Modelling the Structure of Australian Wool Auction Prices

[1]Caccetta, L., [1]C. Chow, [1]T. Dixon and [2]J. Stanton

[1]**Western Australian Centre of Excellence in Industrial Optimisation, Curtin University of Technology,**
[2]**Department of Agriculture, Western Australia, E-Mail:** chi.chow@postgrad.curtin.edu.au

Keywords*: Wool; Pricemaker; Micron; Hauteur; Tree-based regression; Recursive partitioning.*

EXTENDED ABSTRACT

Australia is the largest exporter of wool in the world. Over 500,000 lots of wool are sold in Australian raw wool auctions each season. Wool is a major Australian export worth about AUD $3.5 – 4 billion dollars per year. It constitutes around 17 per cent of all farm exports in Australia.

The auction centres for wool in Australia are located in three regions, Northern (Sydney, Newcastle, Goulburn), Southern (Melbourne, Geelong, Adelaide, Launceston) and Western Australia (Fremantle). In an auction, raw wool is sold using the greasy price, expressed in cents per kilogram. This can be converted to a clean price estimate by multiplying greasy price by 100 and dividing by the yield. Here yield is the estimate of the clean fibre either after washing/scouring or processing of greasy wool. The response variable is the clean price in c/kg, which is the base price less the total discount. The base price for wool for a given style and fibre diameter (micron), assuming there are no faults for strength, length, vegetable matter or colour is expressed in c/kg clean.

The Australian Wool Exchange (AWEx) prepares market reports which contain price tables. However these tables are often sparse when several quality characteristics are combined. In addition, auction price indicators and some premium and discount tables are released regularly (weekly) by AWEx for the local market participants and for the international sectors that rely on their raw wool supplies from the Australian market.

The multiple linear regression software package called "Pricemaker" has been the most common tool for assessing the wool price and the effects of changes in quality on the wool price. The wool characteristics that are taken into account by the Pricemaker are: fibre diameter, staple strength (measure of the strength of a wool staple), vegetable matter content, staple length (measure of the average length of a wool staple), unscourable colour and style. The relative importance of these characteristics appears to change over time, but there have been few attempts to identify or separate the effects of supply changes from changes in market demand (Stanton 1993; Stanton 1994; Stanton and Coss 1995; Stanton et al. 1997). The only observation is that diameter dominates. A limitation of the Pricemaker system is the mean fibre diameter must be within the range of 18.5 – 24.5 μm. The main deficiency of Pricemaker's multiple linear regression method and other available tools is their complexity and the extensive list of underlying assumptions. The assumptions include: linearity, normality (especially the symmetry on the high price side and in the discount region), independence, etc.

Tree-based methods (Breiman et al. 1984; Cheng et al. 2004) are an alternative means to generalised linear (Watters and Deriso 2000) and additive models for regression problems and to linear logistic and additive logistic models for classification problems. These types of model are fitted by binary recursive partitioning i.e. successively splitting a dataset into increasingly homogeneous subsets until it is infeasible to continue, based on a set of "stopping rules".

An important advantage of tree-based regressions is that they are easier to interpret and discuss in contrast to linear models when analyzing a set of independent variables that contain a mixture of numeric variables and factors. Intuitive and interpretable models are given in the form of "rules". In addition, tree-based regressions do not predict or grow nodes when there is insufficient data. They are robust to sparse missing values. Tree-based regressions are also known to be robust to monotonic behaviour of independent variables, so that the precise form in which these appear in the model is irrelevant.

This paper investigates a tree-based regression technique to model the clean price of wool. A number of results from the Fremantle auction data will be presented and discussed.

