Panel data validation using cross-sectional methods

Lance R. Broad[a] and Ted Lynch[b]

Abstract

An examination of the suitability of an Irish Sitka spruce research panel data set for growth modelling purposes was undertaken. The panel data set arose from several repeated measurements in replicated field experiments. When being considered as data for yield modelling several difficulties arise. Simple histograms comparing sampled plots and the underlying forest estate demonstrated a sampling imbalance - whereby site index classes for sampled plots misrepresented the population. The spatial proximity of established plots also meant there was a lack of randomisation at the plot level, which eroded statistical independence between plots and increased plot cross-correlations. However, the availability of independent, non-research volume data permitted the construction of stand-level volume equations for both research and non-research stands. Observed differences in volume equation residuals for research thinned and unthinned stands were then explored. Thinning effects, volume equation inadequacy, or other sampling biases were considered as potential candidates to explain residual differences. It was found that the differences were consistent with a form of sampling bias when measuring volume sample trees. These validation techniques have led to a better understanding of the research data set.

Keywords: Growth modelling, panel data, Sitka spruce, validation.

Introduction

Permanent sample plots are the most common source of data for studies concerning forest growth and yield. Data from permanent sample plots have both spatial and time components. The spatial component is associated with the location of plots within a forest estate, which collectively give rise to a cross-sectional sample at a given point in time. The temporal component refers to repeated measurement of permanent sample plots, giving rise to a longitudinal sample. Data from permanent sample plots that combine cross-sectional and longitudinal samples are known as panel data. Forestry panel datasets tend to be more complicated than those found within econometric literature. The reason is that the longitudinal sections, or time between samples, in forestry data sets are often not evenly spaced.

Standard validation of forestry panel data usually progresses via a range of *ad hoc* computer based techniques that examine both longitudinal attributes, such as the time series associated with individual tree diameters in a plot, and cross-sectional aspects, such as rates of taper along volume sample trees. Standard validation tests tend to mirror the structure of the data, in that checks perform tests on the cross-sectional and longitudinal portions of the panel data. Validation of data is usually performed prior to their use in the fitting of equations.

[a] Technical Forestry Services, New Zealand.

[b] Coillte Research and Development, Newtownmountkennedy, Co Wicklow, Ireland (corresponding author, ted.lynch@coilte.ie).

Obtaining reliable panel data relies as much on good measurement practice as it does on a set of validation tests. Good measurement, data entry and transcription practices (when used), can help ensure a clean run through validation testing. However, it is not generally possible to formulate one set of validation tests that is sufficiently comprehensive to constitute the definitive set of tests for all intended use of forestry panel data. Consequently, validation tests tend to be specific, with consideration given to the prevalent form of stand management.

The standard type of validation tests, focusing on cross-sectional and longitudinal attributes, are not directed towards detecting forms of sampling bias that are associated with the initial creation of the permanent sample plots. Even the cross-sectional aspects of such testing are typically conducted within-plot and are not focused towards using inter-plot information. Similarly, longitudinal testing is generally conducted within-plot. However, many forms of sampling bias arise through a lack of appropriate randomisation. Consequently, bias detection requires an assessment of the underlying sample survey design structure to determine how randomisation issues have been approached and implemented.

This work sets out an investigation of the suitability of an Irish Sitka spruce panel data set for growth modelling purposes. This required an investigation the data set for the presence of sampling bias, through deploying both graphical and analytical techniques. The investigation was performed largely, although not totally, at the cross-sectional level.

Data and methods

Data provenance

Coillte Teoranta (the Irish Forestry Board) maintains the most extensive database on Sitka spruce (*Picea sitchensis* (Bong.) Carr.) in the Irish Republic. Initial measurements on research permanent sample plots used imperial units. In 1972 the metric system was adopted; subsequent measurement continued in metric, the original data were not generally converted. The database includes many silvicultural thinning and spacing trials that have been conducted during the period 1963 to 2001. Data used in modelling within this study were measured during the period 1972 to 2001. The database was computerised in 2000 using Microsoft Access®, thereby allowing investigation through database queries. Other code components have facilitated assembling of data required for growth and yield studies. Table 1 contains a summary of the Sitka spruce data.

In 2003, Coillte established within its production stands a small cross-sectional sample that consisted of volume data only. These are the non-research data within the study - and comprise some 70 observations from thinned stands and 43 from unthinned stands. The location of the non-research plots was determined using a systematic sample survey design. Data were collected according to generally recognised sampling practices – particular attention was given to the collection of volume sample trees. Although designed as a systematic sample it was treated as a random sample in the study. The prospect of there being any form of cyclical bias is considered remote, but since several of the plots were measured in the same dormant season there is a possibility of cross-plot correlation due to climatic effects.

