J. (1990) "An Introduction to Generalized Linear Models", Chapman and Hall.

[11] Draper N.R. and Smith H. (1981) "Applied Regression Analysis", John Wiley and Sons (Second Edition).

[12] Fuller W.A, (1987) "Measurement Error Models", John Wiley & Sons Incorporated.

[13] Fuller Wayne A., (1999) "Environmental Surveys Over Time", Journal of Agricultural, Biological, and Environmental Statistics, Volume 4, Number 4, Pages 331-345.

[14] Gertner G.Z., (1990), "The Sensitivity of Measurement Error in Stand Volume Estimation", Canadian Journal of Forest Research, Volume 20, Pages 800-804.

[15] Hansen M.H, Hurwitz W.N, Madow W.G. (1953) "Sample Survey Methods and Theory, Volume I", John Wiley and Sons, INC.

[16] Hogg & Tanis, (1997), "Probability and Statistical Inference", Prentice-Hall International, Inc.

[17] Lehtonen R and Pahkinen E.J. (1996), "Practical Methods for Design and Analysis of Complex Surveys'", John Wiley and Sons, New York.

[18] Lessler J.T, Kalsbeek W.D. (1992) "Nonsampling Error in Suveys", John Wiley & Sons Incorporated.

[19] Maclaren, J.P., (1993) "Radiata Pine Growers' Manual", New Zealand Forest Research Institute, Bulletin number 184.

[20] McCullagh P. & Nelder J.A. (1989), "Generalized Linear Models", Second Edition, Chapman and Hall.

[21] McRoberts R.E. & Hansen M.H., (1999), "Annual Forest Inventories for the North Central Region of the United States", Journal of Agricultural, Biological, and Environmental Statistics, Volume 4, Number 4, Pages 361-371.

[22] New Zealand Forest Research Institute Ltd. (1990) "What's new in forest research – MicroMARVL, versatile plantation inventory", New Zealand Forest Research Institute Ltd., No. 191.

[23] Reams G.A. & Van Deusen P.C., (1999), "The Southern Annual Forest Inventory System", Journal of Agricultural, Biological, and Environmental Statistics, Volume 4, Number 4, Pages 346-360.

[24] Sarndal C.E., Swensson B & Wretman J, (1992), "Model Assisted Survey Sampling", Springer

[25] Schreuder H.T., Gregoire T.G. & Wood G.B., (1993), "Sampling Methods For Multiresource Forest Inventory", John Wiley and Sons, Inc.

[26] Stein S.K. & Barcellos A. (1992), "Calculus and Analytical Geometry", McGraw-Hill Inc.

[27] Theil, H. (1971), "Principles of Econometrics", New York: Wiley.

[28] Thompson S.K., (1992) "Sampling", John Wiley and Sons, Incorporated.

Appendix A – MINITAB Macro used in simulation study.

```
GMACRO

ME_Sim

#k10=Var(log(DBH))
let k10=0.007
#k11=Var(log(Vh))
let k11=0.009
#k12=Rho
let k12=0

while(k10~=0.009)
let k11=0.001

while(k11~=0.009)
let k12=0

while(k12~=1)
let k13=1

#Generates the sigma_uu matrix (cov matrix of measurement errors)
        let c22=k12*sqrt(k10)*sqrt(k11)
        let c20=k10
        let c21=k11
        Stack c20 c22 c16.
        Stack c22 c21 c17.
        copy c16 c17 m6
        name m6 'sigma_dd'

while(k13~=21)
#Generates random data from a Normal(0,K11) distribution (for log(Vh))
        Random 233 c24;
        Normal 0.0 K11.

#Generates random data from a Normal(0,K10) distribution (for log(DBH))
        Random 233 c25;
        Normal 0.0 K10.


#Creates the x matrix (ie actual value plus the measurement error)
        let c30=c7+c25
        let c31=c8+c24

#Creates the Mean vector (x_bar)
        Mean C30 k30.
        Mean C31 k31.
        copy k30 c12
        copy k31 c13
        Stack  C12 C13 c14.
        copy c14 m1
        name m1 'x_bar'
```

```
#Generates the s_xx matrix
        Name m2 = 's_xx'
        Covariance  C30 C31 's_xx'.

#Generates the s_xy matrix
        Covariance  C30 'Log(Vol)' 'cov_dbh_vol'.
        Covariance  C31 'Log(Vol)' 'cov_vh_vol'.
        copy m3 c40-c41
        copy c40 m50
        Transpose m50 m50
        copy m50 c42 c43
        copy m4 c44-c45
        copy c44 m51
        Transpose m51 m51
        copy m51 c46 c47
        copy c43 c47 m5
        Transpose m5 m5

#Calculates the intercept and beta vector of the ME_model
#Beta_hat Vector

        Subtract 'sigma_dd' 's_xx' m7.
        Invert m7 m8.
        Subtract 'sigma_de' M5 m10.
        Multiply m8 m10 m11.

#Intercept

        Transpose m11 m13.
        Multiply m13 m1 m14.
        Subtract M14 m12 m15.


#Moves the Beta and intercept values to columns
        copy m13 c33 c34
        copy m15 c32

#Calculates the log volume using the measurement error parameter ests
        let c50=c32+(c33*c30)+(c34*c31)
        Stack c51 c50 c51.

        let k13=k13+1

endwhile

let k12=k12+0.2
endwhile

let k11=k11+0.002
endwhile

let k10=k10+0.002

endwhile

ENDMACRO
```