Prediction of Forestry Planned End Products using Dirichlet Regression and Neural Networks

Ciarán Hickey[1], Stephen Kelly[1], Paula Carroll[1], and John O'Connor[2]

[1] Centre for Business Analytics, School of Business, UCD, Ireland. Contact: paula.carroll@ucd.ie

[2] Coillte, Ireland

Abstract We describe a set of non-parametric and machine learning models to forecast the proportion of planned end products (PEP) that can be extracted from a forest compartment. We determine which forest crop attributes are significant in predicting the product proportions (of sawlog, pallet, stake, and pulp) based on an Irish dataset supplied by Coillte, the Irish state forestry company.

Dirichlet regression and neural networks are applied to predict the product proportions and evaluated against a multi-variate multiple regression benchmark model. Based on predictive performance, the neural network performs slightly better in comparison to Dirichlet regression. However assessing the model logic and taking account of user interpretation, the Dirichlet regression outperforms the neural network. Both models are also compared to an existing rule based model used by Coillte. The non-parametric and machine learning techniques provided consistent reliable models to accurately predict the PEP proportions. The two proposed models extend the versatility of non-parametric and machine learning techniques to areas such as forestry.

Keywords

Forestry product forecast model, Dirichlet Regression, Artificial Neural Networks

Introduction

Forest management are facing new challenges caused, in part, by public expectation that forests provide a myriad of services along with products. Managing these expectations and services (e.g. flood control, habitat, water quality, etc.) requires techniques and models to support the process of sustainable management, (Burger, 2009). The introduction of Sustainable Forest Management practice in Ireland, new environmental regulations and planning restrictions have made optimal decision making a challenge for forest managers. Advanced non-parametric and machine learning methods provide an opportunity to allow forest managers to determine optimal solutions within the boundaries of regulations and constraints, (Peng, 2000). This paper demonstrates how such analytical techniques can be used to assist with complex forestry management problems.

Coillte is a state company which operates in forestry land based business producing renewable energy and panel products. It was established under the Irish Forestry Act 1988 and manages approximately 7% of the land

cover of Ireland (445,000 hectares). It holds a timber market share of between 80-85%, being the principal supplier of roundwood to sawmills in Ireland.

Coillte developed a rule-based model to predict the outturn (proportions) of products from a forest compartment based on the following parameters: harvest type, mean-diameter breast height (dbh), average tree volume and species. Named the planned-end-product (PEP) model, it predicts the proportions of sawlog, pallet, stake and pulp that could potentially be extracted from a forest compartment. Computations for the proportions are based on a set of tables in a Forestry Inventory System (FIS). It is a rule-based process similar to a decision tree where downgrade percentages (proportions) are determined in MS Excel using 'if-then' rules by forest managers. The output of the PEP model is a sales proposal (SP), i.e. the proportions of end products for each species in each sub-compartment. The PEP model is an integral part of the planning and decision making process. The sawmill industry relies on the forecast volumes and Coillte constructs its marketing and investment strategy around the forecasts.

Analysis by Coillte showed that actual harvest output varied significantly from the PEP model forecasts. Coillte management felt that the overall PEP predictions at an estate-level were accurate, but that stand-by-stand level precision of product outturn could be low. There was variation in the level of downgrade being carried out leading to a risk of crops being excessively downgraded and a potential under-optimisation of the crop. The process of downgrading the PEP forecasts could lead to inconsistent decision making across the different forest management areas (e.g. a forest manager in county Donegal might downgrade differently to a forest manager in county Mayo).

Predicting forest growth and outturn is an integral part of sustainable forest management, (Soares et al., 1995, Everingham et al., 2009). It facilitates better sales income forecasts and improves management process, (Murphy et al., 2010). Everingham et al., (2009) say "*Early and accurate crop forecasts offer substantial benefits to industry through increased profitability, better logistical arrangements and improved customer satisfaction*".

Coillte sought to develop a more accurate PEP proportion forecast model that would allow consistent decision making, and improve the efficiency of the planning process as a whole. This paper describes our work developing alternative PEP forecast models.

Prediction Approaches

Parametric statistical techniques are typically adopted by ecologists for analysis of the relationships between an observed response and a set of predictors in a dataset, (Hochachka et al., 2007). Aertsen et al.,

(2010) assess the strength of parametric and non-parametric methods in predicting site index, and conclude that multiple linear regression is the easiest and most straightforward tool to use, which with better data preparation can be a suitable technique. Lek and Guégan, (1999) suggest that non-parametric techniques can more accurately represent ecological data. Artificial Neural Networks (NNs) have been used successfully in forest science. Liu et al. (2003) compare the performace of NNs to traditional statistical methods for classifying forestry plots into ecological habitats. In subsequent work, Liu et al. (2005) describe the use of NNs and statistical techniques to predict product recovery based on black spruce tree characteristics.

There has been limited research on applications to forested ecosystems in Ireland. Both parametric and non-parametric models were considered as potential candidates. The next section covers the background to each modelling approach.

Multivariate-Multiple-Regression

One of the most widely used forms of modelling ecological data-sets is multi-variate linear regression (MLR). MLR is a simple method that is commonly used as a benchmark model for evaluating alternative non-parametric models, e.g. NNs. MLR models the relationship between independent variables (IVs) and two or more dependent variables (DVs). MLR is considered the benchmark technique for modelling growth and projected outputs. However it may not capture all the data and relationships accurately, (Soares et al., 1995). Increasing the number of IVs may improve the performance, but this improvement may not compensate for the cost in the degrees of freedom of including more variables and more importantly, would likely lead to overfit. Therefore a finite set of significant IVs must be determined as inputs to the MLR model. This can be achieved through methods like forward elimination, backward elimination, or combination of both with stepwise regression.

Underlying multi-variate procedures is the assumption of normality of the residual errors. That is, the errors of the regression models are normally distributed. When a distribution is normal, the values of skewness and kurtosis are zero. Kurtosis values above zero can indicate a distribution that is too peaked with short, thick tails, and kurtosis values below zero indicate a distribution that is too flat. Non-normal kurtosis will lead to an underestimate of the variance of a variable that either underestimates or overestimates the true value.

The errors should also be independent for each set of IVs and homoscedastic, that is, the errors vary by the same amount for all values of the IVs. These assumptions can be checked by examining the normality and homoscedasticity of the residuals arising in analyses involving prediction. Bradley, (1982) reports that design-based statistical inference becomes less and less robust as distributions depart from normality. It can lead to

misinterpretations of data and theoretical impasses which may produce misleading results, (Smithson and Verkuilen, 2006).

Appropriate transformations e.g. square root, log, and natural log can be applied to normalise the variables to attempt to improve normal distribution fit of the residual errors. We note that the IV's do not need to have a normal distribution. Regression is fairly robust against departures from normality. Only the error around the line of regression must be normally distributed. As long as the distribution of errors is not extremely different from normal, inferences will not be seriously affected. If the assumptions discussed above are not strictly followed for the dependent variables, it does not invalidate the analysis so much as weaken it. When MLR models fail to provide a good fit, generalised linear modelling techniques such as Dirichlet regression may provide a better fit.