1. INTRODUCTION

There have been a number of studies involving the wool market and auctions (Graham-Higgs et al. 1999; Jones et al. 2004; Kemp and Willetts 1996; Simmons and Hansen 1997) and agricultural forecasting (Allen 1994; Bessler 1994; Freebairn 1994; Tomek 1994). In particular, Cheng et al. (2004) have proposed to model the Australian wool auction price by using tree-based regression. They modelled data between July 2000 and December 2000 in Fremantle by using tree-based regression and multiple linear regression. Comparison between two regression methods was made. The standard linear model does not allow interactions between independent variables unless they are in multiplicative form. Tree-based models can detect interaction between parts of levels or parts of the numeric range of independent variables. Based on the results illustrated by tree-based regression, only three predictor variables were used, namely fibre diameter, staple length and staple strength, instead of using six predictor variables as required by Pricemaker system.

This study further develops the idea of Cheng et al. (2004) of modelling the Australia wool auction price with tree-based regression. We examine the prices of clean wool from the Fremantle centre in the Western region for three specified time periods of interest, and investigate the ability of tree-based regression to model wool auction data.

2. TREE-BASED REGRESSION MODEL

Tree-based regression models have been used widely in fields such as social science (Morgan and Messenger 1973; Morgan and Sonquist 1963), statistics (Breiman et al. 1984) and machine learning (Quinlan 1979, 1983, 1986 and 1993). Much research has been done in this area and improvements are continually developed (Scott et al. 2003). Tree-based models belong to statistical classification techniques and are defined by the algorithm used to fit them. The algorithm partitions the space of independent variables (**X**) into homogenous regions such that, within each region, the conditional distribution of y given x, f($y|x$), does not depend on x. Independent variables can be both factors and/or numeric.

The deviance of a regression tree is the usual scaled deviance for a linear model. Similar to the linear regression method, the tree-based regression method also uses residual sum of squares as its measure of fit. Both predict a continuous response and the definition and meaning of residuals and sum of squares is the same in both:

$$D = \sum_{cases:j} (y_j - \mu_j)^2 \qquad (1)$$

where $y_j \sim N(0, \sigma^2)$, $j = 1, \ldots, N$, is the response and μ_j is the mean. The split that gives the largest reduction in deviance will be chosen.

The model in this study is fitted using a binary recursive partitioning that follows Breiman et al. (1984) quite closely. A recursive partitioning package available for the software package R, a statistical programming environment, is utilized whereby the data are successively split along coordinate axes of the predictor variables so that at any node, the split that maximally distinguishes the response variable in the left and right branches is selected. Splitting continues until nodes are pure or data are too sparse; terminal nodes are called leaves, while the initial node is called the root. Since the response variable is numeric, the tree is called a regression tree. The model used for regression assumes that the numeric response variable has a normal (Gaussian) distribution.

In growing a tree, the binary partitioning algorithm recursively splits the data in each node until either the node is homogeneous or the node contains too few observations. The minimum node deviance and the minimum number of observations in fitting a tree-based regression for small datasets varies according to the software and algorithm used. Robust methods to determine the two mentioned criteria are still unknown.

3. DATA

The Department of Agriculture, Western Australia, has kept an extensive data collection of wool that were offered in auctions from the late 1960s to the present. Figure 1 shows the average price per kilogram of clean wool from Fremantle auction data between July 1998 and December 2002. Series 1 represents wool of 19 microns, Series 2 represents 21 microns, while Series 3 represents 23 microns.

It is commonly accepted in the wool industry and community that diameter (micron) is the dominant characteristic that drives the final clean price. Micron is the unit of length equivalent to one millionth of a metre or 0.001 mm. As micron determines the type of product the wool can be turned into, it is very often the first and arguably the most important characteristic an auction buyer would take into account. This is reflected in the majority of Figure 1, where finer wool was sold for a higher average price. However, it can be seen that this price difference began to diminish