Table 1: Coillte's Sitka spruce dataset.

	Sampling method	*Variable(s)*[1]	*Quality*	*Sample size n*
Non-research data (Thinned)	*Systematic Sampling*	*dbh, N, H and sectional volumes (cross-sectional data set)*	*Good*	*70 (volume)*
Non-research data (Unthinned)	*Systematic Sampling*	*dbh, N, H and sectional volumes (cross-sectional data set)*	*Good*	*43 (volume)*
Estate data (Thinned)	*Total enumeration*	*Site index (cross-sectional data set)*	*Good*	*27,887 (site index)*
Estate data (Unthinned)	*Total enumeration*	*Site index (cross-sectional data set)*	*Good*	*16,330 (site index)*
Research data (Thinned)	*Plots in replicated field experiments*	*dbh, N, H, and sectional volumes (panel data set)*	*Stat. ind. compromised*	*819 (volume)*
Research data (Unthinned)	*Plots in replicated field experiments*	*dbh, N, H and sectional volumes (panel data set)*	*Stat. ind. compromised*	*425 (volume)*

[1] *dbh: diameter at breast height (1.3 m above ground level)*
N: stems per plot
H: top height
Site index: top height (m) at age 30 (elapsed growing seasons since planting)

Further site index data were available at the sub-compartment level (the smallest unit of area that can be considered homogeneous for management purposes) across the entire Coillte estate. Consequently they give the most accurate site index representation possible for the estate[2], and are termed estate data.

Coillte's research data base was initially established as a set of spacing and thinning trials, which were mostly established using a randomised block design. Blocked plots were repeatedly measured, thereby creating a panel data set. While treatments were randomly assigned within blocks, having plots in close physical proximity to each other led to highly correlated growth model residuals.

The existence of the non-research and estate data sets allowed comparisons with the cross-sectional component of the research panel data set to be made. Elucidating the cross-sectional nature of the research data can, in some circumstances, also lead to an understanding of its longitudinal behaviour.

[2] At the time of writing Coillte's forest estate had 44,217 sub-compartments, some 27,887 managed as thinned stands with the remainder unthinned.

Graphical analysis

Volume datasets for thinned and unthinned plots were extracted from the research database: the thinned dataset had 819 plot-level volume observations, while the unthinned set had 425. The greater number of thinned observations was due to availability of observation data both before and after each plot was thinned. Site index was also available for each plot.

The volume samples were part of a larger data set that includes plot re-measurement data, allowing construction of growth trajectory data. The volume samples are a subset of those available within the research database. Specifically, they are the volume samples associated with the growth trajectory data. In this sense, they represent the cross-section associated with the panel data set.

An obvious requirement of any sampling design that seeks to provide data for yield modelling is that plots should exist over the range of site indices found within the forest estate where the model is to be applied. Similarly, the frequency with which plots appear in site index classes should, by and large, be the same as the frequency that estate stands appear in that site index class. Conformity to these requirements is easily checked by plotting appropriate site index graphs. Site index histograms for thinned and unthinned stands were therefore plotted for research and estate data (Figure 1).