Dirichlet Regression

Modelling the dependent variables as proportions is a challenging problem. A MLR approach is problematic as it does not take into account that the dependant variables are compositional, i.e. the DV proportions sum to 1. In our case, the outputs are the product proportions (Sawlog, Pallet, Stake, and Pulp) which sum to 1. The bounds of [0,1] are known as the unit variant constraint. MLR also assumes that the conditional expectation function is linear, but this does not hold for combination of skewness and heteroscedasticity in variables with scales bounded at both ends, (Smithson and Verkuilen, 2006).

Dirichlet regression (DR) models are regarded as a generalisation of beta regression models for more than two components, (Gueorguieva et al., 2008). They follow the same assumptions as beta regression but allow multiple DV's which is the case for Coillte's multiple outputs: sawlog, stake, pallet, and pulp. Beta Regression is particularly associated with compositional data, (Simas et al., 2010). Some examples of Dirichlet model applications are: market share analysis, soil composition, election forecasts, and household expenses composition. It is used in a disease history case described in (Carreras et al., 2012) of several mutually exclusive events with each row of the transition matrix modelled by a multivariate Dirichlet distribution to deal with uncertainty in transition problems.

DR is useful for modelling data consisting of multivariate positive observations that sum to the value of 1. Preliminary descriptive data summarisation and analysis of Coillte's response variables showed them to be compositional and skewed suggesting DR may be an appropriate method to create a model for predicting the multiple outputs of sawlog, pallet, stake and pulp.

It is justifiable to use the Maximum Likelihood Estimate (MLE) approach in Dirichlet model fitting since beta distributions form an exponential family that satisfy certain useful regularity conditions which help ensure MLEs exist and are well-defined, (Smithson and Verkuilen, 2006).

The domain of a beta-distributed DV is [0, 1], i.e., the DV lies in the closed unit interval. A logit link function "squeezes" the real line into the unit interval. Two link functions are used for the location parameter μ (the DV proportion to be predicted) and the precision parameter ϕ (the variability of the predicted beta distributed proportion DV).

This section gives some of the mathematical background to Dirichlet models. Let x_i be observations of the set of independent explanatory variables *IV* used to explain the location (proportion) parameter, μ_j. β_{ij} are the regression coefficients to be estimated. Then for every product proportion $j \in$ Products {Sawlog, pallet, stake, pulp}:

$$ln\left[\mu_j/\left(1-\mu_j\right)\right] = \beta_{0j} + \sum_{i\,\epsilon IV} x_i\,\beta_{ij}$$

$$i \,\epsilon\, IV\ (Aspect, Soil\ Type, Elevation, Slope, \ldots)$$

where β_{0j} is the intercept for the j^{th} product.

The logit transformation $ln\left[\mu_j/\left(1-\mu_j\right)\right]$ maps a number $\mu_j \,\epsilon\, (0,1)$ onto a real line. This link is desirable from an interpretational point of view because the regression coefficients obtained are log-odds. The regression coefficient can be transformed to probability odds by back transformation, i.e., taking the exponential of the regression coefficient as demonstrated in the Results section below. The logit link is compared against other link functions and explained in Cox, (1996).

The precision (variability) parameter is also linked via a log function and must be positive as a variance cannot be negative. Let w_i be independent explanatory variables to regress upon the precision parameter ϕ_j. δ_{ij} are the corresponding regression coefficients to be estimated. The design matrices *x* and *w* are separate and do not necessarily have to be disjoint.

$$\ln\left(\phi_j\right) = -\delta_{0j} - \sum_{i\,\epsilon\,IV} w_i\,\delta_{ij}$$

The probability (proportion) and variability values are obtained by inverting the link functions for each model:

$$\mu_j = \frac{\exp(\beta_{0j} + \sum_{i\epsilon IV} x_i\,\beta_{ij})}{1 + \exp(\beta_{0j} + \sum_{i\epsilon IV} x_i\,\beta_{ij})} \qquad \text{Location Model}$$

$$\phi_j = \exp(-\delta_{0j} - \sum_{i \,\epsilon\, IV} w_i\, \delta_{ij}) \qquad \begin{matrix}\text{Precision}\\ \text{Model}\end{matrix}$$

The *DirichletReg* package in R, (R Core Team, 2013), can be used to create a location model (also known as the mean model) for the proportion DVs based on $n-1$ IVs. The precision model is based on the remaining attribute. This allows models to conveniently account for over dispersion by including the precision parameter ϕ_j to adjust the conditional variance of the predicted proportion. The precision controls how concentrated the distribution is around the predicted proportion, (Huang 2012). When the precision is high, the proportion values are predicted over a narrower range giving a more precise indication of the true population proportion, μ_j. When the precision is small, the predicted proportion values are distributed more diffusely.

Neural Networks (NNs)

Certain data mining techniques can model a process without prior assumptions about the forms of relationships between independent and dependent variables, (Hopfield, 1984, Hornik et al., 1989). They offer a powerful and flexible way for exploratory analysis of ecological systems.

NNs are models inspired by an analogy of how the brain works. NNs do not require prior knowledge of the underlying process or structure of the target function. It is not necessary to pre-specify the type of relationship between covariates and response variables as for instance as linear combination, (Frauke and Fritsch, 2010). NN approaches can be used for ecological data because they provide a way to overcome typical difficulties that arise in handling forestry data, such as nonlinear relationships and non-normality, (Ito et al., 2008). One advantage of NN models is their ability to process biological data that contains complex correlations, (Basheer et al. 2000).

NNs employ machine learning algorithms to learn how a set of inputs are connected to a set of outputs. The algorithm creates a network modelling the connection of weighted inputs to the outputs via a set of hidden layers of nodes and transfer functions. This type of NN model is called a multilayer perceptron (MLP). A MLP can be considered as a non-parametric, non-linear regression model. The MLP predicts an output for each input vector and the error between the predicted and the actual value of the output is calculated.

Figure 1 is an unlikely but illustrative diagram of a possible MLP which includes all attributes at our disposal to produce the individual yield proportions. The first layer, the *input layer*, serves to hold the data being input to the MLP. This layer is connected to a *hidden layer* and the nodes in the hidden layer are connected in turn to an *output layer* which represents the processed output from the model. The optimal size of the hidden layer is not known prior to modelling and is determined heuristically by the modeller. Each of the connections (arcs) between the nodes has an associated real-valued weight which is similar in concept to a regression

coefficient. The value passing along a connection is modified by multiplying it by the weight before it reaches the next node. Generally, the processing carried out at each node in the hidden and output layers consists of passing the weighted sum of inputs to that node through a non-linear *transfer* function.

