Estimating Uncertainty in the Kyoto Compliant Carbon Accounting System for New Zealand's Planted Forests

Brown, J.A.[1], Woollons, R.C.[2], and Manley, B.R.[2]

[1]Biomathematics Research Centre
University of Canterbury
Christchurch
New Zealand

[2]School of Forestry
University of Canterbury
Christchurch
New Zealand

20 July 2005

Prepared for Ministry of Agriculture and Forestry, PO Box 2526 Wellington.

University of Canterbury Technical Report number UCMS2005/7

Disclaimer

This report was commissioned by the Ministry of Agriculture and Forestry to provide an initial estimate of uncertainty of national carbon stock change based on currently available data.

The report is only for the use by the entity that commissioned it and solely for the purpose stated above. Canterprise shall have no liability to any other person or entity in respect of this report, or for its use other than for the stated purpose.

CONTENTS

EXECUTIVE SUMMARY

- The calculation of carbon stored in New Zealand Kyoto planted forests is likely to be estimated as some functional combination of variables that describe the land area of forest and forest biomass. The details of the final carbon model are not currently available.

- Uncertainty analysis will be conducted to identify the limits to the accuracy of any carbon estimate.

- Sensitivity analysis will be undertaken to identify and rank model variables in terms of their contribution to the overall uncertainty. Variables that contribute most to the overall uncertainty and cause the greatest effect should be the highest priority for allocation of extra survey resources.

- A combination of analytical and numerical methods (Monte Carlo simulations) will be used in the uncertainty analysis.

- Standard methods assume that the variables are normally distributed and are independent. Our analysis may account for variables that are not normally distributed and display lack of independence. At present there are very little data to test for normality and dependencies.

ACKNOWLEDGEMENTS

Thank you to the following for their advice and their useful discussions: John Moore (Forest Research), Paul Lane (MAF) and James Barton (MfE).

1. GENERAL PRINCIPLES

The object of this Report is to summarise the techniques available to quantify errors associated with calculating carbon stored in New Zealand's Kyoto planted forests.

The Report is divided into two major sections:

(a) A general review of the methods to assess uncertainty in estimating stored carbon. Many of these methods rely on the availability of the data that will be used to estimate carbon. To date these data are not available, but it is anticipated data will be acquired as the overall Kyoto process in New Zealand develops.

(b) A specific description of the method likely to be employed in 2004-2005 to obtain first estimates of the error involved in estimating carbon stored in New Zealand's Kyoto planted forests.

1.1 An Overview

Forest inventories conducted according to good practice are those that neither over- nor under-estimate and where uncertainty has been reduced as far as practicable. Under- and over-estimation is a result of bias where the estimate of carbon change is inaccurate. Good practice implies that any identifiable bias is removed. However, some uncertainty in the final estimate will remain and identification of this is an important part of good practice. Identification and attaining some level of understanding of uncertainty allow resources to be directed to reducing uncertainty.

'Uncertainty' can be regarded as a lack of knowledge about the true value of a quantity (Cullen & Frey, 1999). Uncertainty analysis is concerned with identification of credible limits to the accuracy of an estimate and the extent that an estimate may differ from the true value. It is a structured process where appropriate methods are used to determine uncertainty in each component of the estimation process and to aggregate these. Sensitivity analysis is a valuable tool for identifying and ranking model variables in terms of their contribution to the overall uncertainty. Variables that contribute most to the overall uncertainty and cause the greatest effect should be the highest priority for allocation of extra survey resources.

1.2 Uncertainty Analysis

1.2.1 Terminology and definitions

Some of the current literature on uncertainty analysis uses terminology that differs from standard statistical practice, for example, the terms parameters and variables (see Winiwarter and Rypdal, 2001). We use normal statistical practice and define a variable to be generally, any quantity that varies. More formally a variable is a quantity that can take any one of a specified set of values.

Associated with most variables is the concept of a probability distribution. Any specific mathematical distribution has a set of parameters that are used in defining the distribution. For example, the Weibull probability density function has a form:

$$f(X) = \frac{c}{b}\left(\frac{X-a}{b}\right)^{c-1} \exp\left(-\left|\frac{X-a}{b}\right|^{c}\right) \qquad -8 < X < 8$$

where *a* is a location parameter, *b* a scale parameter, and *c* a shape parameter.

