Statistical ensemble models to forecast the Australian macadamia crop

D.G. Mayer[a] **and R.A. Stephenson**[a]

[a] *AgriScience Queensland, Department of Agriculture and Fisheries,*
EcoSciences Precinct, Dutton Park 4102 QLD, Australia
Email: david.mayer@qld.gov.au

Abstract: Australian production of macadamia nuts has generally been increasing over time. This underlying trend, however, features considerable year-to-year variability – for example, the 2011 crop was 28,500 tonnes nut-in-shell, *vs.* 44,000 tonnes in 2014. This degree of variability is generally attributed to climatic influences, particularly around the key phenology phases of flowering, pollination, nut-set and nut-drop. Of late, some management effects have also tended to become equally important to climatic variation. Accurate crop forecasts for the Australian macadamia industry are required each year, in order to facilitate planning, handling, processing and marketing.

A range of statistical and other forecasting methods have been used in agricultural systems. These forecasts have shown quite mixed results. Where the independent variables represent the underlying agronomic processes (or are proxies for these), the forecasts should be reasonably accurate. However some projects have produced quite disappointing results, as the forecasting process is well-known to be fraught with problems. One major issue here concerns the 'changing nature' of the macadamia industry as the orchards age, resulting in recent and current yields being lower than those that have been achieved in past years.

In this study, two levels of crop predictions were produced for the Australian macadamia industry for each of the six separate production regions. Firstly, the overall longer-term forecast was based on tree census data from growers in the Australian Macadamia Society (AMS), scaled up to include non-AMS orchards. Expected yields were based on historical data provided by the growers, with a nonlinear regression model incorporating the interacting effects of tree age, variety, year, region and tree spacing. Orchard decline amongst older trees, which has recently become more apparent, was also incorporated into the yield model. Long-term forecasts were made out to about 10 years, after which the effects of (unknown) future plantings, tree removals and rejuvenation of orchards begin to have a major impact.

The second level of crop prediction was an annual climate-based adjustment of these overall long-term estimates, taking into account the expected effects of the previous year's climate on production. The dominant climatic variables were observed temperature, rainfall and solar radiation, and modelled water stress. Based on the proven forecasting success of boosted regression trees and 'random forests' statistical methods, the average forecast from an ensemble of general linear regression models was adopted (rather than using a single best-fit model). Exploratory multivariate analyses and nearest-neighbour methods were also used to investigate the annual patterns in the data. In parallel, AMS each year conducts an annual survey of about 20 key industry growers and consultants. Their replies were integrated into a 'growers forecast' for each year, and this is also taken into account when the AMS releases its annual crop forecast.

Overall, the success rate from this 15-year project has been less than desirable. This is attributed to a number of reasons, including incomplete base-data, macadamia varietal differences and their interactions with climate, and variable management approaches within the industry. Out of the fourteen years of forecasting, the targeted ±10% maximum error rate was only achieved in seven years for the climate forecasts, and six for the growers forecasts. The first seven years of the project generally saw a period of 'good crops', and here the absolute error rates averaged 8.2% for the climate forecasts and 11.6% for the growers forecasts. The next four years had notably poor crops due to low prices which lead to less-intensive management, and all forecasts were too high. The climate-adjusted forecast models had optimistically assumed 'about the same production patterns as before', but these yields were clearly not being achieved. Following a return to more normal prices, the forecasts for the more recent years have shown average absolute error rates of 8.6% for the climate models and 6.8% for the growers forecasts. These are within the targeted ±10%, and compare quite well with other crop forecasting applications.

Keywords: *Regression, yield, tree-crop*

1. INTRODUCTION

As new areas are planted and existing trees age, the production of macadamia nuts in Australia has been generally increasing, from around 30,000 tonnes at the start of this century to over 40,000 tonnes NIS (nut-in-shell, at 10% moisture) in recent years. However, this trend is complicated by a high degree of year-to-year variability, with crops ranging from 28,500 tonnes (in 2011) to almost 44,000 tonnes (in 2004, 2006 and 2014). With most orchards having reasonably good management and pest control, this variability is generally attributed to climatic factors in the year prior to harvest. To facilitate efficient handling and processing demands, and to plan for future marketing and export contracts, the macadamia industry needs to anticipate and manage both future production increases and this inherent annual variability.