One common transfer function used in NNs is the logistic sigmoidal function. It transforms the output values into a differentiable, nonlinear function. It acts as a squashing function so as to keep the output values within specified bounds, (Müller and Mburu, 2009):

$$Logistic\ sigmoid(x) \coloneqq \frac{1}{1+e^{-x}}$$

From a modelling perspective, NNs are complicated as the number of hidden layers is difficult to specify. Typically, a sample of the data is used to train the network while the remaining data is then used to test the performance of the constructed NN. Wang et al., (2005) describe a series of trials to determine the number of layers and note that the NN model must be trained until a convergence criterion is satisfied. NN construction is a trial and error process for establishing a suitable model. Resilient back propagation (RPROP) is the most common way of altering the weights in response to an error during the learning phase. It is the most widely used algorithm for supervised learning with multi-layered feed-forward networks, (Riedmiller and Braun, 1993, Müller and Mburu, 2009). Chen and Pollino, (2012) explain that by applying different combinations of inputs and examining the resulting probabilities throughout the network, reviewers can test whether the behaviour of a model is consistent with current understanding about the system. A similar observation is given in (Aertsen et al., 2010), where the authors compare and evaluate five modelling techniques for predicting the site index of three different tree species located in the Taurus Mountains of Turkey. The number of hidden units depends in a complex way on the number of inputs and outputs, the number of training cases, the noise in the targets, the type of activation function and training algorithm used. Some authors feel that there is no structured way to determine the number of hidden units without training several networks and estimating the errors in each, (Sarle, 1995, Swingler, 1996).

Kaul et al., (2005) use MLR and NN models on the same crop yield dataset. They conclude that NN models, like regression models, are applicable only to the conditions for which they were developed. Ultimately the accuracy of their NN crop yield predictions are dependent on the learning rate and number of hidden nodes, which has a significant effect on the development of the model.

Müller (2009) describes a case study where a variety of ecology data are used for the input layer. They conclude that RPROP is a particularly suitable training algorithm for broad-scale investigations as it necessitates

less training runs and is more robust to the choice of initial parameters, as compared to the commonly used standard Backpropagation.

The main objective of our paper is to create a new model to improve PEP prediction and thus the product outturn. NNs provide a suitable approach having a strong predictive performance, (Aertsen et al., 2010).

Data Pre-Processing and Analysis

Data from Coillte's FIS covering some 5,616 closed SPs from 2000 to 2012 were extracted corresponding to approximately 7.9 million m^3 of timber. Only single species SPs were included in the sample as each SP was to be treated as a "sub-compartment" for analysis purposes. The attributes were both categorical and numerical: soil type, mean-dbh, elevation, aspect, harvest type, species, average tree volume, and slope.

Data pre-processing techniques can improve the quality of the data which helps to improve the accuracy and efficiency of the mining process. Data that is 'dirty' can cause confusion during the mining process and lead to unreliable output. Detecting these anomalies and rectifying them early can lead to huge payoffs for decision making, (Han & Kamber 2006). The Coillte data contained some anomalies such as these, exploratory data analysis was conducted to summarise and prepare the dataset for model building.

Statistical analysis was used to determine the relationships between parameters and to check if the assumptions of normality were valid. Central tendency and dispersion measures were calculated. Summaries are displayed in ***Table* 1** and *Figure 2*.

Local outlier factor (LOF) and principal component analysis (PCA) were used in the data pre-processing phase. Most outlier applications consider an outlier as a binary property, but LOF uses the relative density of an object against its neighbours as the indicator of the degree of the object being outliers. As most of the data in the dataset are not outliers, it is meaningful to identify only the top *n* outliers, (He et al., 2003). Attributes' multicollinearity was determined to indicate if the numerical methods to solve regression equations were appropriate. High multicollinearity affects the significance values and the confidence intervals for the regression coefficients.

MLR Methodology

We experimented with several transformations in constructing the MLR benchmark model and we found the MLR model to be more accurate with transformed data, but still not adequate. Several of the attributes were log-transformed to ensure normality. Descriptive statistics of the transformed variables were obtained and showed significant improvement in skewness, kurtosis. Stepwise MLR was carried out through SAS with this dataset, creating four separate linear equations, one for each respective DV. Recall, our aim is to use a MLR as a

benchmark model for comparison. Our approach above sought to find the best fitted MLR model for this dataset but as we will see below, the MLR model failed in predicting the product proportions from a logical perspective.

Dirichlet Regression Methodology

Recall that we have four DVs (proportion of sawlog, pallet, stake and pulp) that we wish to predict. For scales bounded by an interval [0,1], a suitable candidate for models is the Dirichlet distribution. Statistical analysis suggested that our DVs were conditionally beta distributed rather than Gaussian which can be seen in *Figure 2*. We see that the proportions are skewed. Dirichlet distributions handles heteroscedasticity and skewness effectively, (Smithson and Verkuilen, 2006). The predicted (proportion) values for sawlog and stake are concentrated towards the zero boundary of the interval [0,1], highlighting the positive skew of the DVs.

The significant variables were determined and the models were tested and trained via a 10-fold cross validation. A multicollinearity test was conducted and attributes with high correlation removed. During this phase, the attribute 'average tree volume' was removed as it is highly positively correlated with mean-dbh.

The *DirichletReg* package in R, (R Core Team, 2013), provides techniques for building Dirichlet models. The approached used models the location (proportion) μ_j and the precision ϕ_j separately, with $j \in Products\ \{Sawlog, pallet, stake, pulp\}$. The benefit of this approach is that the location model predicts the expected values and the precisions (dispersion ϕ_j) models the variances. The expected values μ_j are the most important output of the Dirichlet regression, as these can be compared directly with the actual proportions. The precision shows which dependent variables are susceptible to variance, and how each one behaves in their respective range.

All viable parameters were run through the Dirichlet regression model. IVs with a '$p\ value \geq 0.05$' level of significance were deemed not to be significant and were removed from the model as inputs. One redundant parameter was removed through backward elimination based on the *p*-value of the attribute at each pass.

Neural Networks Methodology

As the ecological dataset contained a broad range of distinct forest crop attributes, there was a high likelihood of non-linear and unstructured relationships. The outputs (DVs proportions) and inputs (IVs) relationship structures do not have to be predefined in NNs. This allowed more flexibility in the model building as the exact functional forms are controlled by parameters that are determined in the training process of the NN. The NN was built using the *neuralnet* package in R, (R Core Team, 2013), and is a MLP trained with an RPROP algorithm. The connections (arcs) are randomly assigned weights initially and as the RPROP algorithm trains the network to reduce the error, the weights are adjusted accordingly, thus the order of inputs being added is insignificant. The activation function was set to the logistic function, with the differentiable error function left at the default of 'sum of squared errors'.

