

# DERIVATION OF A PERENNIAL VEGETATION DENSITY MAP FOR THE AUSTRALIAN CONTINENT

Joanne Chia, Min Zhu, Peter Caccetta, and Jeremy Wallace

Mathematics for Mapping and Monitoring  
CSIRO Mathematical and Information Sciences  
65 Brockway Road, Floreat, W.A 6014

joanne.chia@csiro.au, min.zhu@csiro.au, peter.caccetta@csiro.au, jeremy.wallace@csiro.au  
<http://www.cmis.csiro.au/rsm>

**KEY WORDS:** canopy density, random forests

## ABSTRACT:

Multi-temporal Landsat coverages of the entire Australian continent have been assembled under the Australian Greenhouse Office (AGO) National Carbon Accounting System (NCAS) Land Cover Change Program. The major purpose of this program was the production of spatially detailed classifications of the extent and change in area of forest cover for Australia since 1972 for input to carbon modelling. Changes in vegetation density may also contribute significantly to the carbon budget, and are also of major significance in natural resource management. Landsat has frequently been used for the estimation of biophysical parameters, such as cover or crown density. Regression analyses of image and ground data have commonly been applied. Operational monitoring of vegetation trends from time series of cover/density indices has also been demonstrated in rangeland and forested environments. The AGO imagery provided an opportunity to produce a Landsat-derived estimate of perennial vegetation density for the entire continent using a consistent methodology. Standard methods were defined to provide density 'ground truth' from Ikonos images acquired close to or near 2002. Over 1000 Ikonos images from AGO archives were used. Vegetation density was recorded for multiple sample areas within each image, and nearly 2000 training sites were selected from this process. Ground and image data from multiple 1:1m mapsheets were analysed together. Broad stratification zones and scene boundaries were also used in the modelling process. A variety of analysis techniques were examined, including linear models and 'random forests'. Random forests is a powerful non-parametric tree ensemble technique. Methods and results were compared on the basis of 'goodness of fit' and comparison of output maps. Random forests performed better than linear models in a range of test areas, and this method was applied nationally. The process has produced a map of vegetation density for Australia, limited at present to areas within the AGO forest mask. Details of the random forest procedure, the ground data, and the analysis will be discussed.

## 1 INTRODUCTION

Canopy cover, also known as crown cover, is defined as the percent of area occupied by tree crowns. Estimation of canopy cover is of wide interest in scientific studies for management and policy issues. The Australian Greenhouse Office (AGO) is a leader in developing a credible monitoring of Greenhouse gas production at the national level. In particular, land use change under current Kyoto rules, and under anticipated developments of those rules, requires a credible quantitative methodology and an operational system to deliver land-cover change greenhouse monitoring. The AGO NCAS Land Cover Change Program was specially set up to produce spatially detailed classifications of the extent and change in area of forest cover for Australia since 1972 for input to carbon modelling. As a result of this program, multi-temporal Landsat coverages of the entire Australian continent have become available and this gives an opportunity to produce a Landsat-derived estimate of perennial vegetation density for the entire continent using a consistent methodology.

Spectral mixture analysis (SMA) or linear regression techniques have often been used in the past to estimate tree canopy density. Some of these studies include (Iverson, 1989, Zhu and Evans, 1994, and De Fries et al., 2000). However, since tree canopy is not an end member, it cannot be estimated directly using the SMA method (Roberts et al., 1993). In addition, the model used in the linear regression and SMA is not adequate to model the relationship between spectral signals and canopy density which can get very complex and highly variable especially over large areas (Ray and Murray, 1996). A study conducted by (Huang

and Townshend, 2001) has found that using regression tree technique that includes a set of linear models produced more accurate estimates as compared to using just a single linear model. Some recent studies such as (Huang et al., 2001 and Xian et al., 2002) have successfully used regression tree techniques to model the complex relationships and handle large data sets.

Here, we describe a study carried out to estimate canopy density over the Australian continent for year 2002 using a recently developed regression tree technique known as *random forests*. Random forests is a tree ensemble technique. It is a powerful non-parametric technique that has numerous advantages such as the ability to handle very large data sets and large numbers of variables, it does not over fit and it is computationally efficient (Breiman, 2001).