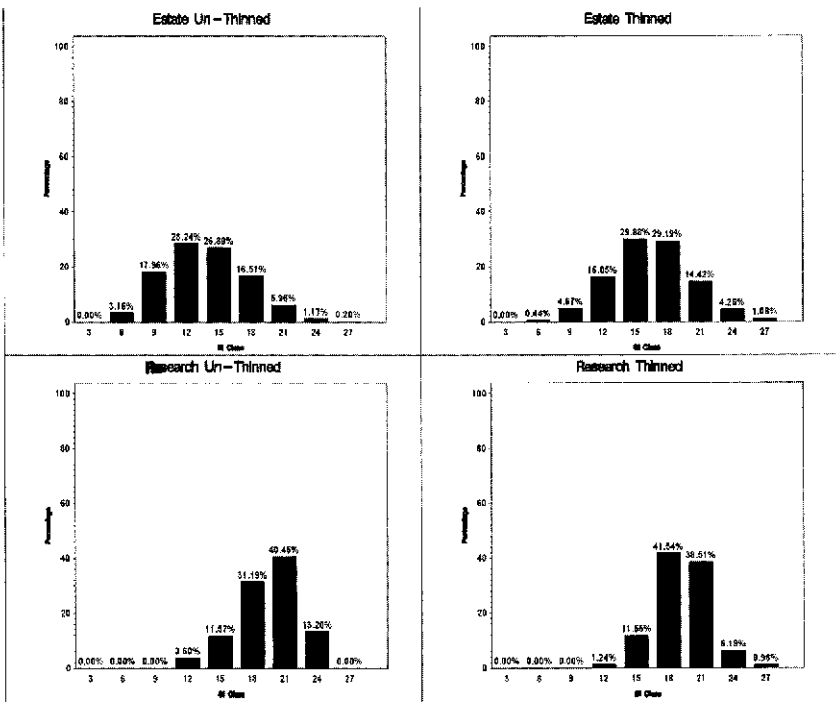

Figure 1: Site index distributions for (clockwise from bottom left) Research Unthinned, Estate Unthinned, Estate Thinned and Research Thinned.

It is apparent from a comparison of research and estate site index histograms in Figure 1 that research plots have been located predominantly in higher site index stands and there were no research (thinned or unthinned) plots in the two lowest site index classes that occurred within Coillte's estate.

Corroborative evidence as to the difference in site index distribution between research and estate data was obtained by testing the multinomial hypothesis $H_0 : p_i = p_i^o$ versus $H_0 : p_i \neq p_i^o$ for some i. Here p_i^o are the proportions appearing in each site index in the estate histogram of Figure 1, while p_i are the corresponding proportions appearing in the research histogram. The test was conducted using Pearson's chi-squared test with test statistic

$$Q_k^0 = \sum_{i=1}^{k+1} \frac{(N_j - np_j^o)^2}{np_j^o}$$

having k degrees of freedom (Mood et al. 1974). Strictly, the test requires independence of the underlying multinomial trials, a requirement that was not fully met due to the layout out of research plots within the experiment design blocks. Performing the test for the thinned crop was $P(\chi^2_7 > 625.5) < 0.001$ and for the unthinned crop $P(\chi^2_7 > 1648.2) < 0.001$. Figure 1 and the chi-squared tests therefore jointly indicated that the site index distribution within the experiment plots was not reflective of site index across the Coillte estate.

The graphical analysis identified that the selection of sites for field experiments was biased towards higher productivity sites. Site index is not influenced by thinning practice (inclusive of no thinning) when thinning is conducted from below. Consequently many of the plots in experiments would have similar site indices due to the spatial proximity of the plots. This also suggests a lack of randomisation at the plot level.

The overall sampling design for the experiments has not been recorded. It is highly unlikely that the pattern of site indices indicated within the research data of Figure 1 would emerge from a sampling design based on a simple random sample of plots. More sophisticated, randomly based plot sampling designs, such as sampling via a probability proportional to size within a site index class, would serve to lessen the probability of observing the outcome associated with the research data in Figure 1. These statements must be qualified, in that the establishment of permanent sample plots, for reasons of limitation of resources, generally takes place over a number of years. Forest estates can change over time with respect to their site index distributions, depending on factors such as soil fertility, effects of multiple rotations, and land transfers to, and from, other land uses.

A number of methods are available to address data imbalance. Provided data exist within all site index classes, weighting can be used to when fitting models. However, where classes are not represented in the sample, then data imbalance can only be addressed by using additional data obtained by further sampling, or by use of data from an external source.

Quantitative techniques for examining a panel dataset are applicable when either an independent panel or a cross-sectional dataset exist; these allow equation fitting

and model comparisons to be made. Here, volume equations over the research and non-research cross-sectional data were used to further examine the cross-sectional structure of Coillte research database.

Volume equation analysis

The volume equations used come from a general class of volume/basal area quotient equations introduced by García (1984). The equations are fitted in quotient form so as to improve error variance homogeneity. Their general form is

$$\frac{V}{B} \cong \beta_0 + \sum_{i=1}^{n} \beta_i \, g_i(B, N, H, S) \qquad (1)$$

Where

β_0, β_1, ..., β_n is a set of parameters, $g_i(B, N, H, S)$ is a predictor formed from its arguments basal area B (m^2 ha^{-1}), stocking N (stems ha^{-1}), top height H (m) and site index S (top height (m) at age 30), through the operations of multiplication, division, taking a power or multiplication by a constant.

Here site index (S) may be used to account for any site effects (García personal communication). Provided the predictors are formed as indicated above, the resulting equation will always be linear in its parameters. Stepwise regression is then a particularly useful technique for model identification and fitting purposes.