Agricultural production systems can be affected by many factors. Statistical model-selection methods are one means to determine the relative importance of these independent influences, and to estimate their effects (Garcia-Paredes et al., 2000; Deng et al., 2005). The fitted coefficients of these statistical models can then also be used for forecasting purposes (Chatfield, 2005).

In this study, macadamia production was forecast in two stages. Firstly, the longer-term ‘expected’ yields were estimated from existing tree numbers, estimated yields and assumed new plantings. Because of the considerable delay in achieving significant levels of production after planting, reasonable predictions out to about six years are possible (Scott, 1992). Beyond this time frame, the effect of (unknown) future plantings starts to impact on the accuracy of forecasts. The second stage was to take these estimates and refine them for each year by considering the effect of the climate during the 13 months prior to harvest. In parallel to this second stage, crop estimates made by surveyed growers and industry consultants were integrated into an alternate ‘expert opinion’ annual forecast.

The objective of this project was to produce forecasts for the industry’s total production each year, with a target of ±10% maximum deviance.

2. METHODS

The Australian Macadamia Society (AMS) has a good working relationship with Australian macadamia processors. Each year the processors provide the AMS with confidential production data by defined regions (Fig. 1). This forms the basis for all forecasts. For each region, these actual production amounts are then standardised to an annual percentage deviance by comparing them with expected production (Mayer and Stephenson, 2000). The expected production for each historical year is derived by formulating and then hind-casting the long-term model.

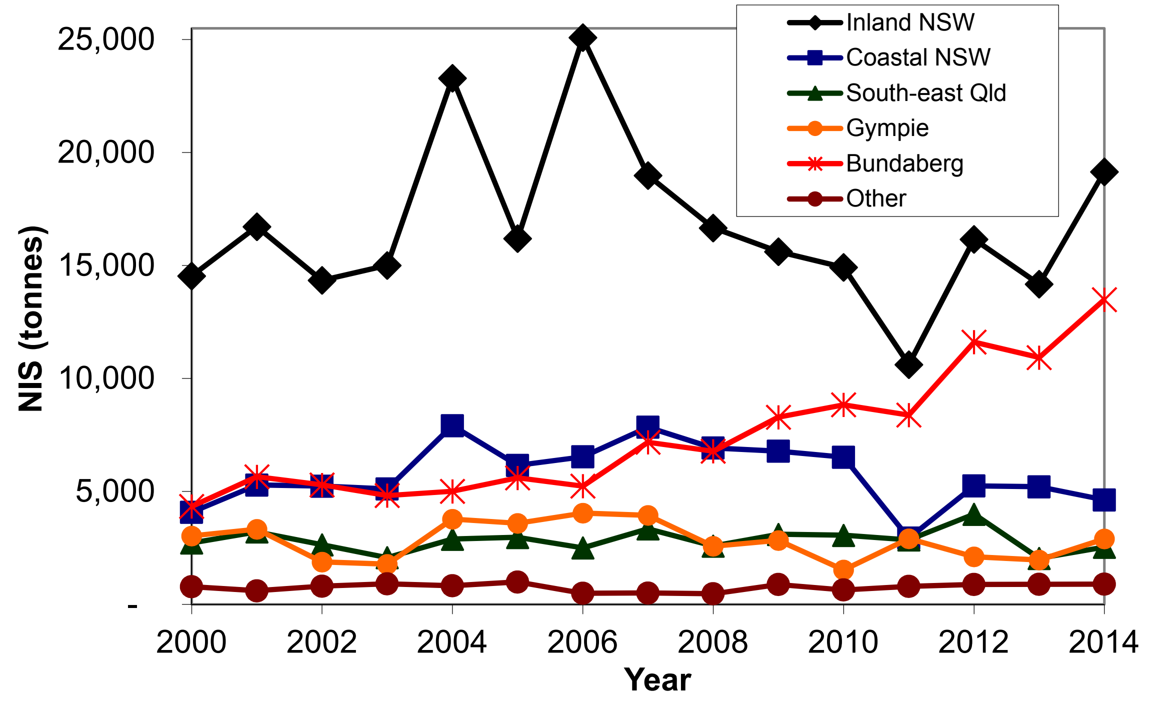

Figure 1. Annual macadamia nut-in-shell (NIS) production by regions.