This paper is organized as follows. Section 2 describes the model inputs such as image data, stratification zones and image date boundaries. It also summarises the study conducted to compare the performance of random forests with other models such as linear regression model, logistic model and the regression tree model as implemented in the statistical software *R* (R Development Core Team, 2006) and describes the application of random forests to Landsat data. Section 3 presents and discusses the results achieved and section 4 ends with conclusion and future work.

## 2 MATERIALS AND METHOD

The overall strategy for data processing in this study consists of three basic steps namely

- The study area is stratified into broad zones reflecting variations in vegetation types due to rainfall, soil and geology;
- Ground data are selected using Ikonos images and with stratification zones as a guide so that they represent a range of densities within each zone;
- Random forests is applied to Landsat data and training sites to obtain estimates of canopy densities.

## 2.1 Landsat Data

The study area covers the forested area of the entire Australian continent. A total of 37 1:1m map sheets covering the continent of Australia namely, SD54, SE54, SE55, SF54, SF55, SG54, SG55, SH54, SH55, SH56, SI54, SI55, SI56, SJ54, SJ55, SH53, SI53, SK55, SI50, SI51, SH50, SH51, SH52, SG52, SG51, SG50, SE51, SF51, SE52, SF52, SF50, SD52, SD53, SE53, SF53, and SG53 were used in the process to produce the final density map. The final map of vegetation density for Australia is limited at present to areas within the AGO forest mask. The AGO forest mask is built from multi-temporal Landsat images for the NCAS Land Cover Change Program. It includes woody areas that are have canopy cover above a nominal 20% (Furby, 2002).

In practice, data from several mapsheets were combined and analysed together to obtain sufficient ground data. The process was iterated using different combinations of geographically adjacent mapsheets.

## 2.2 Ground Data

Standard methods were defined to provide density 'ground truth' from the available Ikonos image archives. Nearly 1000 high spatial resolution Ikonos images acquired in 2002/03 across the country were used to select approximately 2000 training sites. Canopy density for each site was estimated by independent experts using a documented protocol. Stratification zones are used as a guide in selecting training sites so that they represent a range of densities in each zone. For each site, a grid is superimposed over each sample areas in the Ikonos images, and absence and presence of vegetation is counted. The proportion of trees in each grid is then recorded as the density for that site.

It should be noted that the current set of training sites are being revised. The processing described here identified particular areas where the present ground data coverage is inadequate. To improve the coverage and representativeness of the ground data, new Ikonos images are being purchased at the time of writing this paper.

## 2.3 Stratification

In this study, we used the stratification zones developed for the Australian Greenhouse Office Project (Furby, 2002). Each map sheet was divided into different zones reflecting variations in vegetation, geology, soil, rainfall and land use associations for the AGO LCCP. These zones were used in the density mapping, with like zones often being merged. In addition, the zone information is included in the modelling process as a factor.

## 2.4 Image Date Boundary

Seasonal differences between image date boundaries are often visible in the image mosaics. This is caused by different dates of acquisition of images. The image date boundaries are used as factors in the modeling process.

## 2.5 Model Comparison

A range of methods were considered and applied to data from test areas and compared. Two parametric methods (linear regression and the logistic regression), and a non-parametric method (decision trees) were applied. The results obtained were compared with those achieved by random forests. Model performance is measured by the mean absolute prediction errors (MAPE), square root of the value of cross-validation mean square errors (CVME), R-squared value which is calculated by  $(1 - \frac{\sum(resid^2)}{\sum(obs - mean)^2})$ , where "resid" is the residuals and "obs" is the observations, and by visual inspection of output maps.

For model comparison, data from map sheets SD54, SE54 and SE55 were pooled. There were 144 training sites in total. For each model considered, the training sites data were fitted using the spectral signals, stratification zone numbers and image date boundaries where the latter two independent variables were treated as factors. The results achieved by each model are summarized in Table 1 below.

Model	CVME	MAPE	R-square
Linear	22.0	14.876	0.610
Logistic	22.7	13.243	0.618
Tree	25.5	12.873	0.678
Random forest	23.4	8.069	0.878

Table 1: Values of CVME, MAPE and R-square achieved by the linear model, logistic model, tree model and random forest model using data taken from map sheets SD54, SE54 and SE55.