Fitting equation (1) to the combined research thinned and unthinned data resulted in the model specified in Table 2.

The model has an R^2 of 0.9627 and an adjusted R^2 of 0.9625.

The dynamic re-weighting scheme available within SAS® PROC REG (SAS Institute 1990b) was used to screen observations during model fitting. Observations having an absolute studentised residual of more than 3 were excluded from model fitting – this resulted in 14 observations being removed.

The model appears under a range of fitting diagnostics to be adequate but upon closer scrutiny of the volume predictions, biases are observed as described below.

Table 2: Sitka spruce research data – volume/basal area quotient regression.

Variable	*Parameter*	*Std Error*	*Pr > |t|*
Intercept	*4.11729*	*(0.3627)*	*< 0.0001*
H	*0.26687*	*(0.0144)*	*< 0.0001*
H/√N	*4.71047*	*(0.5277)*	*< 0.0001*
1/H	*-33.32451*	*(2.6576)*	*< 0.0001*
H/N	*-64.28313*	*(7.8327)*	*< 0.0001*
*100/B*H*	*1.00465*	*(0.3152)*	*0.0015*
S/B	*1.69270*	*(0.4542)*	*0.0002*
*S*S/B*	*-0.09036*	*(0.0139)*	*<0.0001*

Applying the equation specified in Table 2 to predict volumes (to 7 cm small end diameter (SED)) for each unthinned research observation results in the predictions of Table 3.

Here V7 denotes mean observed stand volume to 7 cm SED, V7hat is estimated mean stand volume to 7 cm SED, V7rsd denotes the mean residual when estimating to 7 cm SED. Finally, V7rpe is relative prediction error for V7 being calculated as the mean of quotients of the form (V7-V7hat)/V7. For the unthinned research data, the result indicates that there is a 2.33% over-estimation of volume to 7 cm SED as assessed by a mean relative prediction error (equivalently mean of ratios).

Applying the equation in Table 2 to predict volumes for the research thinned stands leads to the results in Table 4.

For the research thinned data there was a 0.37% under-estimation of volume to 7 cm SED via a mean relative predictive error statistic. It is apparent that the research equation in Table 2 is better at predicting the thinned data than the unthinned data. This is a reflection of the observation weighting employed. During fitting some 819 thinned and 425 unthinned observations were used.

Having observed the biases arising from the research volume equation the question arises as to whether they are statistically significant? The lack of statistical independence within the research data (see Table 1) precludes the direct testing of biases via volume equation parameters. Such an approach would test for common parameters for thinned and unthinned stand volume equations. A probabilistic analysis can, however, be conducted using a related set of findings. The same pattern of biases was observed when (uncalibrated) volume equations were fitted to research data from four additional species: Douglas fir, lodgepole pine, Norway spruce and Scots pine (unpublished work). In each case, the thinned volume observations were, on average, under-predicted, while the unthinned observations were over-predicted.

Table 3: Research unthinned data mean predictions.

Variable	*Observations*	*Mean*
V7	*425*	*364.6903*
V7hat	*425*	*373.5528*
V7rsd	*425*	*-8.8625*
V7rpe	*425*	*-0.0233*

Table 4: Research thinned data mean predictions.

Variable	*Observations*	*Mean*
V7	*819*	*312.3422*
V7hat	*819*	*309.9420*
V7rsd	*819*	*2.4003*
V7rpe	*819*	*0.0037*

With respect to each volume basal area quotient equation fitted over thinned and unthinned observations, the sum of the theoretical residuals may be partitioned as

$$\sum_i \varepsilon_i = \sum_j \varepsilon_j + \sum_k \varepsilon_k$$

where the first summation on the right is taken over the thinned observations and the second over the unthinned. The observed residuals on any volume basal area quotient equation are required to sum to zero. This follows from the normal equation associated with the constant parameter (see Seber 1977, p 47). A null hypothesis of there being no difference in the predictive ability of each volume equation over its respective thinned and unthinned stands requires the expected values of the sum of theoretical residuals be zero i.e.

$$E(\sum_j \varepsilon_j) = 0 \qquad \text{and} \qquad E(\sum_k \varepsilon_k) = 0$$

By appealing to the central limit theorem (Capinski and Kopp 1999) we can assume that the theoretical residual sums

$$\sum_j \varepsilon_j \qquad \text{and} \qquad \sum_k \varepsilon_k$$

are normally distributed. Normal distributions are symmetric and consequently the mean, median and mode assume a common value of zero. From normal distribution theory we can make the statement

$$P[\sum_j \varepsilon_j > 0] = 1/2$$

or equivalently in terms of the mean

$$P[n_1^{-1} \sum_j \varepsilon_j > 0] = 1/2$$

where n_1 is the number of thinned observations. A similar statement holds with respect to sum and mean for the unthinned errors.