2.1. Long-term Model

The long-term model for the expected production for each region in each year integrates tree numbers with expected yield-per-tree, with both parts being based on data from the 2010 AMS survey of producers. The yield-per-tree model (as detailed in Mayer et al., 2006) incorporates the effects of region, variety, and the interaction between age and planting density. The actual tree yields utilised to estimate this multiplicative equation are from 2007 to 2010, as these data are felt to be more representative of the current yields in the industry.

Subsequent to the initial development of the yield-per-tree models of Mayer et al. (2006), 'orchard decline' has become evident amongst the older orchards in the industry. Recent data from the industry benchmarking report (AMS, 2015) were used to estimate this effect. This benchmark surveyed 250 farms, covering approximately 55% of both the area planted and production of the industry. Considering the cross-years averages for tree yields by age groups and the individual-years data, we have adopted a 2.5% decline per year from age 20, plateauing at a 20% decline after age 27. Under this assumption, the '25+' year class (nominally taken as 25 to 35 inclusive) averages 13% lower than the 20-24 year old group. This agrees with the benchmark report, where the cross-years average for this decline was 12%.

For the long-term model, assumed new plantings of 40,000 trees per year are added in with the existing tree numbers. The majority of these new plantings are allocated to the Bundaberg and central Queensland ('other') regions. The expected production amounts for each year are calculated for each region, both forecasts and hindcasts. The latter are scaled-up to the actual amounts (see Fig. 2) to account for the farms that were not included in the AMS census. This scale-up factor is similarly applied to the forecasts, to obtain 'whole-of-industry' totals. The revised long-term forecasts are for 46,000 tonnes this year (2015), rising to an expectation of just over 50,000 tonnes in 2020.