1.2.2 Steps in uncertainty analysis

Central to uncertainty analysis is the concept of a probability distribution function. Such a distribution describes the range and likely variation in the possible values of any variable. Using this distribution and its associated error term, confidence intervals can be calculated and defined, with a specified probability, the range of values within which the true variable will lie.

The major steps in uncertainty analysis are:

- Defining the model.
- Listing all input variables.
- Specifying the maximum likely range of potential values for the unknown parameters that will be estimated.
- Specifying a specific probability distribution for values occurring within this range.
- Determining and accounting for correlations among the input variables.
- Using either analytical or numerical procedures, and estimating the uncertainty in each of the model variables.

To perform a quantitative uncertainty analysis, probability distributions are assigned to each of the uncertain variables. The chosen distributions preferably result directly from data but sometimes are selected from subjective judgment. There are a number of different distributions that are commonly used, including the normal distribution, and the assumption of these or other empirical distributions is usually dependent on the availability of relevant data. Distributions that possess explicit expressions for estimation of their parameters are clearly to be preferred. Therefore, when empirical data exists a normal distribution is a good first approximation (or lognormal or truncated if negative values are unrealistic, and supplemented by uniform or triangular distributions). If there is good, compelling reason, other more complex distributions can be considered.

If the appropriate probability distribution is unknown, uncertainty analysis needs to include the uncertainty of the unknown distributions. This is easiest to do in Monte Carlo simulations by using a number of different distributions in multiple simulations. Monte Carlo simulation is described in section 1.2.4.

1.2.3 General methodology for fitting a probability density function

The choice of probability distribution starts by inspecting the data. If the data are from a valid random probability sample and can be assumed to be representative of the range of conditions, then classical statistical methods can be used to estimate distribution-parameter values (e.g., the mean and variance for a normal distribution) using the method of moments or maximum likelihood (Meyer, 1965). Specific details for some distributions are given in section 2.3. The goodness of fit of the distribution to the sample data needs to be assessed using the standard methods with probability plots and other statistical tests, for example, χ^2, Anderson-Darling, or Kolmogorov-Smirnov goodness-of fit tests (Stephens, 1986)

1.2.4 Assessing total errors: Monte Carlo simulation

Error propagation is the calculation of the cumulative errors from the model-equation. For a simple additive model with no dependencies among variables the variance of the model can be estimated directly by the weighted sum of the variances of the variables (see section 1.5.2). When dependencies exist among the variables analytical methods become more complex and for non-normal distributions numerical methods can be used, for example, Monte Carlo simulation.

There are various explanations and descriptions for what are Monte Carlo simulations. We use the definition that Monte Carlo simulations are a numerical method to sample the distributions of input variables to generate a representative distribution of predicted model values.

Monte Carlo simulations, in their simplest form proceed by taking a random value from a specified probability density function of each variable in the model. The process is repeated for a very large number of iterations resulting in a probability distribution of the model estimate. More sophisticated methods of choosing the random values include latin-hypercube-sampling (McKay *et al.*, 1979), a method that is statistically more efficient. Instead of choosing each value randomly, the probability density function is systematically divided into (small) equiprobable intervals and samples taken from each interval thus ensuring representation of a full range of values. The actual algorithm for drawing the random values uses the inverse cumulative density function rather than the probability density function.

One of the advantages of Monte Carlo simulations is that if the model becomes more complex, the complexity of the Monte Carlo technique largely does not necessarily change. In comparison, with analytical methods variance estimation can become very difficult (Hammonds *et al.*, 1994). Another advantage is that when the probability

distribution of any input variable is completely unknown Monte Carlo procedures can include simulation over several different distributions, in effect adding in uncertainty due to the lack of knowledge about any variable's distribution.

When dependencies exist among the variables Monte Carlo methods can still be used. In the case of two variables that are normally distributed simple algorithms exist to generate random deviates allowing for the existence of correlation between them (see section 2.5.2). For non-normal distributions the methods become approximate and complex, although is still possible. More generally, correlations among any type and number of distributions can be simulated using an approximation based on rank correlation (Cullen & Frey, 1999, Frey, 1992).

1.3 Sensitivity Analysis

Sensitivity analysis is used to assess the relative importance or contribution of variables in a model to the overall uncertainty and can be used to prioritise effort to reduce uncertainty. Uncertainty will be reduced by expending extra effort in the estimation of the variables with the highest contribution.