The observed mean residuals for any volume basal area quotient regression, although theoretically correlated, should also tend to follow a normal distribution. Consequently, we anticipate the observed mean residuals for thinned and unthinned stands to be normally distributed with mean, median and mode of zero. It is important to note that the probability of $^1/_2$ that the observed residual sum, or mean, exceeds the median (50% quantile) is the same for each of the five volume equations. Given that all five equations examined had observed thinned mean residuals exceeding zero, then under the null hypothesis the probability of this event happening on the basis of chance is determined from the binomial distribution as . This differs markedly from the observation that 100% of trials had thinned mean residuals above their median value, and is strongly suggestive of the presence of bias associated with thinned and unthinned predictions.

The non-research data do not suffer from the same lack of statistical independence as do the research data. In terms of plot selection these data are known to have been collected in a manner that was virtually equivalent to a random sample. Volume sample trees were randomly selected across the range of merchantable trees. The non-research data consist of a volume cross-section over some 70 thinned plots and 43 unthinned plots.

In Table 5 results are presented from fitting a volume/basal area quotient regression to non-research volume observations.

The model was fitted over the non-research thinned and unthinned stand volume observations. It has an R^2 of 0.9627 and an adjusted R^2 of 0.9625.

The non-research volume equation allowed the construction of Table 6, which shows mean observed, predicted, residual and relative prediction error values for thinned stands.

The mean observed, predicted, residual and relative prediction error values for the volume equation for unthinned non-research stands are shown in Table 7.

The hypothesis model given in Table 5 can be tested against a maximal model, where separate volume equation parameters are fitted for thinned and unthinned stands. The test provides an indication as to whether separate regressions are required for volume predictions in thinned and unthinned stands.

Table 5: Non-research data – volume/basal area quotient regression.

Variable	*Parameter*	*Std. Error*	*Pr > |t|*
Intercept	*0.92527*	*(0.4201)*	*< 0.0001*
H	*0.41616*	*(0.0143)*	*< 0.0001*
*N*H/1000*B*	*-1.97212*	*(0.2799)*	*< 0.0001*
B/H	*0.19608*	*(0.0575)*	*< 0.0001*

Table 6: Non-research thinned data mean predictions.

Variable	*Observations*	*Mean*
V7	*70*	*394.2157*
V7hat	*70*	*393.9792*
V7rsd	*70*	*0.2364*
V7rpe	*70*	*-0.0042*

Table 7: Non-research unthinned data mean predictions

Variable	*Observations*	*Mean*
V7	*43*	*378.5654*
V7hat	*43*	*379.1120*
V7rsd	*43*	*-0.5466*
V7rpe	*43*	*-0.0050*

Table 8: F-test of common volume equation parameters for thinned and unthinned stands.

Source of variation	*degrees of freedom*	*sum of squares*	*mean square*
Resid. hypothesis model	*109*	*30.5236*	
Resid. maximal model	*105*	*29.5297*	*0.2812*
Difference for test	*4*	*0.9939*	*0.2485*

From Table 8 the quotient of the Difference mean square and the Maximal Model Residual mean square is 0.8835 and is distributed as $F_{4, 105}$. The test indicates that the null hypothesis of a common volume equation for thinned and unthinned stands cannot be rejected.

The conditions required for implementation of the test in Table 8 can be questioned (see Discussion). However, further support for the test result comes from the comparatively small magnitudes of the biases and the near zero relative prediction errors in Tables 6 and 7.

Discussion

The statistical test used for the non-research volume observations requires statistical independence between observations on plot volume. Also required is homogeneity of error variance. It is possible that the volume equation residuals would exhibit some degree of auto-correlation associated with within plot predictions over small time intervals in the absence of thinning. Cross-correlation effects between plots arise largely through climatic effects and are most pronounced when plots are re-measured in the same year. Factors that should reduce the impact of the correlation structure are longer periods between subsequent re-measurement of plots and the fact that not all plots are measured in the same year. The effect of the correlation structure has been ignored here.

A more comprehensive analysis would involve building an error component model that included serial- and cross-correlation terms and testing the significance of these effects. In a study of error component models for forestry yield models Gregoire (1987) found that ordinary least squares had lower prediction errors than models fitted with error components attached. Gregoire's work suggests that it may be difficult to formulate appropriate error component models.