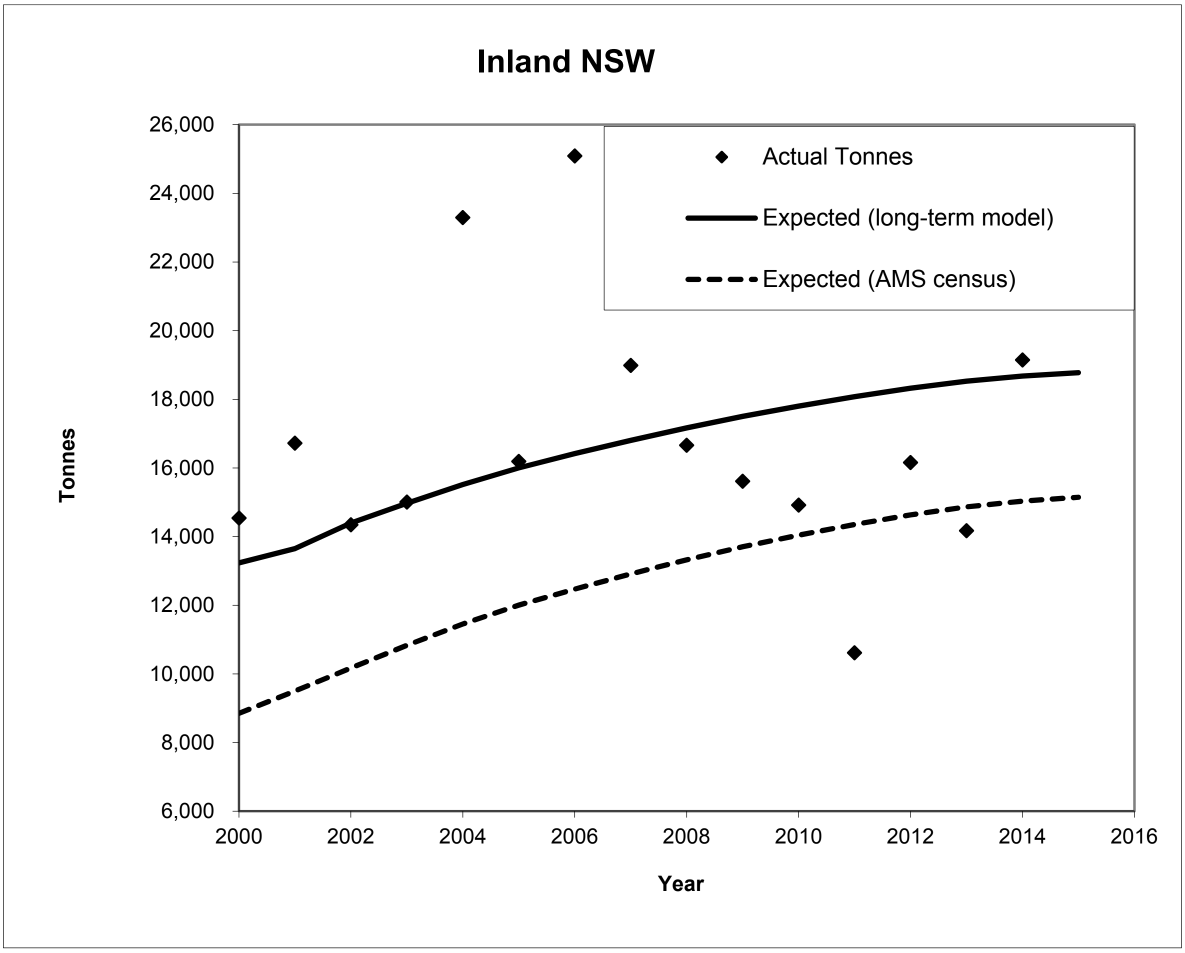

Figure 2. Actual annual production for the 'inland NSW' region, and expected production (from just the AMS-census trees, and scaled-up to the overall long-term model).

2.2. Climate-adjusted Model

This process uses regression ensemble models to adjust the long-term or expected annual production for the effects of the previous year's climate, also including some additional non-climatic influences. The dependent (Y) variable is the percentage deviance (positive or negative) from the expected production for each historical year.

Monthly meteorological data are extracted on a regional basis (Jeffrey et al., 2001), from the approximate centroid for each defined region – Hinkler Park for the Bundaberg region, Woolvi for the Gympie region, Beerwah for south-east Queensland and Newrybar for coastal NSW. For inland NSW, we take the average from Alstonville, Clunes, Dunoon and Lismore. No meteorological data (nor climate adjustment) is used for the minor 'other' region, as it ranges widely from tropical Queensland to Western Australia.

The key variables used for the regression ensembles include maximum and minimum temperatures, monthly rainfall ($\log_{10}$ transformed), adjusted monthly rainfall (capped at monthly evaporation), pan evaporation rate, solar radiation, and cumulative day-degrees either side of 26° C (the optimal temperature for photosynthesis in macadamias; Allan and De Jagar, 1979). In addition, the monthly averages for some modelled climate indices are included. These are calculated from a calibrated soil-water-balance model (McKeon et al. 1990), based on 'an average' macadamia orchard (in terms of soil type and depth, and tree age and planting density). A number of agronomic indices were also considered over the years of this project, and following discussion with industry experts, we adopted the average monthly transpiration-efficiency index, the number of water-stress days per month (days with plant-available-water-capacity < 15%), and the soil-water-index.

In the initial years of the project, monthly climate data were used in model selection. This did cause some problems regarding the number of potential predictors, correlations amongst these, and some selection of adjacent months in different models which were probably accounting for the same climatic effect. To somewhat alleviate these problems, we investigated and then adopted a move to integrating the monthly data into key macadamia physiological periods. These are 'last summer' (the previous January), 'floral initiation' (April and May of the previous year), 'winter' (June to August), 'flowering / nut set' (September and October), 'premature nut fall' (November), 'nut growth' (December), and 'oil accumulation' (January of the current year).

For each region, the important 'non-climatic' effects are also screened, namely the biennial-bearing effect (where a large crop suppresses the crop of the following year, and *vice-versa*), and CPI-adjusted nut prices (direct, plus lagged by one or two years). Of the price variables, the lag of two years consistently had the best degree of fit, being significant for the regions of inland NSW, coastal NSW and Gympie. Notably, no price variable had an effect for the regions of south-east Qld or Bundaberg, with this being a consistent result for the ensemble models over the past few years. Current prices then had no additional effect for any region, possibly because these are correlated with prices two years ago. The biennial-bearing term was generally significant for inland NSW for all the years of this project, and recently has also become significant for Bundaberg.