Spearman rank correlation coefficients can be used to quantify individual variable's contribution. This method involves ranking the values simulated in the Monte Carlo process. The model estimates simulated in the Monte Carlo process are also ranked. The association between the rankings is measured by the Pearson rank correlation coefficient, *R*. The correlation coefficient between the two rank orderings (*x* and *y*) is calculated as:

$$R = \frac{\sum_{i=1}^{n}(R_{ix} - \bar{R}_x)(R_{iy} - \bar{R}_y)}{\sqrt{\sum_{i=1}^{n}(R_{ix} - \bar{R}_x)^2}\sqrt{\sum_{i=1}^{n}(R_{iy} - \bar{R}_y)^2}}$$

where n is the number of values ranked and R_{ix} is the rank of the i^{th} value in the x data set and R_{iy} is the rank of the i^{th} value in the y data set, and $\bar{R}_x, \bar{R}_y$ are the respective means (Iman and Conover 1982).

The Pearson rank correlation coefficient varies between +1 and -1. A value of +1 is when the rankings are identical, and -1 when the rankings are as greatly in disagreement as possible (that is, one rank order is the opposite of the other). The significance of R can be tested as:

$$t = R\sqrt{\frac{n-2}{1-R^2}}$$

with (n - 2) degrees of freedom and compared to the Student's *t* distribution (McBean and Rovers, 1998). Higher coefficients indicate higher relative contribution.

Other methods for sensitivity analysis are to graph the Monte Carlo estimates against the simulated variables. Graphs with obvious trends are indications of higher contribution to the uncertainty. Tornado graphs (Winston, 2001) are a very useful summary of the ranking of variables. Many other methods exist including multivariate linear regression and probabilistic sensitivity analysis (Cullen and Frey, 1999). Smith and Heath, (2001) discuss the use of two 'Importance indices' for ranking the contributions of 10 carbon inputs to an uncertainty model.

1.4 Sources of Uncertainty: Components of Error

The contribution to the uncertainty in the estimate of carbon stored in New Zealand's Kyoto planted forests from the assessment process has many components. Errors will occur from measurement, sampling, classification, estimation, model errors and uncertainty from expert judgment.

Measurement errors include incorrect measurement in the field, limitations in instrument accuracy or discrimination threshold, and errors in recording and transcribing data. Sampling uncertainty is introduced when there is inherent natural variation and only a fraction of the entire population is measured. An example of classification error is the map estimates of forest land. Classification errors can occur when forest land is mapped as non-forest or vice versa, or is incorrectly aged, thus introducing errors in estimation of the total forested area. Model uncertainty is used to represent lack of confidence that the mathematical model is a "correct" formulation for carbon-change assessment. Model uncertainty exists if there is a possibility of obtaining an incorrect result even if exact values are available for all of the model parameters (variables) and above all if the structure of the model is incorrect.

1.5 Estimating Uncertainty

1.5.1 Analytical methods for combining uncertainties

When uncertainties are combined by addition, that is, the model Z is the linear and the sum of variables,

$$Z = \sum_{i=1}^{n} a_i Y_i$$

where a_i are constants, the variance of Z, when there are no dependencies or correlations among the variables, is:

$$V(Z) = \sum_{i=1}^{n} a_i^{\,2} V(Y_i),$$

where $V(Y_i)$ is the variance of the variable Y_i.

However when dependencies are existent the covariance ($Cov(Y_i Y_j)$) between every pair of variables is included

$$V(Z) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2\sum \sum a_i a_j Cov(Y_i Y_j),$$

where,

$$Cov(Y_i Y_j) = r_{ij} s_i s_j$$

and $?_{ij}$ is the correlation coefficient between Y_i and Y_j and s_i and s_j the respective standard deviations (Lindgren, 1993)

When uncertainties are combined by multiplication, that is, the model Z is the product of variables, the overall uncertainty can be more complicated to calculate. If there are only two variables, Y_1 and Y_2,

$$Z = \prod_{i=1}^{2} a_i Y_i$$

the overall uncertainty when there is no dependency or correlation among the variables is

$$V(Z) = V(Y_1)V(Y_2) + a_1^2 V(Y_2) + a_2^2 V(Y_1).$$

When there are dependencies between two variables the correlation, r, between the variables is included,

$$V(Z) = V(Y_1)V(Y_2) + a_1^2 V(Y_2) + a_2^2 V(Y_1) + r^2 V(Y_1)(Y_2) + 2Y_1 Y_2 r\sqrt{V(Y_1)V(Y_2)}.$$

These formulae can be extended to three or more variables.