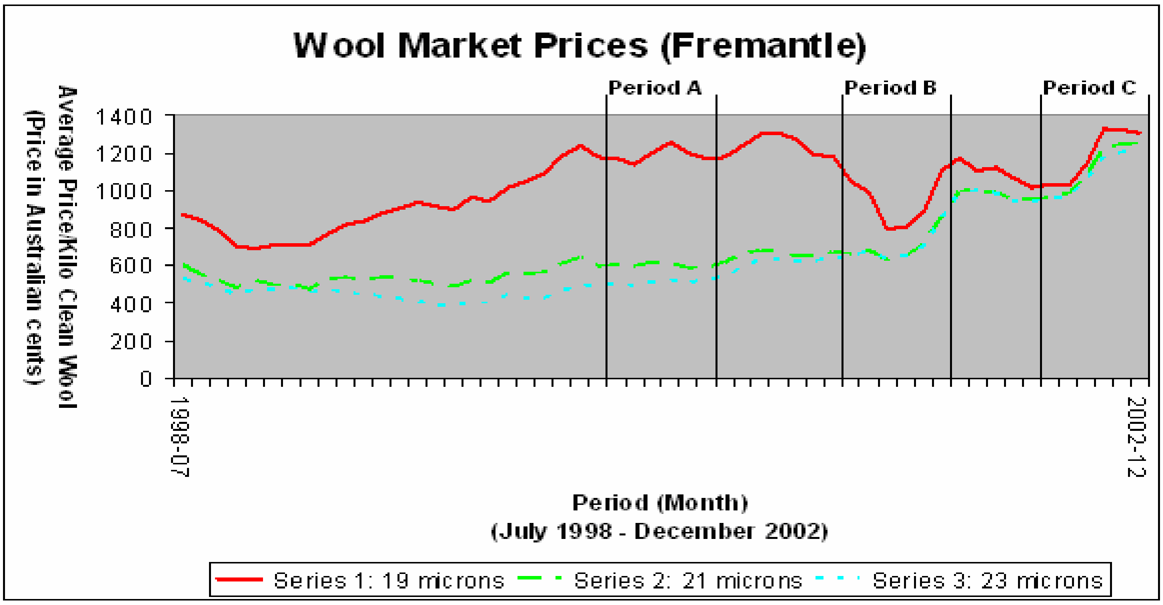

Figure 1. Average prices of wool of three different microns.

sometime around August 2001.

We choose to examine three different 6-month periods: July – December 00 (Period A), August 01 – January 02 (Period B), and July – December 2002 (Period C). We examine Period A because the price difference due to micron was the greatest in this period. We examine Period B because this was when the price difference began to diminish. We examine Period C because the price difference was minimal in this time.

The datasets used combine background information (including sale data, broker, lot sequence, weight, brand and region), appraised quality characteristics (style, unscourable colour) and measured quality characteristics of the raw wool (diameter (µm), coefficient of variation (or cv) of diameter (cvd %), staple length (mm) (SL), cv of staple length (cvsl %), staple strength (N/ktex) (SS), vegetable matter content (%), curvature (°/mm), comfort factor (%), yield (%), point of break (%), and hauteur (mm)). Here, hauteur is an industry measure of the average fibre length.

Tree-based regression is used to fit the wool auction datasets and model the clean price. The most common predictor variables: diameter, yield, staple strength, staple length, point of break, vegetable matter and hauteur are utilised in the tree-based regression.

4. RESULTS

Figures 2 – 4 show the plots of price vs. diameter in the three periods.

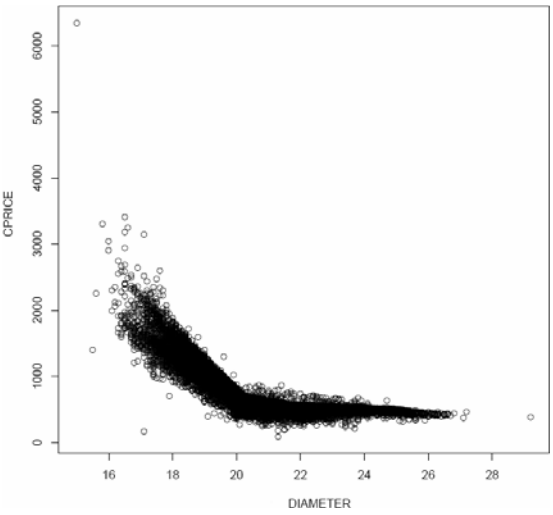

Figure 2. Price vs. Diameter in Period A.