The biases found in the research data volume predictions are likely to be consistent with any of the following causes:

1. there is a positive thinning effect which occurs when thinning is properly performed,
2. some form of volume equation inadequacy (other than a thinning effect or sampling bias which can lead to volume equation difficulties) or
3. there is a volume sampling bias.

It is possible that these causes may act in concert. However, the investigation was restricted to single effects under the assumption – that, at most, one of these causes is active. This restriction is very stringent but it does facilitate discussion of the

possible causes. It can be argued that the likelihood of more than one cause being active is small.

Any thinning effect is envisaged to act through facilitating stem diameter growth above the point of dbh measurement, and result in higher stem volumes for thinned stands. The mechanism for any such effect is envisaged as a photosynthetic response of the remaining tree crowns to higher light intensity following removal of competition in thinning. Thinning in the research plots was closely specified and supervised to ensure consistent levels of thinning over plots.

If a thinning effect were the cause of the differences observed in the research data then it should also express itself within the non-research data and be capable of detection using statistical tests. The fact that no statistical difference was observed between thinned and unthinned volume equations for the non-research data suggests however that a thinning effect cannot be active.

If there are inadequacies in the volume equation that led to the observed differences in the research data then these should be apparent when fitting volume equations to the non-research data and again be capable of detection using statistical tests. The lack of statistical difference between thinned and unthinned volume equations for the non-research data suggests that the argument for volume equation inadequacy is not strong. Further support for the volume equation comes from this class of equation being widely used for both thinned and unthinned stands without encountering bias problems (García 1984, 2003).

The suggested cause of additional volume sampling bias is most likely associated with the subjective selection of larger volume sample trees. Selecting larger volume sample trees leads to a larger mean volume per tree, which in turn leads to, a larger plot volume or volume per hectare, when multiplied by the appropriate stocking. Stated alternatively, any selection bias towards larger volume sample trees will be reflected in the parameters of the individual tree volume versus basal area straight line (regression) used to calculate mean volume per tree, and also in the parameters of the volume/basal area quotient regression used to calculate volume per hectare.

Plot location could also act as a form of sampling bias contributing to the differences in volume equation residuals. Many of the experiments used as data sources were established using replicated experiment designs. Any plot selection bias in this instance would be expressed at the block level (block selection bias) and possibly even the experiment level.

If a sample bias occurred with respect to research data it would not be anticipated to occur with non-research data. Plot selection for the non-research data used a systematic sampling design. Although such designs are capable of admitting cyclical forms of bias, their behaviour is expected to closely parallel a fully randomised design. Moreover, the volume sample trees for the non-research plots were selected across the range of observed diameters, thus ensuring acceptable determination of the parameters in the individual tree volume versus basal area regression.

The statistical test conducted on the non-research volume equation is consistent with a sampling bias not being present, but suggests a sampling bias in the research volume data.

Although this analysis suggests a bias in the research volume data it does not provide a definitive indication as to its source. In an endeavour to trace the source of the bias the relative prediction errors for the research thinned data were calculated and subsequently sorted. Of particular interest were observations with observed values (plot volumes as determined through sectional measurement and volume/basal area regression) being under-predicted by the volume equation, as these may suggest that the observed values are too high. These observations were examined to determine the range of the breast height diameters associated with the volume sample trees compared with the breast height diameter range for the crop trees (Table 9).

Table 9: Crop and volume sample breast height diameter range data[1].

		Sample		*dbh (cm)*		
Plot	*Year*	*Crop*	*Volume*	*Min*	*Max*	*Range*
CCA057901	*1979*	*MC*		*6.4*	*20.2*	*13.8*
	1979	*TH*		*7.2*	*16.0*	*8.8*
	1979		*TH*	*11.1*	*15.2*	*4.1*
CCA057910	*1978*	*MC*		*7.2*	*18.5*	*11.3*
	1978	*TH*		*7.5*	*17.5*	*10.0*
	1978		*TH*	*9.4*	*14.8*	*5.4*
CCA057902	*1978*	*MC*		*5.7*	*16.8*	*11.1*
	1978	*TH*		*8.0*	*17.5*	*9.5*
	1978		*TH*	*9.9*	*14.5*	*4.6*
CCA057909	*1978*	*MC*		*8.2*	*17.8*	*9.6*
	1978	*TH*		*8.1*	*17.5*	*9.4*
	1978		*TH*	*10.1*	*15.2*	*5.1*
CCA057906	*1978*	*MC*		*5.9*	*17.8*	*11.9*
	1978	*TH*		*7.1*	*16.9*	*9.8*
	1978		*TH*	*9.6*	*15.9*	*6.3*
CAL018109	*1980*	*MC*		*2.1*	*16.9*	*14.8*
	1980	*TH*		*7.7*	*13.8*	*6.1*
	1980		*TH*	*8.0*	*14.1*	*6.1*
CAL018108	*1980*	*MC*		*7.3*	*15.3*	*8.0*
	1980	*TH*		*3.5*	*13.5*	*10.0*
	1980		*MC*	*9.0*	*13.9*	*4.9*
	1980		*TH*	*8.3*	*12.2*	*3.9*