From Table 1, it can be seen that the random forests model has overall performed better than the other models in terms of the mean absolute prediction error and the R-squared value. These two measurements directly compares the predicted densities with those of the training sites. In addition, a scatter plot of estimated densities against the spectral signals revealed that these two variables have a far more complex relationship than that of a simple linear one. Thus, the linear model is not suitable for the data in this case. Visual inspection of output maps was also done and it was found that the map produced using random forests was most consistent with the original satellite image. As an example, Figure 1 below shows a sample area taken from the resulting density maps produced by the linear model and the random forest model. In the example, it can be seen from the left picture that there are variations in the woody area at the top left hand side which is consistent with the Landsat image. The map produced by linear model as shown in the middle picture in Figure 1 does not show such variation. Also, the estimated densities produced by random forests model are more consistent with the supplied ground data as compared to those yielded by the linear model.

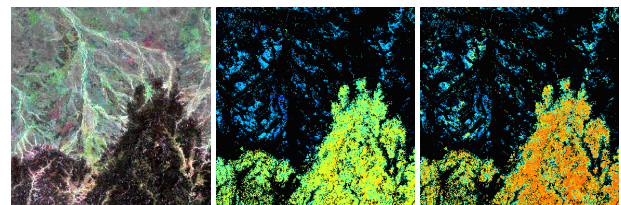


Figure 1: Left: Sample area taken from Landsat image (mapsheet SD54) in the enhancement of bands 5, 4 and 2 in red, green and blue respectively. Centre: Corresponding area taken from density map achieved by linear model. Right: Corresponding area taken from density map achieved by random forest model

## 2.6 Applying Random Forests to Data

Random forests is a classification and regression approach developed by (Breiman, 2001). It is an ensemble of unpruned decision

trees constructed by using bootstrap samples of training data and random feature selection. The basic algorithm of random forest works as follows, assuming training data is of size  $N$ .

- Draw  $N$  observations from the training data with replacement;
- Using data from step 1, construct a decision tree to its maximum depth without pruning it. At each node, a random subset of  $m$  predictor variables from the total set of  $M$  available predictors ( $m < N$ ) are chosen as candidate splitters, and the best split of these  $m$  variables is used to split the node. Perform new random selection for each split;
- Repeat steps 1 and 2 a large number of times to generate a forest of trees;
- Predictions are calculated by averaging for regression or by majority vote for classification.

Some advantages of random forest model include

- in regression-type problems it does not over fit;
- in classification applications it produces highly accurate results;
- it gives estimates of variables that are important in the classification;
- it can handle large number input variables;
- it runs efficiently on high dimensional data sets.

We applied the random forests regression model as implemented in the statistical software *R* (R Development Core Team, 2006) to the data. Operationally, repeated runs of the model were conducted using combined data from different spatial groupings within a number of geographic regions. Typically, data from several map sheets in the same region were pooled and analysed as a single data set. For each run of the model, the estimated densities of the training sites were fitted using on the corresponding Landsat digital counts from six bands, stratification zones and scene boundaries. The regional results obtained were then combined to produce a map of vegetation density for Australia. Areas outside the Australian Greenhouse Office forest extents were masked.

In applying the procedure described above to obtain a density map of the country, the 37 map sheets were merged into 9 groups, resulting in 9 random forests models being used. All 9 models yielded R-square values of between 0.8 – 0.9. Two examples of the groupings, together with the results achieved are presented in the next section.

### 3 RESULTS AND DISCUSSION

The results for two of the random forests models fitted to data from different areas are described in this section. In the first example, data from map sheets SF55, SG55 and SG56 were combined and analyzed. Training data from 194 sites was available. The results are shown in Table 2 where an R-squared value of 0.878 has been achieved, indicating that random forests has fitted the data well. In addition, the diagnostics plots in Figure 2 show that there is good agreement between the estimated and predicted densities (left plot), and that there is no systematic pattern or trend between residuals and predicted densities, indicating that

the model is appropriate for the data. Visual inspection was also performed on the resulting density maps. Figure 3 shows a sample area taken from the density map of map sheet SF55 together with its corresponding Ikonos image. The training site in this area has a recorded density of 71% while the predicted density is 72.5%, which means there is good agreement between the estimated and predicted densities.

In the second example, data from map sheets SH53,SH54,SI53,SI54, and SJ54 were combined and analyzed. The total number of training observations was 295. The results and diagnostic plots are shown in Table 3 and Figure 4. The model has achieved R-squared value of 0.910. The diagnostics plots show good agreement between the estimated and predicted densities, and that there is no systematic pattern or trend between residuals and predicted densities, indicating that the model is appropriate for the data. Figure 4 shows a sample area taken from the density map of map sheet SI54 together with its corresponding Ikonos image. The training site in this area has a recorded density of 44% while the predicted density is 45%, which means there is good agreement between the estimated and predicted densities.