1.5.2 Numerical methods for combining uncertainties

The advantage of Monte Carlo methods is that dependencies can be more readily incorporated into the uncertainty estimation process. When there are no dependencies, given some distributional form for each variable in the carbon estimation model, values are randomly selected from each distribution and combined using the model. When dependencies exist, the method of selecting the random variables is effectively constrained. If two variables, Y_1 and Y_2, are correlated, then the range of likely values of Y_2 is conditional on the value of Y_1.

1.5.3 The effect of dependencies

Correctly specifying dependency and correlations may be less important for variables that do not have a high contribution to the uncertainty. The relative contribution of variables to the uncertainty can be identified in the sensitivity analysis. Similarly, weak dependency among variables that do have high contribution to the uncertainty will have little consequence to the analysis. Dependencies can be incorporated into the uncertainty

analysis by modeling it explicitly, or to use a range of methods that manipulate the data directly including aggregating the input variable values into categories and re-sampling techniques.

1.6 Changes in Kyoto Carbon: estimation of differences

Many carbon statistics will be estimated in terms of change in variables or as differences over time. Assuming no dependencies, the variance of the difference of a variable measured in time 1 and then in time 2 is,

$$V(a_1Y_1 - a_2Y_2) = a_1^2V(Y_1) + a_2^2V(Y_2),$$

and with dependencies,

$$V(a_1Y_1 - a_2Y_2) = a_1^2V(Y_1) + a_2^2V(Y_2) - 2r_{12}s_{y_1}s_{y_2}.$$

In estimating change in carbon uncertainty is reduced when the value of the variable when measured at time ($t + ?t$) is correlated to the value at time t. An example of this is when sample plots are repeatedly surveyed for the same variable. Tree volume at time ($t + ?t$) will be correlated with tree volume at time t etc. Growth variables typically have strong correlations.

It is likely that variables will be measured in installed permanent sample plots then re-measured later, maybe several times. In these situations, and dependent on the sample survey technique chosen, the variance of the average difference (for example, at two different ages) of a variable can be surprisingly precise.

There are several variants to the appropriate survey methods but most give good precision by virtue of strong correlations frequently existing between variables at two distinct ages; growth variables give especially strong correlations. For example, if we have a set of sample plots and we estimate volume/ ha at ages T and ($T + ?T$) then correlations in excess of 0.9 can be expected (Some details and an example are given in Appendix 1).

2 METHODS

2.1 Uncertainties in Carbon Estimation: Current Situation in July 2004

The major sections above summarise strategies for the estimation of total stored carbon together with potential methodologies to calculate the likely errors. These methods assume the on-going availability of data collected over several years and a scenario of continuing estimation of changes in carbon stored in New Zealand's Kyoto planted forests.

However, it is certain relatively few of these methods will be able to be used in 2004-2005 or will only be able to be used if various assumptions are made. Essentially the data are or will not be available. It is therefore difficult to describe specific and detailed methodology to measure uncertainty when the specific methods to estimate carbon changes are yet to be decided.

2.2 A Simple Method of Estimation

One basic method to estimate carbon stored in New Zealand's Kyoto planted forests is as a function of the following major variables:

C_1 = area x stem volume x *BEF* x density x carbon fraction (1)

or,

$C`_1$ = area x (stem mass + stem bark mass + live branch mass + dead-branch mass + total foliage mass + total root mass) x carbon fraction (2)

C_2 = area x soil carbon (3)

C_3 = area x dead organic matter x carbon fraction (4)

where *BEF*, the biomass expansion factor, is

BEF = (above ground biomass + below ground biomass) / stem biomass.

Thus depending upon whether C_1 or $C`_1$ is used the appropriate error formulae (addition or products, given above) are applicable.

Equations (1) to (4) are simplistic in the sense in that most of the components later should be able to be expanded in many ways or broken down into several sub-components. For example, in the simplest case 'area' might be defined to be the total estimated area of Kyoto defined forest. An extension is Kyoto forest area, but broken down by age classes. In the same way, volume/ha could be a simple average or alternatively expressed in the form of a yield-age volume table. In both cases, the introduction of age should significantly enhance accuracy but when these additional data will become available is not known.