Different combinations of potential attributes and varying numbers of hidden nodes and layers were tested to determine the best representation of the Coillte dataset, as there is no definitive structural form to build a MLP. Similar to MLR and DR, sampling is done to prevent over-fitting of the model. As the number of permutations of NN design parameters is large, three heuristics, summarised below, were used to explore the NN topology design space. The mean squared error (MSE), the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were recorded as performance measures. AIC is used to aid model selection of a statistical model, measuring the relative quality of the model. AIC and BIC deal with the trade-off between the complexity of the model and the goodness of fit of the model.

As the input variables are on different scales, the numerical attributes (i.e., mean dbh, slope and elevation) were standardised to ensure the contribution of all variables in the model. Three heuristics listed below were used to select the best NN topology for the Coillte dataset.

1. **Input Independent Variable Selection**: IVs from the training data-set were incrementally added to the NN model and the residuals were plotted relative to each proportion.
2. **Hidden Nodes**: Starting with a relatively low number of hidden nodes, the number was increased methodically for each trial over a range of $2 \rightarrow 9$ nodes. AIC and BIC measures were used to determine the models performance.
3. **Sample Size**: Different sized subsets of the training data ranging from 500 to 3,500 observations, in increments of 500, were randomly sampled.

In the IV Selection stage, the IVs *harvest clearfell*, *harvest first thinning*, *harvest second thinning*, *species SS/NS*, *species LP/LPS/OC*, *elevation*, and *mean-dbh* were selected.

The *Hidden Nodes Selection* heuristic showed that as the number of hidden nodes increased the performance measures and runtime increased, while the error metrics decreased. However the reduction in error with a larger amount of hidden nodes was seen to be an over fitted NN, with predicted proportion results sometimes having negative values. By the second heuristic, the number of hidden nodes to best train the NN was found to be four.

A random sample of 2,000 observations provided the lowest MSE in the sample size heuristic to determine the training sample size. Another random sample of 2,000 observations was taken and the NN was retrained. This was done to ensure stability of the NN. These heuristics determined the NN design parameters. The NN model was then built and run in R, (R Core Team, 2013).

Results

SPs from the year 2012 were used as the validation test set to evaluate the models. To avoid over fitting and overtraining the SP's for 2012 were originally extracted from the main dataset. The test set consists of 158 SP's.

Soares et al., (1995) believes that model evaluation should comprise of a critical appraisal of model logic as well as theoretical and biological realism of the model. Consideration of user interpretation must also be a factor as well as the fundamental goal of predictive performance. Qualitatively, the models were examined in terms of logic and from theoretical and biological views.

Although the multi-variate multiple regression produced similar outputs to the actual product proportions, it failed from a logical perspective. The logic of the model is not accurate in that it is susceptible to predicting negative proportion values at times and it does not follow the unit constraint for proportion as shown in *Figure 3.* A SP cannot have a negative proportion from a forest for a particular product as it is infeasible and unrealistic.

Both the Dirichlet regression and NN models were evaluated both qualitatively and quantitatively. Quantitatively, the models were assessed against a test set comprising of all the SP's for the year 2012. The goodness of fit, predictive performance, weights, standardized residuals, and variance were analysed.

Dirichlet Regression Results

Backward elimination was carried out for attribute selection with a significance level of 0.05. The significant attributes can be seen in *Table 2*. The percentages next to each individual regression coefficient are the percentage increase/decrease for a unit increase for the corresponding IV when other parameters are fixed. As noted above, each regression coefficient 'β_{ij}' is represented by the log odds it can be transformed to probability odds by back transformation taking the exponential of the regression coefficient.

For example, we see results for the sawlog proportion in the alternative model: Product: Sawlog, attribute: Sitka Spruce/Norway Spruce: $\beta_{11} = 1.1016$. Taking the exponential; $exp(1.1016) = 3.0089$. Therefore there is a *200.89%* increase for the proportion value of sawlog when the species of the tree is Sitka Spruce/Norway Spruce (holding all other IVs constant). We see in *Table 2* that for a unit increase in slope, there is a *1.1%* decrease in the proportion of Sawlog, *2.3%* increase in proportion of pallet, *3.9%* decrease in Pulp, and a *3.4%* decrease in stake. The positive and negative influences are indicated by the arrows beside the percentage values.

This table not only represents the influence of the regression coefficients but can be used as an insight for harvest managers into patterns not typically detected. For example, Sitka Spruce/Norway Spruce has a significant effect across the products of sawlog, pallet and stake.

Lodgepole Pine has the greatest effect on pulp when other parameters are kept fixed. If Lodgepole Pine is the selected species, there is a *327%* increase in the proportion for pulp for the alternative model. The logic of this result was verified by Coillte. They indicated that Lodgepole pine would generally have a high proportion of pulp, if not 100% pulp.

Neural Networks Results

The training process needed 30,398 steps until all absolute partial derivatives of the error function were smaller than 0.01, the threshold which was set. As mentioned previously consideration for the user interpretation must be considered a factor as well as the fundamental goal of predictive performance. The AIC and BIC were 286.53 and 555.36 respectively for the final NN.

Figure 4 shows the input (IVs) and output (DVs) nodes along with the internal structure of the trained NN, i.e. the network topology. The plot includes the trained weights determined by the transfer function (not visible in the figure).

The *neuralnet* package in R (R Core Team, 2013) calculates and summarizes the output of each node, i.e. all nodes in the input, hidden and output layer. Thus, it can be used to trace all values or signals passing the MLP for given IV combinations, (Günther and Fritsch, 2010). This aids in interpreting the network topology of the trained NN. It was used to calculate predictions for new IV combinations.

Model Comparison

Predictive performance is one of the essential evaluation criteria specified by Coillte. Future forecasts are continuously being carried out by management to determine the most suitable forest compartments to harvest in order to reach pre-defined product quotas. These forecasts and decisions to harvest particular forest compartments rely on the accuracy of the PEP model. Any significant improvement on the original model is considered a success.

Both Dirichlet regression and NN offered useful methods to predict the forest proportions. The line plots in ***Figure 5*** show the proportions predicted by the Dirichlet and NN models, with the actual proportions, for a sample SP from 2012. The Dirichlet and NN model predict the actual proportion more accurately with smaller error in difference than Coillte's rule based PEP model. The error can be inspected visually from the trace of the predicted and actual proportions.

The NN has a slight advantage if models are strictly based on predictive accuracy. The NN slightly outperformed the DR. Achieving this high accuracy can have other obstacles that vitiate its attraction. The

Dirichlet regression approach offers more when model logic is taken into account. *Table 3* summarises a comparison of the modelling approaches.

Tables similar to *Table 3* represent the influential attributes of the each model giving a more effective insight into the components of the model. Both the PEP model and NN exhibit weaknesses in this particular area. The interpretation of the NN hidden units/layers is not as straight forward as interpretation of regression coefficients.

Conclusion

From the comparison of models, we see that Dirichlet regression and Neural Network modelling approaches are efficient and reliable techniques for predicting the proportions of a SP, based on the predictive performance and metrics in *Table 3*. Both models demonstrated accuracy, consistency and robustness.