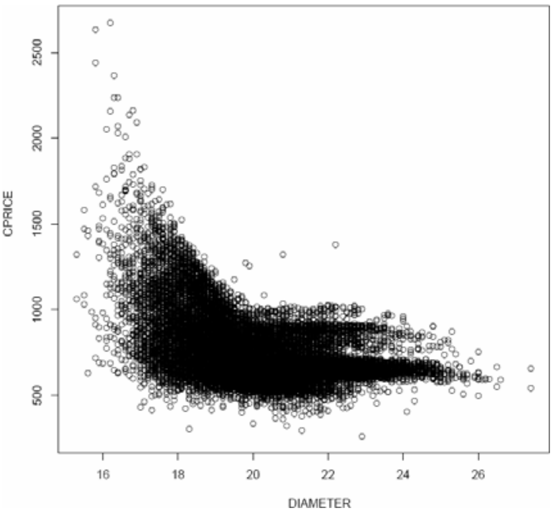

Figure 3. Price vs. Diameter in Period B.

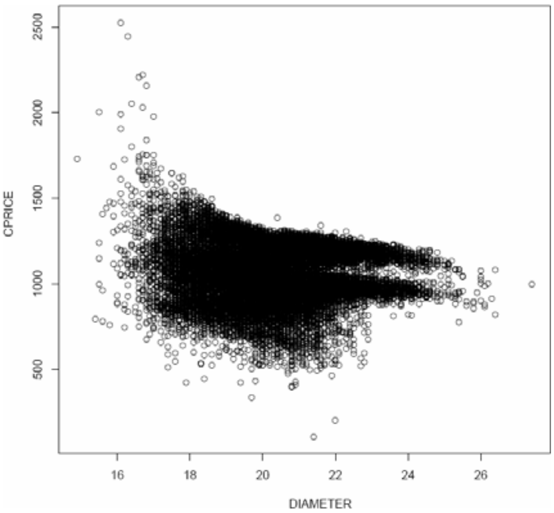

Figure 4. Price vs. Diameter in Period C.

Figure 5 shows the tree generated for Period A, where price difference due to diameter (micron) was great, and diameter is expected to be the dominating variable. Figure 2 also shows a clear pattern, which reinforces our expectation. The decision splits on the actual tree are dominated by diameter, which agrees with the expectation.

Figure 6 shows the tree for Period B, where price difference due to diameter starts to diminish. The pattern in Figure 3 is also less distinct and less visible. On the tree, diameter is still the dominating variable as it determines the top split and plays an important role in subsequent splits. However, other variables (characteristics) have taken up a number of the splits, indicating their growing importance in driving the price as we go from 2001 to 2002. Especially interesting of these variables is the yield, whose splits have displayed their significance in determining the prices of a high percentage of the overall sale.

Figure 7 shows Period C, where price difference due to diameter had become minimal. Here, diameter is expected to lose its dominance, as supported by the further lack of distinct or visible pattern in Figure 4. The tree shows that yield has indeed taken over diameter as the top split. Other variables continue to show their significance on the tree. Diameter appears further down the tree, but it is clear that its significance has taken a back seat.