2.3 Estimating Parameters of Probability Density Functions

Theoretically, there are a wide range of probability density functions available to summarize data. Most important in this context is the symmetric Normal distribution but also relevant are the log-normal, Weibull, Gamma and Beta distributions that can depict asymmetric data. For other datasets, the uniform and exponential distributions may be relevant (Johnson *et al.*, 1994).

Table 1 *Common probability density functions (p.d.f.'s) with their mathematical expressions and limits.*

Distribution	p.d.f	Limits
Normal	$f(X) = \frac{1}{s\sqrt{2p}} \exp\left(-\frac{1}{2}\left\| \frac{X-m}{s} \right\|^2 \right)$	- 8 < X < 8
Weibull	$f(X) = \frac{c}{b}\left(\frac{X-a}{b} \right)^{c-1} \exp\left(-\left\| \frac{X-a}{b} \right\|^c \right)$	a = X < 8
Uniform	$f(X) = \frac{1}{(b-a)}$ = 0 elsewhere	a = X = b
Exponential	$f(X) = a^{-1} \exp\left(-\left(\frac{X}{a} \right) \right)$ = 0 elsewhere	X > 0

Given suitable data, for any selected distribution this leads to the method to estimate the distribution's parameters. Well-known methods of estimation include moments, maximum-likelihood, and percentiles (Kendall *et al.*, 1994). Statistically, maximum likelihood estimation can be shown to be the most robust technique, possessing minimum variance properties.

In practice, however, the advantages of maximum-likelihood estimation can be minimal because biological data is frequently characterised by high variation and lack of strict adherence to distribution models and many of the above distributions do not have analytic maximum-likelihood solutions.

We consider the most important functions to be the Normal, Weibull and occasionally the Uniform and Exponential distributions (Table 1). Collectively, they are capable of reasonably depicting most biological datasets, and all have practical methods of parameter estimation (Table 2).

Table 2 *Methods of estimation of common probability density functions, where* $\bar{X}$ *is the sample mean and n the sample size.*

Distribution	Estimation method	Estimates
Normal	ML,	$\mu = \bar{X}$ $s^2 = \frac{\sum_{i=1}^{n} X^2 - (\sum_{i=1}^{n} X)^2 / n}{(n-1)}$
Weibull	Garcia(1981)	a = min (X_i) $(1/b) = \left(\frac{\Gamma(1+1/c)}{\bar{X} - a} \right)^c$ Define $z = s/(\bar{X} - a)$ $\frac{1}{c} = z(1 + (1-z)^2 \sum_{i=0}^{3} k_i z^i)$ k_0 = -0.22106417 k_1 = 0.010060668 k_2 = 0.117358987 k_3 = -0.050999126
Uniform	ML	b = Max(X_i) a = Min(X_i)
Exponential	ML	a = $\bar{X}$

2.4 Methods for Combining Uncertainties

2.4.1 Analytic methods

Given access to the average values of each major input variable and an estimate of each corresponding standard deviation then assuming independence among the variables, uncertainty can be estimated using the methods described previously.

It is likely that some of the input variables will show some dependence to each other. We anticipate, for example, that volume/ha and density may show some dependence. Both change with age so some association is expected. The linear relationship between any two variables can be measured using Pearson correlation coefficients ($?_i$) estimated by r,

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

If relationships between variables are curvilinear or non-linear then transformations can be useful (for example, logarithmic or square-root).

See section 1.5.1 for details of analytical methods to combine uncertainty.

2.4.2 Numerical methods

Simulation methods will use Monte Carlo approaches. Monte Carlo simulations can be generated using SAS with a purpose-built program. Initial simulations will assume that each input variable is normally distributed and independent of each other. The normal random number generator in SAS will be used, RANNOR, to generate a large number (j) of random normal deviates, Z_{ij}, for the i distributions, each with an estimated mean (μ_i) and standard deviation (s_i) .

$$Z_{ij} = \mu_i + s_i \, RANNOR\,(seed_i)$$

where $seed_i$ is a set of (different) random numbers

Total carbon can be estimated for each of the j simulations. The associated total error is estimated by the distribution of the j estimates of carbon.

In this basic form the method is likely to produce very similar results to those outlined in 2.4.1. However, small extensions to the methodology are likely to give more information. For example, as suitable data becomes available some of the major predictor inputs should be able to be split into a series of sub-classes, each represented by a set of distributions, and combined through appropriate addition or multiplication. For example, volume may be able to be divided into age-classes. Moreover the methodology here easily extends to subsequent sensitivity analyses (see section 2.7).