Using a quantitative evaluation approach, the NN was shown to be more accurate in predicting the proportion than Dirichlet. Three heuristics were used to select the NN topology design. This contributed more to the NN model building than a traditional trial and error process. It addressed the principal criteria that govern NN topology design to ensure a NN model best suited to the Coillte dataset.

From a qualitative viewpoint the NN may be more difficult for an organisation to adopt and implement due to its lack of transparency and the difficulty in interpretation the topology.

The logic of a Dirichlet regression model is a more straight forward method to understand. It offers the possibility of formulating the model within MS Excel which is a common application in the day to day work for forestry managers in Coillte.

Using an equation to model the proportions may be preferable to the current practice of 'lookup tables' followed by discretionary down-grading. Dirichlet regression not only offers an accurate prediction but it distinguishes patterns within the data that may not be obvious from other analyses. An influential parameters table can be produced which gives an insight into the influences of particular attributes on the product proportions. Charts like that in *Table 2* give an insight into the interactions of attributes and could assist in the deciding the most suitable species to grow in a particular forest compartment during forest plantations planning.

Further Research

There is potential to further develop the models proposed and implemented in this study. Additional research into the application of Dirichlet processes and generalised linear models to forest crop data may offer an innovative perspective and further insight into the roles that these attributes play in predicting end-product outturn.

Handling multiple dependent variables contributes to the complexity of the models. Various experimental analyses were conducted to justify the methods described in this paper. Further investigation into structured robust methods for preparing the data for analysis could enhance the model building process.

More experimental trials of various NN could also be carried out by the modeller. For example, a second hidden layer could be useful when trying to learn complicated target functions, particularly multimodal functions, i.e. those with multiple local maxima (peaks) and local minima (valleys).

References

AERTSEN, W., KINT, V., VAN ORSHOVEN, J., ÖZKAN, K. & MUYS, B. 2010. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecological modelling,* 221**,** 1119-1130.

BRADLEY, J. V. 1982. The insidious L-shaped distribution. *Bulletin of the Psychonomic Society*.

BURGER, J. A. 2009. Management effects on growth, production and sustainability of managed forest ecosystems: Past trends and future directions. *Forest Ecology and Management,* 258**,** 2335-2346.

CARRERAS, G., BACCINI, M., ACCETTA, G. & BIGGERI, A. 2012. Bayesian probabilistic sensitivity analysis of Markov models for natural history of a disease: an application for cervical cancer. *Italian Journal of Public Health,* 9.

CHEN, S. H. & POLLINO, C. A. 2012. Good practice in Bayesian network modelling. *Environmental Modelling & Software*.

COX, C. 1996. Nonlinear quasi-likelihood models: applications to continous proportions. *Computational Statistics & Data Analysis***,** 449-461.

EVERINGHAM, Y. L., SMYTH, C. W. & INMAN-BAMBER, N. G. 2009. Ensemble data mining approaches to forecast regional sugarcane crop production. *Agricultural and Forest Meteorology,* 149**,** 689-696.

FRAUKE, G. & FRITSCH, S. 2010. Neuralnet: Training of Neural Networks. *The R Journal*.

GUEORGUIEVA, R., ROSENHECK, R. & ZELTERMAN, D. 2008. Dirichlet component regression and its applications to psychiatric data. *Computational Statistics & Data Analysis,* 52**,** 5344-5355.

GÜNTHER, F. & FRITSCH, S. 2010. neuralnet: Training of neural networks. *The R Journal,* 2**,** 30-38.

HE, Z., XU, X. & DENG, S. 2003. Discovering cluster-based local outliers. *Pattern Recognition Letters,* 24**,** 1641-1650.

HOCHACHKA, W. M., CARUANA, R., FINK, D., MUNSON, A. R. T., RIEDEWALD, M., SOROKINA, D. & KELLING, S. 2007. Data-Mining Discovery of Pattern and Process in Ecological Systems. *The Journal of Wildlife Management,* 71**,** 2427-2437.

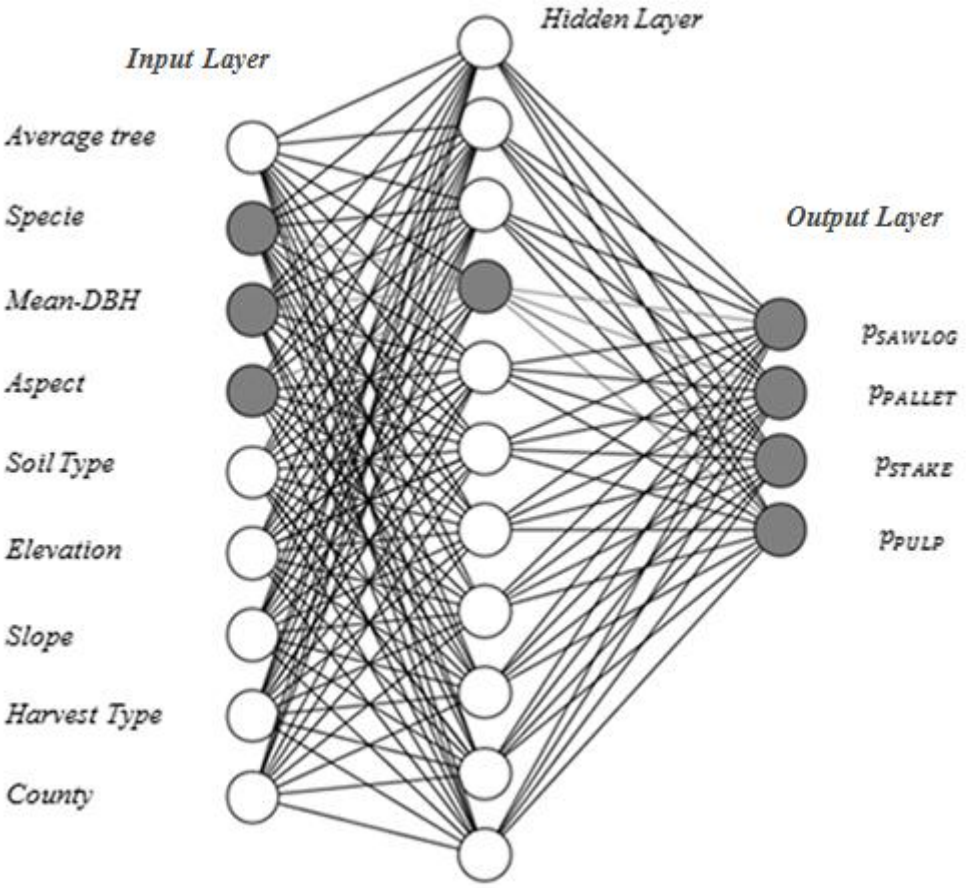
HOPFIELD, J. J. 1984. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences,* 81**,** 3088-3092.