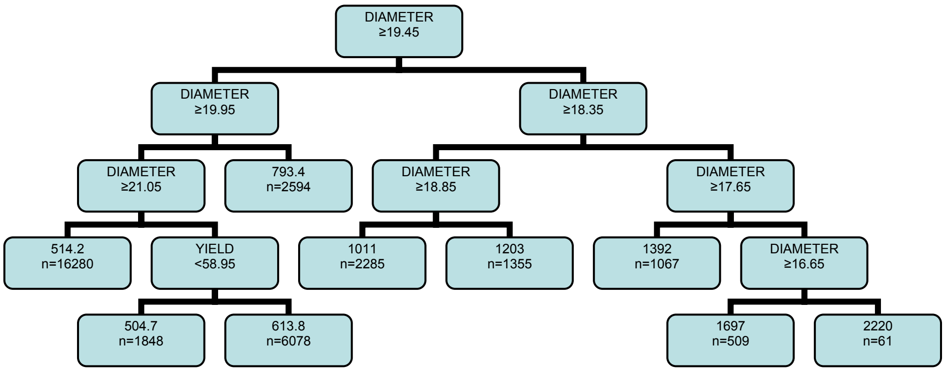

Figure 5. Tree for Period A.

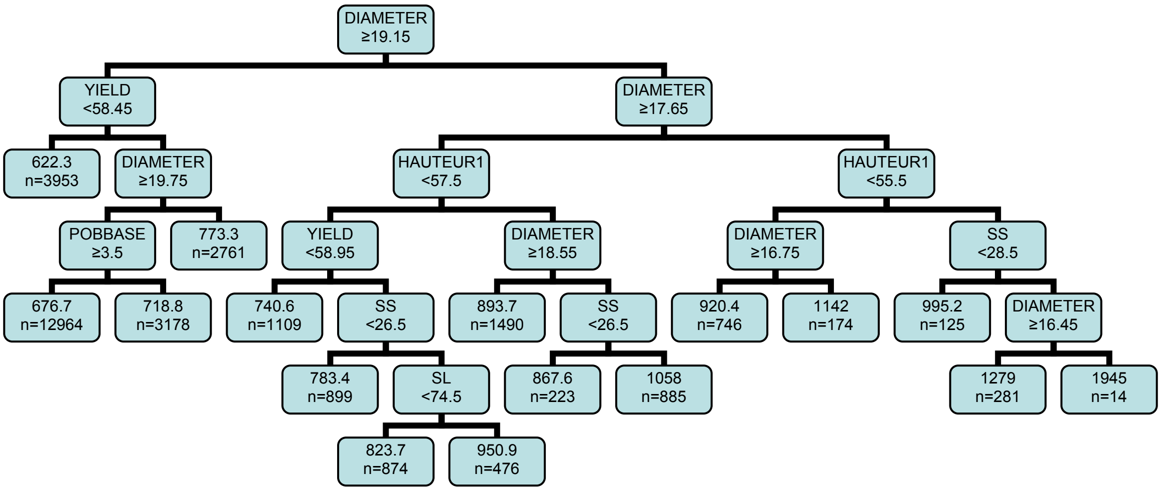

Figure 6. Tree for Period B.

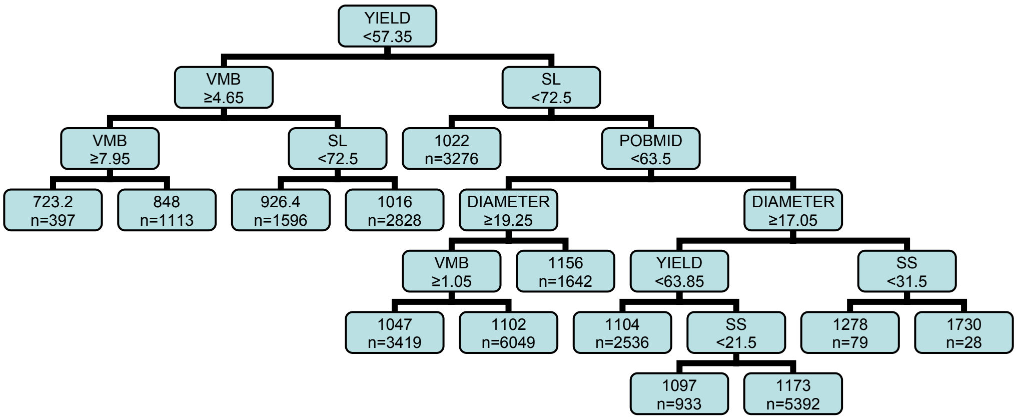

Figure 7. Tree for Period C.