Forecasting from statistical models 'is fraught with problems and is not for the faint-hearted' (Chatfield, 2005, p. 133). These exercises often produce disappointing results (Chatfield, 2005). Macadamias are recognised as a difficult crop for research – a number of industry workshops and forums held during this project have struggled to define the key influences on production. As exemplified in McFadyen et al. (2004, 2005, 2013), even mature and well-managed orchards display varying yield patterns. Our statistical models provide evidence of the more important influences, but these have varied somewhat over the years and between regions. Linear regression models have previously been used to screen for correlations between yields and meteorological effects, for data from Hawaii (Liang et al., 1983) and Australia (Stephenson et al., 1986). In these studies, temperature, rainfall and stress-days proved important.

Ensemble regression methods are based on the relatively new boosted regression techniques (Elith et al. 2008, Hastie et al. 2009), or 'random forests', which come from the data-mining and machine-learning sciences. These methods develop self-tuning ensembles of regression tree models, and these have recently shown improved predictive behaviour in a number of areas (Hoerl et al. 2014). Song et al. (2013) shows how the successful 'multiple models' concept of these tree-based regressions can effectively be extended into the general linear modelling (GLM) framework. In weather forecasting, the average of multiple models has repeatedly been shown to outperform any of the individual models. These ensemble predictors 'are known to lead to highly accurate predictions' (Song et al. 2013). We are yet to investigate and incorporate the more complex operational parameters of these techniques available in the R language (R Core Team 2013), however our climate-adjustment models have been using a baseline implementation of GLM ensembles in

GenStat (VSN 2014). For each region, around 30 to 50 step-forward multiple regression models are formed, each with different combinations of 'the best' climate terms (and 2nd and 3rd best, at each step). A maximum of four climate terms was imposed for each model, to prevent over-fitting. The models in these ensembles usually agree reasonably well, for example 32 of the 35 forecasts for the 2015 crop in Bundaberg were negative (a lower crop than expected). The overall average forecast from the ensemble of models is adopted for each region.

To assist with the interpretation of the forecasts from the climate adjustment model ensembles, the meteorological data are also subjected to principal components analysis. The dominant two vectors are used to determine which historical years were 'closest' (climatically) to each year to be forecast, effectively identifying analogue years. These annual climate patterns are also used to investigate nearest-neighbour methods, using a two to eight-dimensional representation of the Euclidean distances, and a range of different weighting schemes and numbers of neighbours. However, these exploratory nearest-neighbour forecasts proved to be disappointing (Mayer and Stephenson 2008), and are now used more as confirmation of the climate-adjustment models.

2.3. Growers Forecasts

As a complementary and parallel exercise to the statistical climate-adjustment models, we also instigated an annual survey of key industry growers and consultants. Their estimates of how each year's crop compares with the crop of the previous year are averaged regionally and applied to our estimated regional production breakdown, with these forecasts then being summed to provide an overall industry total. The annual forecast from these sources is taken as at March, which is the same time as the climate-adjustment model forecasts are made. Between 2004 and 2010 the replies from the consultants were tabulated separately. However as these showed no real advantage over the growers, these two groups were subsequently pooled.

3. RESULTS AND DISCUSSION

Table 1 lists the relevant forecasts and error rates for the duration of this project. Out of fourteen years, the forecasts were within the targeted ±10% on only seven years for the climate forecasts, and six for the growers forecasts.

Table 1. Results for the project forecasts.

Year	Actual crop	Climate forecast	% error	Growers forecast	% error	Consultants forecast	% error
2001	34,800	36,000	3.4	33,400	–4.0		
2002	30,200	32,600	7.9	32,300	7.0		
2003	29,700	34,200	15.2	33,800	13.8		
2004	43,700	35,065	–19.8	33,400	–23.6	34,000	–22.2
2005	35,500	35,200	–0.8	38,650	8.9	40,330	13.6
2006	43,900	41,800	–4.8	39,300	–10.5	38,000	–13.4
2007	41,800	39,400	–5.7	36,300	–13.2	37,600	–10.0
2008	36,000	45,600	26.7	40,100	11.4	43,300	20.3
2009	37,500	47,600	26.9	47,500	26.7	45,800	22.1
2010	35,500	41,600	17.2	40,800	14.9	35,600	0.3
2011	28,500	38,900	36.5	33,000	15.8		
2012	40,000	37,070	–7.3	38,280	–4.3		
2013	35,200	39,180	11.3	38,173	8.4		
2014	43,600	40,500	–7.1	40,293	–7.6		