HORNIK, K., STINCHCOMBE, M. & WHITE, H. 1989. Multilayer feedforward networks are universal approximators. *Neural networks,* 2**,** 359-366.

ITO, E., ONO, K., ITO, Y. M. & ARAKI, M. 2008. A neural network approach to simple prediction of soil nitrification potential: A case study in Japanese temperate forests. *Ecological Modelling,* 219**,** 200-211.

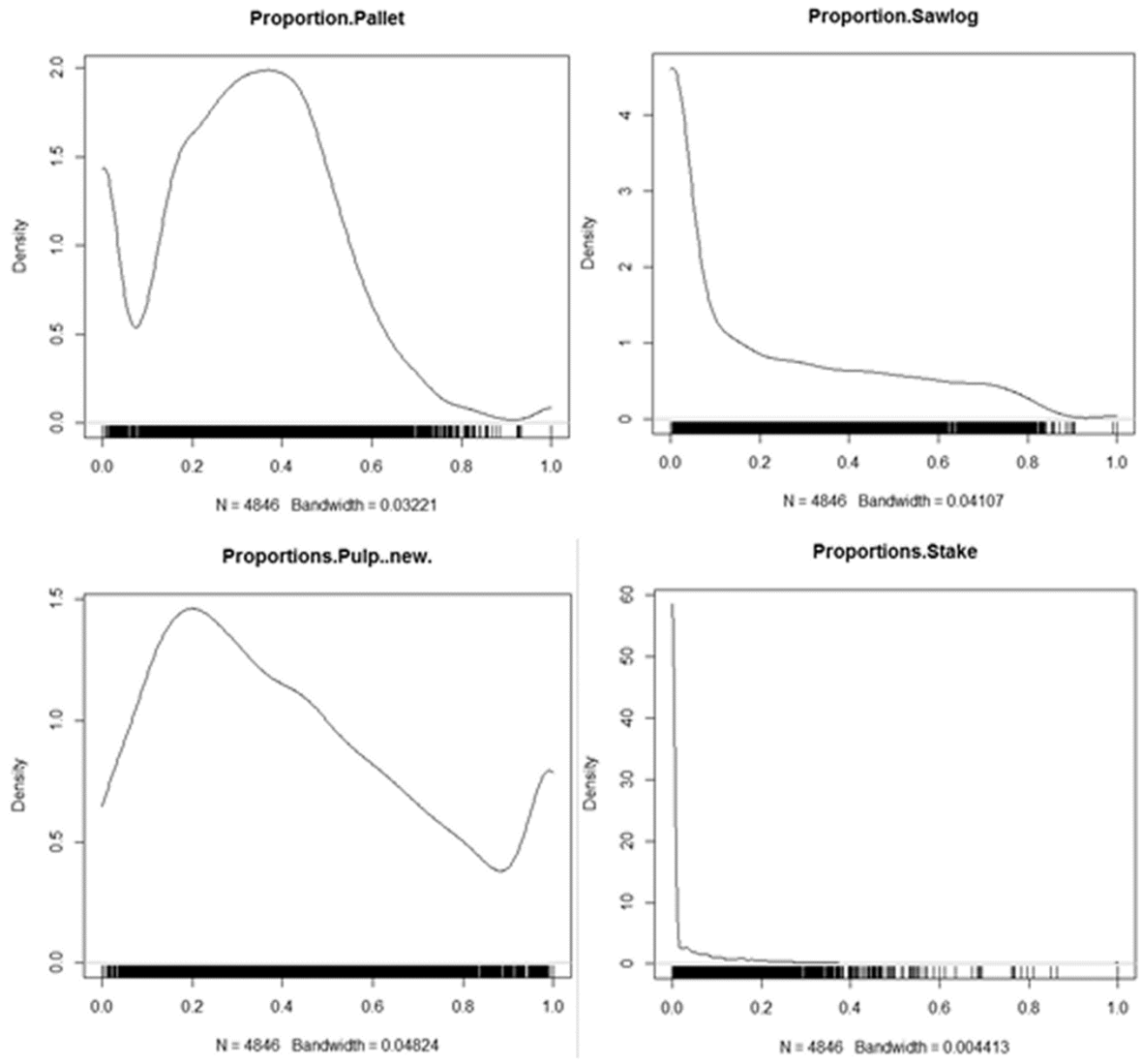
KAUL, M., HILL, R. L. & WALTHALL, C. 2005. Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems,* 85**,** 1-18.
LEK, S. & GUÉGAN, J. F. 1999. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling,* 120**,** 65-73.
MÜLLER, D. & MBURU, J. 2009. Forecasting hotspots of forest clearing in Kakamega Forest, Western Kenya. *Forest Ecology and Management,* 257**,** 968-977.
MURPHY, G., LYONS, J., O'SHEA, M., MULLOOLY, G., KEANE, E. & DEVLIN, G. 2010. Management tools for optimal allocation of wood fibre to conventional log and bio-energy markets in Ireland: a case study. *European Journal of Forest Research,* 129**,** 1057-1067.
PENG, C. 2000. Growth and yield models for uneven-aged stands: past, present and future. *Forest Ecology and Management,* 132**,** 259-279.
R Core Team. 2013. R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing, Vienna, Austria.URL* http://www.R-project.org
RIEDMILLER, M. & BRAUN, H. 1993. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. *Proceedings of the IEEE Internation Conference on Neural Networks (ICNN), 1:586-591, 1993.,* 1**,** 586-591.
SARLE, W. S. Stopped training and other remedies for overfitting. Proceedings of the 27th Symposium on the Interface of Computing Science and Statisfi (.'. _\'. pp. 352-360. Interface Foundation of North America, Fairfax Station. VA, USA, 1995.
SIMAS, A. B., BARRETO-SOUZA, W. & ROCHA, A. V. 2010. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis,* 54**,** 348-366.
SMITHSON, M. & VERKUILEN, J. 2006. A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables. *Psychological Methods,* 11**,** 54-71.
SOARES, P., TOMÉ, M., SKOVSGAARD, J. P. & VANCLAY, J. K. 1995. Evaluating a growth model for forest management using continuous forest inventory data. *Forest Ecology and Management,* 71**,** 251-265.
SWINGLER, K. 1996. *Applying neural networks: a practical guide*, Morgan Kaufmann.
WANG, Y., RAULIER, F. & UNG, C.-H. 2005. Evaluation of spatial predictions of site index obtained by parametric and nonparametric methods—A case study of lodgepole pine productivity. *Forest Ecology and Management,* 214**,** 201-211.