A very important characteristic of wool used in the industry is the hauteur. Hauteur is a measure of the average fibre length. It can be used to determine the quality type and value of wool.

Figures 8 – 10 show the plots of hauteur vs. diameter for each predicted price (an end node on the tree) in the three periods. Each predicted price is represented by a different colour.

The relatively minor or lack of overlapping of the predicted prices in Figure 8 demonstrates that each node on the tree for Period A is made up of a single quality type of wool. Hence we have a reasonable model for this period

However, the overlapping of predicted prices in Figure 9 and 10 demonstrates that each node on the trees in Period B and C is actually made up of several different quality types of wool, which happen to share a similar level of price. This is expected since the response variable of the model is price, not the quality type of the wool.

It is possible that there are fundamental slopes between some of the variables and clean price, so only successive and excessive partitioning will generate a good result. Hence, further care must be taken and an algorithm needs to be developed to improve the model.

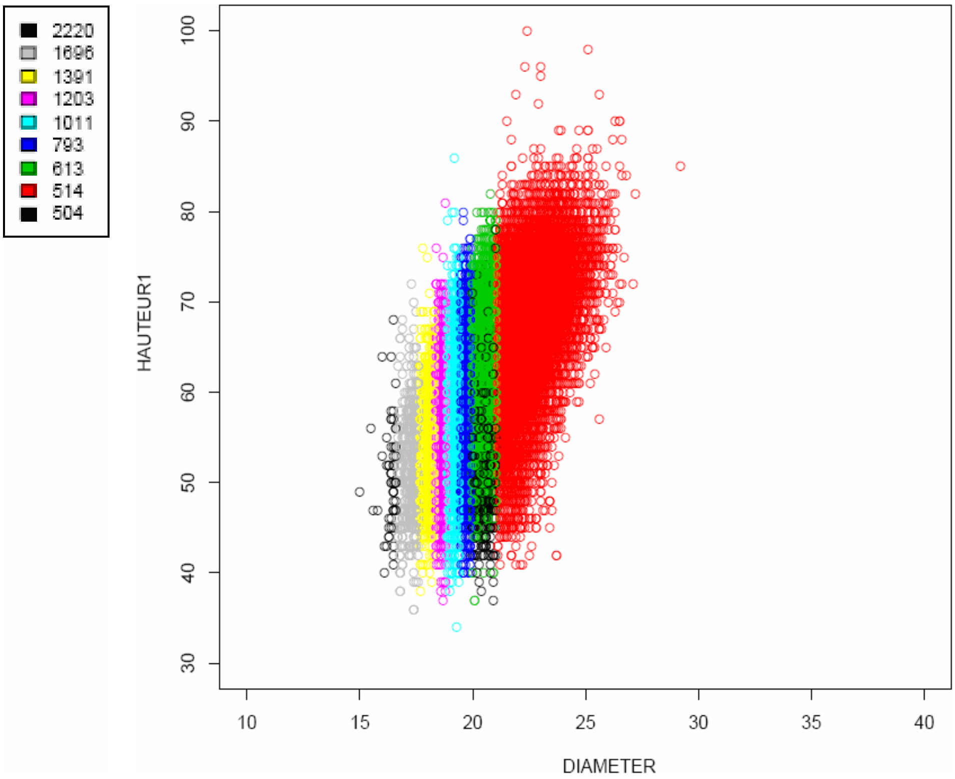

Figure 8. Hauteur vs. Diameter in Period A.