Some input variables are likely not to be normally distributed, that is, they exhibit some degree of asymmetry or are better modelled on theoretical grounds by alternative distributions. Provided independence is still assumed, the methods described still holds, but rather than using the algorithm given, alternative algorithms to generate random deviates specific to each distribution will be used. Distributions will be expressed in inverse cumulative form and uniform random deviates used to generate the appropriate deviates (Neelamkavil, 1987). Table 3 lists the appropriate inverse functions for the commonly used distributions. The various parameters shown are identical to those in Table 1.

Monte Carlo methods can be modified to account for dependencies. If a correlation (?) exists between any two variables then the suitable algorithm can be used to generate two correlated normal deviates Z_1 and Z_2, e.g.,

$$Z_1 = \mu_1 + ?s_1 \text{ RANNOR(seed1)},$$

then

$$Z_2 = \mu_2 + ?s_2 \text{ RANNOR(seed1)} + s_2(1 - ?^2)\text{RANNOR(seed2)},$$

where seed1, seed2 are two randomly chosen positive numbers.

When the probability density functions are not normal then the Monte Carlo process is considerably more complex although still possible through approximate methods (Cullen and Frey, 1999). One method is to use approximations based on rank correlations.

Table 3 *Inverse cumulative distribution functions*

Distribution	Function
Normal	No closed form (use SAS RANNOR)
Weibull	$X_p = a + b(-\ln(1-p))^{1/c}$
Uniform	$X_p = a + p(b-a)$
Exponential	$X_p = -a\ \ln(1-p)$
	where p is a random uniform deviate.

2.5 Uncertainty in the Estimated Change in Carbon

The primary interest is in changes in carbon estimates over time (flux), rather than in estimates of carbon at a given time (stocks). If estimates in two different years are made in the same manner using the same data collection procedures and analysis techniques then the variance of the difference between the two years is likely to be considerably smaller than the variance of the individual estimates. Monte Carlo simulation methods will be used to estimate uncertainty in the estimated change in carbon stocks.

2.6 Sensitivity Analyses

For initial analyses, inspection of the respective coefficients of variation will give some insight into the relative importance of any input variables. In general, the larger the coefficient the more the variable's influence on estimated carbon. This can be used on ranking variable importance.

Sensitivity analyses with the Monte Carlo simulations will include scatter plots of the estimated carbon versus the individual variable values of each simulation. These have the advantage of not only providing visual evidence of dependence but more importantly they reveal the shape of any associations. The estimated strength of the relationship between the variable values and estimated carbon can be used to rank the variables by relative importance. Pearson correlation coefficients are suitable for measuring linear relationships, and when there is nonlinearity still provide monotonic dependency rank correlations that can be used.

3 CONCLUDING COMMENTS

In this report we describe the various techniques available to quantify the size of the various errors associated with carbon stock changes in New Zealand Kyoto planted

forests. Initial analysis of linear models with variables that are normally distributed and have no dependencies will be relatively straightforward. Complexity will be introduced when non-normal and dependent variables are analysed.

We have been restrained in this preliminary report to describing potential methods. When we have been provided with information on how carbon stocks changes are to be estimated, and access to pilot data the analysis will be conducted. The estimation of carbon is likely to be based on a model with input variables for area, volume, density, and biomass. The degree of complexity and detail intended to be included in each component is currently unknown. Thus, and especially for initial estimates of carbon uncertainty, we cannot be totally specific; similarly it would have been advantageous to demonstrate some of the techniques with pilot data.

For an initial estimate of carbon uncertainty we have indicated we intend to use SAS software to construct the various algorithms or simulation routines we require. For this initial development of the analysis SAS is appropriate to use especially because of the ability in SAS to produce deviates from any distribution). In the longer run, however we may recommend application orientated software such as @RISK (Palisade Corporation, 1997) which is specifically designed for uncertainty analysis.

4 REFERENCES

Cullen, A.C. and Frey, H. C., 1999. Probabilistic Techniques in Exposure Assessment. Plenum Press, New York and London.

De Vries, P. G., 1986. Sampling Theory for Forest Inventory. Springer-Verlag, New York.

Freese, F., 1962. Elementary Forest Sampling. Agriculture Handbook No. 232., U. S. Dept of Agriculture.

Frey, H. C., 1992. Quantitative Analysis of uncertainty and variability in environmental policy making. American Association for the Advancement of Science, Washington, DC.