456 Figure 1 Multi-Layer Perceptron example



457

458 Figure 2 Distribution of DV (Sawlog, pallet, stake, pulp)



459
460

Figure 3 Summed DV Proportions (Unit Constraint) MLR

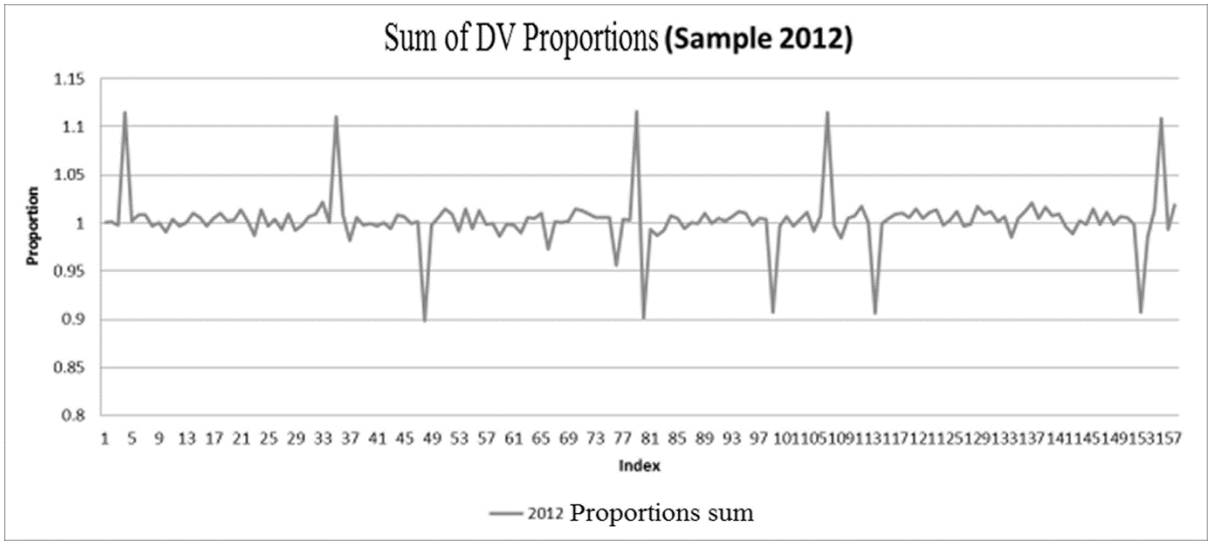


Figure 4 Neural Network Structure

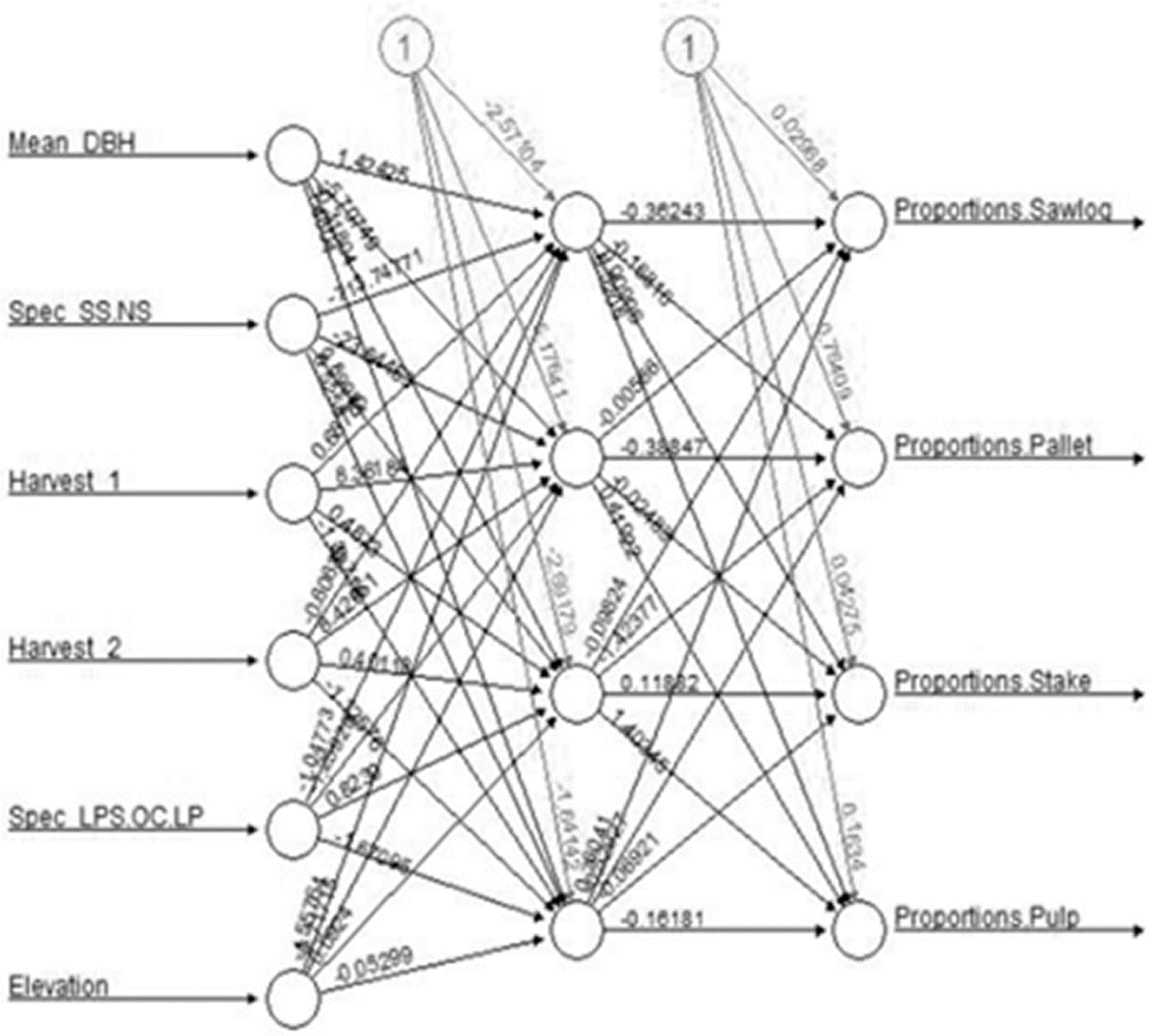
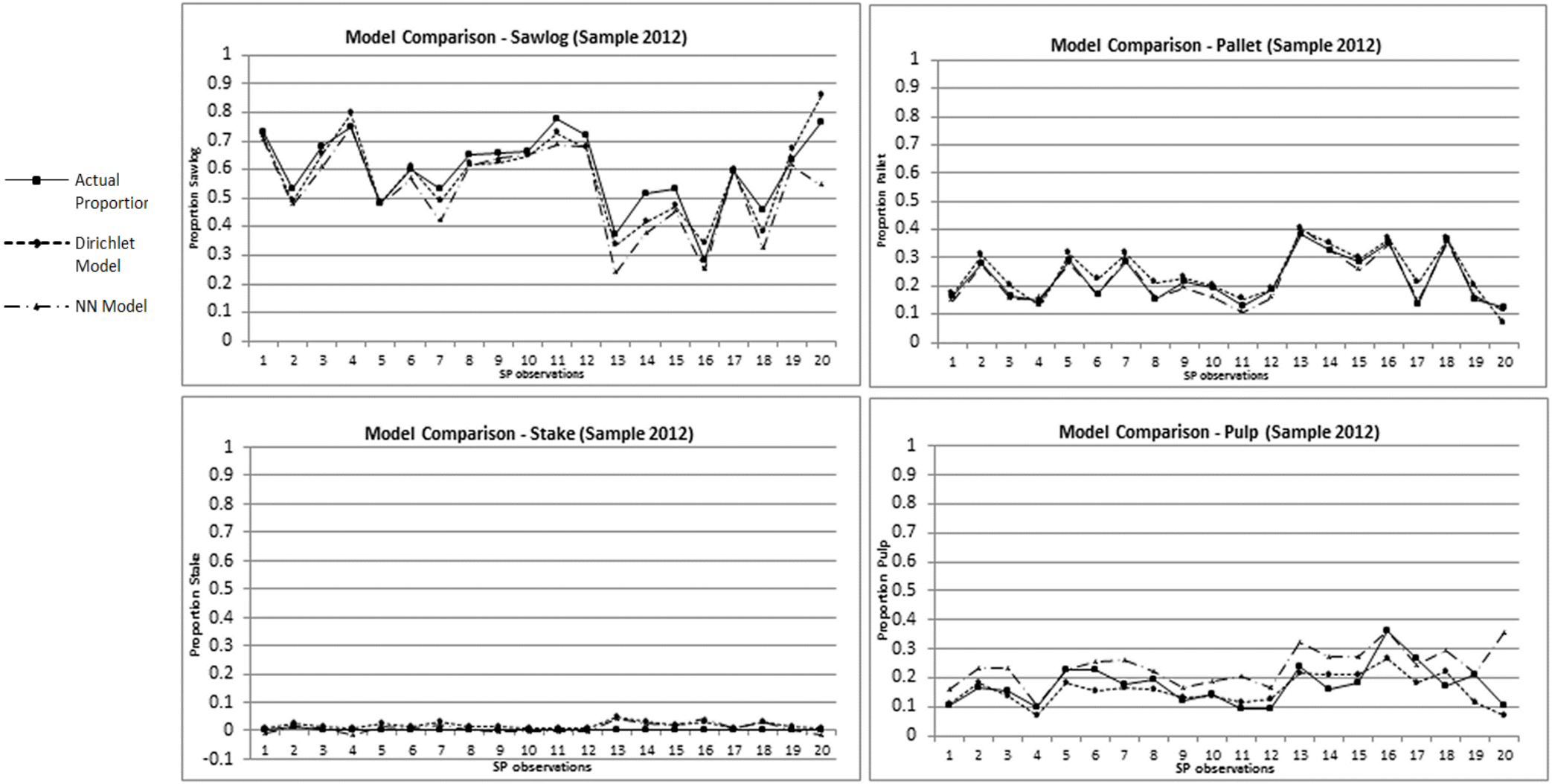