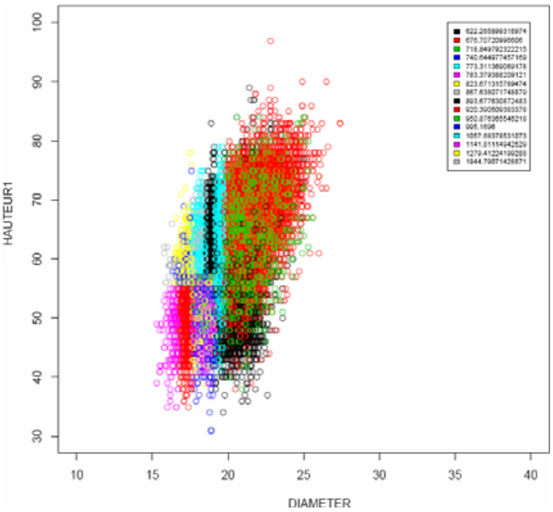

Figure 9. Hauteur vs. Diameter in Period B.

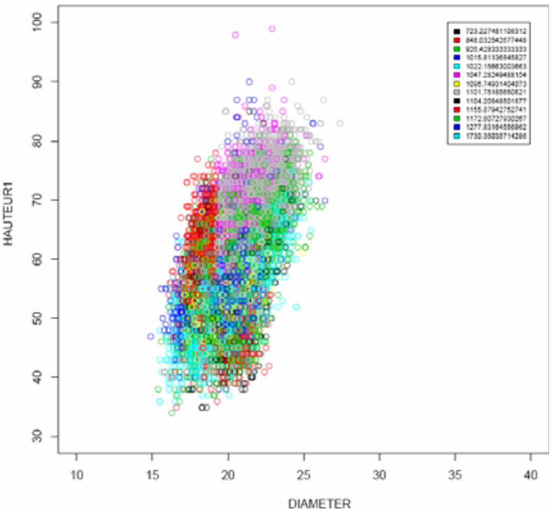

Figure 10. Hauteur vs. Diameter in Period C.

Figures 11 – 13 show the plots of predicted price vs. clean price for each of the three periods. In each period, the variance in price within the end nodes is very high. The individual nodes are evident especially when they span a large actual clean price range. Hence this is another area to be improved and developed.

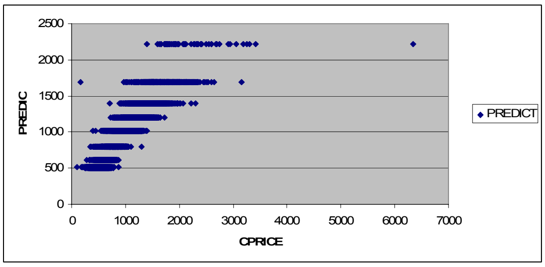

Figure 11. Predicted Price vs. Clean Price in Period A.

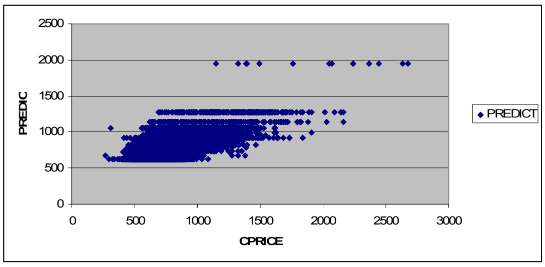

Figure 12. Predicted Price vs. Clean Price in Period B.

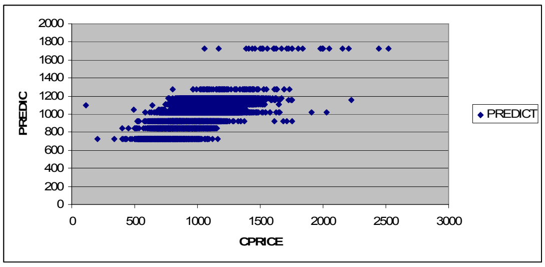

Figure 13. Predicted Price vs. Clean Price in Period C.