Garcia, O., 1981. Simplified method-of moments estimation for the Weibull distribution. *New Zealand Journal of Forestry Science* 11: 304-307.

Hammonds, J. S., Hoffman, F.O., and Bartell, S. M., 1994. An Introductory Guide to Uncertainty Analysis in Environmental Health and Risk Assessment. SENES Oak Ridge, Inc, Oak Ridge, TN, April.

Iman, R.L and Conover, W.J. 1982. A distribution-free approach to inducing rank correlations among input variables. *Communations in Statistics* B11(3): 311-334.

Johnson, N. I., Kotz, S., and Balakrishnam, N., 1994. Continuous Univariate Distributions. 2 nd. Edition. Volumes 1 and 2. Wiley, New York.

Kendall, M. G., Stuart, A., Ord, J. K. and O'Hagan, A., 1994. The advanced theory of Statistics, 6 th. Edition. (in 3 Volumes), Edward Arnold, London.

Lindgren, B. W., 1993. Statistical Theory, 4 th. Edition. Chapman and Hall New York.

McBean, E.A., and Rovers, F.A. 1998. Statistical procedures for analysis of environmental monitoring data and risk assessment, Prentice Hall, New Jersey.

McKay, M.D., Conover, W.J., and Beckman, R.J. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21(2): 239-245.

Meyer, P. L., 1965. Introductory Probability and Statistical Applications. Addison-Wesley, Massachusetts, U.S.A.

Neelamkavil, F., 1987. Computer simulation and modeling. Jon Wiley and Sons.

Palisade Corporation, 1997. Risk Analysis Excel Software. 31 Decker Road, NewField, New York, USA.

SAS Institute Inc., SASQC Software: Usage and Reference, First Edition, Volume 1, Cary, NC: SAS Institute Inc, 1995.

Shiver, B. D., and Borders, B. E., 1996. Sampling Techniques for Forest Resource Inventory. John Wiley and Sons Inc., New York.

Smith, J. and Heath, L. S., Identifying Influences on Model Uncertainty: an application using a Forest Carbon Budget Model. *Environmental Management* 27: 253-267.

Stephens, M. A., 1986. Tests based on EDF statistics. In Goodness-of – fit Techniques (R. B. D'Agostino and M. A. Stephens eds.), Marcel Dekker, New York.

Winiwarter, W. and Rypdal, K., 2001. Assessing the uncertainty associated with national greenhouse gas omission inventories: a case study for Austria. *Atmospheric Environment* 35: 5425-5440.

Winston, W. L., 2001. Simulation Modelling using @RISK. Duxbury Press.

APPENDIX

Relevant sample survey techniques include sampling with partial replacement and double sampling (De Vries, 1986; Shiver and Borders, 1996). There are several variants of double sampling including ratio rather than regression estimators and/or incorporating stratification. The variance formulae needed to part estimate uncertainty are given in Shiver and Borders, 1996. A common application of double sampling is utilizing a straight-line regression estimator without stratification. The variance of the mean (response variable) is given by:

$$s^2_{\bar{y}Rd} = s^2_{y.x}\left(\frac{1}{n_2} + \frac{(\bar{x}_1 - \bar{x}_2)^2}{\sum x^2}\right)\left(1 - \frac{n_2}{n_1}\right) + \frac{s^2_y}{n_1}\left(1 - \frac{n_1}{N}\right)$$

where,

n_1 = number of observations in the large sample
n_2 = number of observations in the sub-sample
N = number of sample units in the population
$\bar{x}_1$ = large sample mean
$\bar{x}_2$ = small sample mean
s^2_y = sampling variance of y, the response variable
$s^2_{y.x}$ = error mean square of the regression.
$\sum x^2 = \sum X^2 - (\sum X)^2$

An example of acalculating the variance of average volume/ ha using a regression estimator and double sampling (adapted from Freese, 1962) follows.

In 1995, two hundred (200) sample plots of 0.1 ha in a 20000 ha forest showed a mean volume of 10.52 m^3/ plot. A sub-sample of 40 plots was subsequently re-measured in 2000.

The mean of the 40 plots in 1995 = 10.46 m^3/ plot
The mean of the 40 plots in 2000 = 13.32 m^3/ plot

The data are illustrated in Figure 1. The correlation coefficient = 0.94.