[1] *MC (main crop), TH (thinning)*

The results in Table 9 were obtained from the first 11 plot volumes considered after sorting on relative prediction error. It is apparent that the diameter range associated with the volume sample trees is generally small when compared with the range of diameters associated with either the main crop or the thinning trees. In addition, an examination of the dbh measurements for the volume sample trees invariably shows that they have been sampled towards the upper end of the plot distribution. The first six plots in Table 12 contain information for the first year of neutral systematic thinnings. The volume sample trees are for the thinned crop portion, but in this context they are also used to calculate volumes for the main crop. The lack of adequate range for dbh measurements illustrated above is not confined to neutral thinning schemes as the last observation in Table 9 illustrates. Nor is it confined to thinned plots as the same problem can be identified in unthinned plots.

The impact of sampling volume sample trees over a restricted range can be understood by examining the variance associated with parameter estimates in the tree volume v basal area regression. The equation has form:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad E(\varepsilon_i) = 0 \qquad E(\varepsilon_i^2) = \sigma^2$$

Where the dependent variable y_i denotes tree volume and the independent variable x_i denotes tree basal area. The variance associated with the intercept parameter is

$$Var(\hat{\alpha}) = \frac{\sigma^2 \frac{1}{n} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

and that of the slope parameter is

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Parameter standard errors may be obtained by taking square roots of the above expressions. Larger standard errors are directly linked to sampling over a restricted dbh range as they become larger as the denominator expressions become smaller. The denominators are an expression that depends on the range of the sampled data. The denominators can be maximised by taking data at the extremes of the ranges. Choosing half the values at each end of the range results in a D-optimal design (Seber 1977). Although sampling at either extreme of the range results in a D-optimal estimation of the slope and intercept parameters, this approach faces a difficulty within a forestry setting. Merchantable trees have a limiting lower dbh and non-

merchantable trees may fall below this. It is sensible to restrict the volume sample trees to merchantable trees. At each end of the merchantable dbh range there may be abnormal trees either suffering from suppression at the lower end, or having enhanced upper stem diameters at the upper dbh range. Consequently, a more pragmatic approach may be to select volume sample trees randomly across the merchantable range. The important point is to sample across the merchantable range.

Given the information to hand, therefore, it is possible to conjecture that the observed biases found in volume equation residuals arise through taking volume sample trees towards the upper end of the dbh distribution. For thinned stands, it is envisaged that the selection has also been biased towards trees with enhanced upper-stem diameters. This conjecture could account for the pattern of observed biases.

A further issue relating to Coillte's database is the intention of the original design versus its intended use in a yield modelling setting. The database is aggregated, in the main, from repeated measurements in replicated field experiments that were originally designed to examine thinning and spacing effects. These designs are capable of analysis using mainly analysis of variance techniques (e.g. split-plot or repeated measures analysis) for analysis of cumulative and incremental growth. In the experiment design, treatment combinations are randomised to plots, within homogenous blocks. This allows more accurate and precise treatment comparisons to be made, as biases and error variances are reduced. In the context of yield modelling, however, longitudinal observations are made on plots that are treated as being statistically independent. True statistical independence will never hold in a forestry context as different plots are always subject to similar climatic effects over any given time period.

Even if the established set of experiments were randomly located across the Irish forest estate, the set of plots within them would not be. Blocking plots in the experimental design setting requires homogeneity within blocks – a desirable feature when detecting treatment differences. Whereas blocking, in terms of yield modelling, restricts randomisation at the plot level, increases the cross-correlation between plots, and further erodes statistical independence between plots, which is already compromised through climatic effects. Parameter estimation techniques for modelling growth equations are also compromised in that the vast majority of growth component fitting techniques are developed around the assumption of statistical independence between plots.