5. CONCLUSIONS

Tree-based regression, with its many advantages over classical multiple linear regressions is full of potential in modelling and forecasting wool auction price. We have demonstrated the method's ability to display (as a tree) the degree of importance of each wool characteristic in influencing the final auction price. However, it is possible that there are fundamental slopes between some of the variables and clean price, so only successive and excessive partitioning will generate a good result. Also, the variance in price within the end nodes is very high. The individual nodes are evident especially when they span a large actual clean price range. Further care must be taken and an algorithm needs to be developed to improve the model.

6. REFERENCES

Allen, P.G. (1994), Economic forecasting in agriculture, *International Journal of Forecasting*, 10, 81–135.

Bessler, D.A. (1994), Economic forecasting in agriculture: Discussion, *International Journal of Forecasting*, 10, 137–138.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone (1984), *Classification and Regression Trees*, Wadworth and Brooks/Cole, Monterey.

Cheng, Y.W., J. Stanton, and L. Caccetta, (2004), Predicting the Australian wool auction price by tree-based regression, in *Proceeding of Industrial Optimisation Symposium*, Curtin University of Technology, Western Australia.

Freebairn, J. (1994), The agricultural commodity market forecasting game, *International Journal of Forecasting*, 10, 139–142.

Graham-Higgs, J., A. Rambaldi, and B. Davidson (1999), Is the Australian wool futures market efficient as a predictor of spot prices?, *Journal of Futures Markets*, 19(5), 565–582.

Jones, C., F. Menezes, and F. Vella (2004), Auction price anomalies: evidence from wool auctions in Australia, *The Economic Record*, 80(250), 271–288.

Kemp, S., and K. Willetts (1996), Remembering the price of wool, *Journal of Economic Psychology*, 17, 115–125.

Morgan, J.N., and R.C. Messenger (1973), *THAID: A sequential search program for the analysis of nominal scale dependent variables*, Technical report, Survey Research Center, Institute for Social Research, University of Michigan, Michigan.

Morgan, J.N., and J.A. Sonquist (1963), Problems in the analysis of survey data, and a proposal, *Journal of the American Statistical Association*, 58, 415–434.

Quinlan, J.R. (1979), Discovering rules by induction from collections of examples, in *Expert Systems in the Microelectronic Age*, ed. D. Michie, Edinburgh University Press, Edinburgh.

Quinlan, J.R. (1983), Learning efficient classification procedures and their application to chess end-games, in *Machine Learning*, eds R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, Tioga, 463–482, Palo Alto.

Quinlan, J.R. (1986), Induction of decision trees, *Machine Learning*, 1, 81–106.

Quinlan, J.R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Scott, C.D., R.M. Willett, and R.D. Nowak (2003), CORT: Classification or regression trees, in *Proceeding of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).*

Simmons, P., and P. Hansen (1997), The effect of buyer concentration on prices in the Australian wool market, *Agribusiness*, 13(4), 423–430.

Stanton, J.H. (1993), Analyses of auction prices from Fremantle and their comparison with Eastern State prices, *Wool Processing Research Opportunities*, Department of Agriculture WA.

Stanton, J.H. (1994), *Western Australian wool production. Part 1: Analysis by weight and characteristics*, Department of Agriculture WA.

Stanton, J.H., and L.R. Coss (1995), Characteristics of wool from shires in the Northern Region, *Rural Research for Farm Profit*, Department of Agriculture WA, 157–158.

Stanton, J.H., K. Curtis, and L.R. Coss (1997), Application of auction information to wool processing, *IWTO Conference*, Boston.

Tomek, W.G. (1994), Economic forecasting in agriculture: Comment, *International Journal of Forecasting*, 10, 143–145.

Watters, G., and R. Deriso (2000), Catches per unit of effort of bigeye tuna: a new analysis with regression trees and simulated annealing, *Inter-American Tropical Tuna Commission*, 21(8), 531–571.