Figure 5 Comparison of Models (2012 test set) - Sawlog/Pallet/Stake/Pulp



469 Table 1 Summary of Dependant Variable summary statistics

Proportions	Valid Range	Prediction	SD
Sawlog	0-1	0.198	0.249
Pallet	0-1	0.325	0.195
Stake	0-1	0.046	0.109
Pulp	0-1	0.43	0.293

470

471 Table 2 Dirichlet alternative model regression coefficients showing influence (in %) on planned end products

472

Attribute \ Product	Sawlog		Pallet		Pulp		Stake	
	$\beta_{iSawlog}$	% change	$\beta_{iPallet}$	% change	β_{iPulp}	% change	β_{iStake}	% change
Intercept	-4.0172	-5,454.54	-0.8249	-128.17	0.3994	49.09	-0.9643	-162.29
Sitka Spruce/Norway Spruce	1.1016	200.90	1.1829	226.38	0.4852	62.45	1.0943	198.71
Douglas Fir/Larch/Scots Pine	0.4375	54.89	0.3493	41.81	0.3800	46.23	0.3906	47.79
Broadleaf	0.0325	3.08	-0.5815	-78.87	0.7604	113.92	-	-
Lodgepole Pine	-0.2117	-24.28	0.4434	55.80	1.4529	327.55	0.6587	93.23
Elevation	0.0000	0.00	0.0000	0.00	0.0000	0.00	0.0000	0.00
Windthrow	0.6439	90.38	0.1269	13.53	-0.1216	-12.93	-0.3160	-37.16
Clearfell	1.0348	181.44	0.1820	19.96	-0.1759	-19.23	-	-
Second Thinnings	-0.0765	-7.95	0.2469	28.01	-0.0489	-5.01	-0.3120	-36.62
Subsequent Thinngs	0.3167	37.26	0.2601	29.71	-0.4347	-54.45	-0.4834	-62.16
Premature Thinnings	0.9691	163.56	0.2529	28.78	-0.1873	-20.61	-	-
Slope	-0.0112	-1.13	0.0230	2.33	-0.0381	-3.88	-0.0333	-3.38
Mean-dbh	0.1553	16.80	0.0543	5.58	0.0299	3.04	0.0459	4.70
First Thinnings							-0.3723	45.11

473
474

475 Table 3 Comparison of Models

Model	PEP Model	Dirichlet Regression	Neural Networks	MVMR
Software*	Excel	R /Excel	R	R/Excel
Running Time	Fast (<10sec)	Fast (<10sec)	Slow (>1hr)	Fast (<10sec)
Interpretation	% downgrades did not show influencing attributes, taken at face value that downgrade is correct.	Influential variables displayed in chart for unit increase of variable. Very easy to interpret results. Creating new model is most difficult part.	Difficult to justify output as interpretation of hidden units/layers can be difficult. Does not display influential variables coherently like Dirichlet.	Interpretation can be little tricky as different transformations per attribute. Does display influential variables.
Accuracy*	**65%/62%/79%/50%**	**69%/66%/96%/56%**	**66%/67%/96%/58%**	**59%/52%/93%/51%**
Complexity (AIC/BIC)	Low	Low	High	Low
Ave MSE*	**.027/.019/.005/.048**	**.018/.016/.004/.035**	**.017/.018/.003/.038**	**N/A**
User Friendliness*	High	High	Low	High
Reliability	Proportions not always sum to 1	High	High	Proportions not always sum to 1

Software*	Many other software packages will be able to model these selected techniques. The applications selected represent the software applications used for this project and also used in Coillte.
Accuracy*	Accuracy of model is based on the test set of 2012 and what percentage of the total predictions across each **DV Sawlog/Pallet/Stake/Pulp** fell within 0.10 raw residual of the actual proportion of that SP
Ave MSE*	**Sawlog/Pallet/Stake/Pulp**
User Friendliness*	Based on non-expert manipulating results for further assessment

476

477