Considering these data by periods, the first seven years saw a period of 'good crops'. Price/kg for macadamias increased steadily from $2.45 in 2001 to $3.60 in 2005, before falling back to $2.60 in 2006. The absolute error rates for 2001 to 2007 averaged 8.2% for the climate forecasts, and 11.6% for the growers forecasts. During this period the regression ensembles worked quite well, despite these methods only being 'in their infancy'. The growers did not perform all that well, but were learning from the feedback, and taking some pride in trying to improve their estimates over time.

The next four years (2008 to 2011) had notably poor crops, and all forecasts were too high. These poor crops were associated with lower macadamia prices – $1.50 in 2007 to $1.90 in 2009; noting here the lag effect as was found in the regression models. During these years the mean error rates were 26.8% (climate) and 17.2% (growers), showing that the growers were more 'attuned' to what was really going on in their orchards and in the overall industry. The climate-adjusted forecast models had optimistically assumed 'about the same production patterns as before', but these yields were clearly not being achieved. Prices started picking up again in 2010, rising from $2.65 in that year to $3.10 in the following year and then increasing steadily to $3.60 in 2014.

For the past three years (2012 to 2014), the absolute error rates have averaged 8.6% (climate models) and 6.8% (growers forecasts). Over the time of this project the growers estimates have shown steady improvement, and overall have been better than the climate forecasts in 7½ of 14 years (2009 was effectively a tie).

Overall, the macadamia industry has proven to be quite difficult to forecast, for a number of reasons, including –

- Incomplete data on tree numbers, ages and densities, requiring the use of scale-up factors for the long-term model.
- Possibly incomplete production totals (mainly concerning exports) which were not included in the processors' data, particularly in the earlier years of the project.
- Varietal differences – anecdotal information suggests short-term climatic influences on pollination success and nut-set which can be quite variable across varieties and regions, and we simply do not have the detailed data to capture these effects.
- Variable management, both across and within regions and years. Our models assumed 'approximately constant management' which would have resulted in steadier yields. However during the period of low prices it was quite obvious that management was less rigorous, as the trees did not deliver anywhere near their potential (as was evident in the preceding and following years).
- The macadamia tree being notoriously 'difficult to quantify', and known to respond to many influences. As a key industry figure initially told us, 'There are many ways to ruin a potentially good crop' (John Wilkie Snr., pers. comm., 2000).

Whilst appearing somewhat disappointing, the relative accuracy of the macadamia forecasts is reasonably similar to other tree-crop forecasting exercises. The USDA annually issue forecasts for their almond crop, based on extensive crop-sampling and with only a two-month lead-time. From 2001 to 2014 these forecasts had an average absolute error rate of 7.8%, with the worst being 13.8% (in both 2002 and 2011). Peiris et al. (2008) predicted coconut production in Sri Lanka using seasonal climate information (primarily rainfall), with an average error rate of 6.8% (for the two years of the study only).

4. CONCLUSIONS

Overall, this project led to a greater understanding of the mechanisms and climatic influences on macadamia production. A range of alternate statistical developments were investigated, including nearest-neighbour methods (not so successful) and regression ensembles (somewhat successful). The 'simpler' and low-cost growers forecasts have shown steady improvement over time, as these personnel become better educated and 'take pride and ownership' of this process.

ACKNOWLEDGMENTS

We thank the staff and members of the Australian Macadamia Society who have assisted us greatly, in particular Kim Jones, Jolyon Burnett, Robbie Commens, Kim Wilson, David Bell, Lyndsay Bryen, Kevin Quinlan, Chris Searle and Paul O'Hare. We are grateful to the macadamia producers who volunteered their intake data. This project has been funded by Horticulture Innovation Australia Limited using the macadamia levy and funds from the Australian Government.

REFERENCES

Allan, P. and J. De Jagar (1979). Net photosynthesis in macadamia and papaw and the possible alleviation of heat stress. *Californian Macadamia Society Yearbook*, 25, 150.