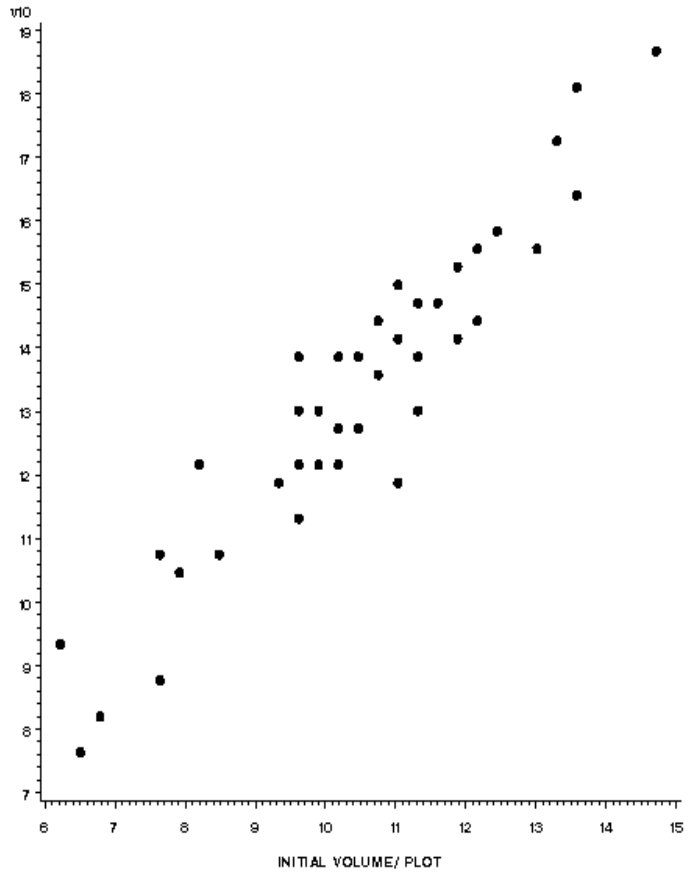

Figure 1: *Volume/ plot measured five years apart*

To work out the average volume/ plot at the second occasion and it's error, we have (from section 1.7)

$n_1 = 200$, $N = 200\,000$, $\overline{x}_1 = 10.52$, $n_2 = 40$, $\overline{y}_2 = 13.32$, $\overline{x}_2 = 10.46$

The corrected SS for the data are:

$$\sum x^2 = 154.27 \qquad \sum y^2 = 242.23 \qquad \sum xy = 182.21$$

so the regression coefficient ß = 1.1811

$$s_y^2 = 242.23/39 = 6.2110$$

and $s_{y.x}^2 = (\sum y^2 - (\sum xy)^2 / \sum x^2)/(n-2)$

$= (242.23 - (182.21)^2 / 154.27)/38 = 0.7111$

so the double sample estimate of the volume/ plot on the second occasion is

$$\bar{y}_{Rd} = \bar{y}_2 + b(\bar{x}_1 - \bar{x}_2)$$

$$=13.32 + 1.1811(10.52 - 10.46)$$

$$= 13.39 \text{ m}^3\text{/ plot}$$

The variance of this estimate (from above) is:

$$s^2_{\bar{y}Rd} = s^2_{y.x}\left(\frac{1}{n_2} + \frac{(\bar{x}_1 - \bar{x}_2)^2}{\sum x^2}\right)\left(1 - \frac{n_2}{n_1}\right) + \frac{s^2_y}{n_1}\left(1 - \frac{n_1}{N}\right)$$

$$= 0.7111\left(\frac{1}{40} + \frac{(10.52 - 10.46)^2}{154.27}\right)\left(1 - \frac{40}{200}\right) + \frac{6.211}{200}\left(1 - \frac{200}{200000}\right)$$

$$=0.7111[(0.025 + 0.00002333)(0.8) + 0.031055(0.999)]$$

$$= 0.7111[0.0200187 + 0.0310239] = 0.03629$$

so the standard error of the mean = ± 0.1905 m^3

Note that if the auxiliary information of the initial plot estimates had not been utilized then the variance is:

$$s^2_{\bar{y}} = \frac{s^2_y}{n_2}\left(1 - \frac{n_2}{N}\right)$$

$$= (6.211/40)(1 - 40/200\,000)$$

$$= 0.1552$$

giving a standard error of ± 0.3939 m^3, that is, the uncertainty estimate is doubled.