The sampling biases considered impact on the panel data set in different ways. The impact of selecting volume sample trees over a limited diameter range is purely cross-sectional in nature. In terms of fitted models this would affect volume and assortment equations. The impact of blocking plots impacts on all fitted models as does the absence of data identified through graphical analysis.

These volume sample tree result has been deduced by undertaking comparative statistical tests and making reasonable assumptions as to how those tests should behave between research and non-research data sets. Detection of the plot cross-correlation result arises through appreciating the different requirements in assembling data for treatment comparison and yield modelling purposes.

It is important to realise that the data omission problem and the increased plot cross-correlation issue are not associated with the set of experiments, per se, rather they arise when the repeated measurements are aggregated across experiments to form a panel data set - the experiment data have only been shown to be deficient with respect to the estimation of plot volumes. Further, aggregation across valid repeated measures experimental design data does not guarantee the creation of a valid panel data set useful for yield modelling purposes.

The prospect of a thinning effect giving rise to increases in upper-stem diameters would have implications for growth modelling in that it suggests some measure of upper-stem development would be required to fully account for stand structure and growth. If higher upper-stem diameters were observable through cross-sectional sampling then there should be some associated behaviour in the longitudinal direction.

Growth modelling theory suggests that growth increments for plots with higher upper stem diameters should be enhanced. This follows from the fact that most univariate growth models can be decomposed multiplicatively with the structure in (2), which indicates that the *increment* will be proportional to the size, at any specified time. Consequently, it could be anticipated that greater growth increments would arise from plots exhibiting higher upper stem diameters.

$$\frac{d\,\mathrm{y}}{dt} = \mathrm{f}(t)\,\mathrm{y} \qquad (2)$$

Where y is some measure of size and f(t) is a declining, or an eventually declining, function of time, so as to provide a declining relative growth rate (y'/y).

Models that can be classified in this way include Bertalanffy-Richards, Gompertz, Levakovic I, Levakovic III, Korf and Sloboda (see Table 1, Zeide 1993).

Conclusion

In assessing Coillte's database for yield modelling purposes three forms of sampling bias were identified. The first, through graphical analysis, indicates that Coillte's research data for Sitka spruce contains a sampling bias that omits lower site index material. The second, identified through the analysis of volume equation predictions, is associated with volume sample tree selection. The third is experiment blocking, which reduces randomisation at the plot level, increases cross-correlation between blocked plots, leads to the eroding of statistical independence between plots and adversely affects parameter estimation in yield modelling equations.

In the context of fitting growth models the data omission issue influences all types of equations in any growth modelling system. The volume sample tree bias impacts on only the volume related components in a growth modelling system – these are the volume and assortment equations. The lack of randomisation at plot level typically impacts on all types of equation within any growth modelling system. In the context of analysing the established experiments as experiment designs, only the volume sample tree bias is of concern. The data omission and randomisation

issue only arise once yield modelling is contemplated and data aggregation ensues. Clearly, the aggregation of data across repeated measures experiments has led to unanticipated consequences that have impacted negatively on the data requirements for yield modelling.

Acknowledgements

The authors acknowledge comments from a referee.

References

Capinski, M. and Kopp, E. 1999. *Measure Integral and Probability.* Springer Verlag, London.

García, O. 1979. Modelling stand development with stochastic differential equations. In Elliott D.A. (Comp.), *Mensuration for management planning of exotic forest plantations,* p 315-333. FRI Symposium No. 20. New Zealand Forest Service.

García, O. 1984. New class of growth models for even-aged stands: Pinus radiata in Golden Down's forest. *N.Z. J. For. Sci.* 14:65-88.

García, O. and Ruiz, F. 2003. A growth model for cucalypt in Galicia, Spain. *For. Ecol. & Mgmt.* 173:49-62.

Gregoire, T. G. 1987. Generalized error structure for forestry yield models. *For. Sci.* 33:423-444.

Mood, A.M., Graybill, F.A. and Boes, D.C. 1974. *Introduction to the Theory of Statistics,* 3rd ed. McGraw-Hill, Boston.

SAS Institute, 1990a. *SAS® Procedures Guide,* 3rd ed., v6.0.

SAS Institute, 1990b. *SAS/STAT® User's Guide*, Vol. 2, 4th ed., v6.0.

Seber, G.A.F. 1977. *Linear Regression Analysis*. John Wiley & Sons, New York.

Zeide, B. 1993. Analysis of growth equations. *For. Sci.* 39:594-616.