AMS (2015). Benchmarking report – Improving farm productivity and competitiveness in the Australian macadamia industry, 2009 - 2014 seasons. Australian Macadamia Society, Lismore.

Chatfield, C. (2005). Time-series forecasting. *Significance* 2, 131–133.

Deng, X., Y. Luo, S. Dong, and X. Yang (2005). Impact of resources and technology on farm production in northwest China. *Agricultural Systems*, 84, 155–169.

Elith, J., J.R. Leathwick, and T. Hastie (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77, 802-813.

Garcia-Paredes, J.D., K.R. Olsen, and J.M. Lang (2000). Predicting corn and soybean productivity for Illinois soils. *Agricultural Systems*, 64, 151–170.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The elements of statistical learning: data mining, inference and prediction, Second edition, Springer.

Hoerl, R.W., R.D. Snee, and R.D. De Veaux (2014). Applying statistical thinking to 'Big Data' problems. *WIREs Computational Statistics*, 6, 222–232.

Jeffrey, S.J., J.O. Carter, K.M. Moodie, and A.R. Beswick (2001). Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling and Software*, 16, 309–330.

Liang, T., W.P.H. Wong, and G. Uehara (1983). Simulating and mapping agricultural land productivity: an application to macadamia nut. *Agricultural Systems*, 11, 225–253.

McFadyen, L.M., S.G. Morris, C.A. McConchie, and M.A. Oldham (2005). Effect of hedging and tree removal on productivity of crowding macadamia orchards. *Australian Journal of Experimental Agriculture*, 45, 725–730.

McFadyen, L.M., S.G. Morris, M.A. Oldham, D.O. Huett, N.M. Meyers, J. Wood, and C.A. McConchie (2004). The relationship between orchard crowding, light interception, and productivity in macadamia. *Australian Journal of Agricultural Research*, 55, 1029–1038.

McFadyen, L., D. Robertson, M. Sedgley, P. Kristiansen, and T. Olesen (2013). Production trends in mature macadamia orchards and the effects of selective limb removal, side-hedging, and topping on yield, nut characteristics, tree size, and economics. *HortTechnology*, 23, 64-73.

McKeon, G.M., K.A. Day, S.M. Howden, J.J. Mott, D.M. Orr, W.J. Scattini, and E.J. Weston (1990). Northern Australian savannas: management for pastoral production. *Journal of Biogeography*, 17, 355–372.

Mayer, D.G., and R.A. Stephenson (2000). Macadamia crop forecasting. In: Proceedings Annual Conference, Australian Macadamia Society Ltd., 26–28 October 2000, Gold Coast, 27–30.

Mayer, D.G., and R.A. Stephenson (2008). Comparison of regression suite and nearest-neighbour forecasts for the national macadamia crop. Proceedings Australasian GenStat Conference, 2 – 5 December 2008, Marysville, 46.

Mayer, D.G., R.A. Stephenson, K.H. Jones, K.J. Wilson, D.J.D. Bell, J. Wilkie, J.L. Lovatt, and K.E. Delaney (2006). Annual forecasting of the Australian macadamia crop – integrating tree census data with statistical climate-adjustment models. *Agricultural Systems*, 91, 159-170.

Peiris, T. S. G., J.W. Hansen, and L. Zubair (2008). Use of seasonal climate information to predict coconut production in Sri Lanka. *International Journal of Climatology*, 28, 103–110.

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.

Scott., F.S., Jr. (1992). Methodology for projecting orchard crop production: A case study of macadamias. In: Bittenbender, H.C. (Ed.), Proceedings First International Macadamia Research Conference, Kaiua-Kona, Hawaii, 28–30 July 1992, 30–37.

Song, L., P. Langfelder, and S. Horvath (2013). Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC Bioinformatics*, 14, no. 5, http://www.biomedcentral.com/1471-2105/14/5

Stephenson, R.A., B.W. Cull and D.G. Mayer (1986). Effects of site, climate, ciltivar, flushing, and soil and leaf nutrient status on yields of macadamia in south east Quensland. *Scientia Horticulturae*, 30, 227-235.

VSN (2014). GenStat for Windows 16th Edition. VSN International, Hemel Hempstead, UK. Web page: GenStat.co.uk