Model	CVME	MAPE	R-square
Random forest	10.97	8.04	0.898

Table 2: Values of CVME, MAPE and R-squared achieved by random forest model using data taken from map sheets SF55, SG55, and SG56.

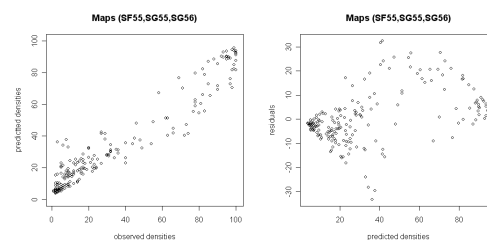


Figure 2: Left:Plot of predicted densities against observed densities for map sheets SF55, SG55 and SG56. Right: Corresponding plot of residuals against predicted densities.

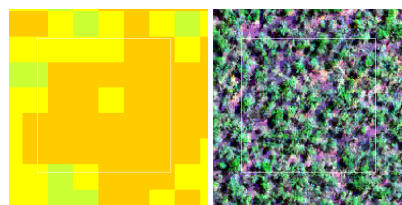


Figure 3: Left: Density map achieved by random forest model for map sheet SI54. The colours red - blue represents high - low densities. Box outlined in white represent the training site. The estimated density of this site is 71%, and the predicted density yields a density of 72.5%. Right: Corresponding Ikonos image displayed in the enhancement of red, green and blue representing the red, near infra red and blue layers respectively.

Model	CVME	MAPE	R-square
Random forest	10.63	7.80	0.910

Table 3: Values of CVME, MAPE and R-squared achieved by random forest model using data combined from map sheets SH53, SH54, SI53, SI54 and SJ54.

### 4 CONCLUSIONS AND FUTURE WORK

An approach has been developed to predict the vegetation cover density of the Australian continent. This approach uses a tree en-

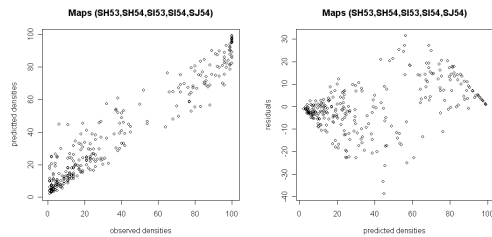


Figure 4: Left:Plot of predicted densities against observed densities for map sheets SH53,SH54,SI53, SI54, and SJ54. Right: Corresponding plot of residuals against predicted densities.

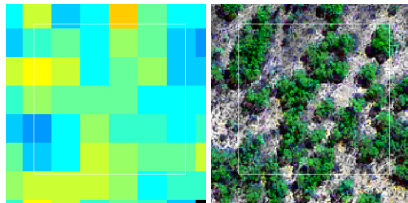


Figure 5: Left: Density map achieved by random forest model for map sheet SI54. The colours red - blue represents high - low densities. Box outlined in white represent the training site. The estimated density of this site is 44%, and the predicted density yields a density of 45%. Right: Corresponding Ikonos image displayed in the enhancement of red, green and blue representing the red, near infra red and blue layers respectively.

semble technique known as random forests, applied to purpose-collected ground data and Landsat images. A total of nine random forests models were used to produce the final first pass density map for the country. All models yielded a R-squared value between 0.8 – 0.9 which indicates that the models are adequate for the data in each case. In addition, diagnostic plots and visual inspections were performed to further confirm the appropriateness of the models. However, it should be noted that the random forests has no spatial knowledge of the area under study. It relies heavily on the input training data to calculate the predictions. Thus, good ground truth and stratification of the region are paramount to produce accurate estimates.

It should be noted that the current set of training sites are being revised as a result of the work described in this paper. New Ikonos images are being purchased at the time of writing, and new training sites will be selected for resolving uncertainty where present data are sparse and to make the distribution of the training sites across the country more comprehensive.

Future work includes processing the revised data to obtain a new density map for the Australian continent for year 2002. Extension of the current methodology to produce time-series density maps is planned.

## REFERENCES

- Breiman, L., 2001. Random forests. *Machine Learning* 45(1), pp. 5–32.
- DeFries, R., Hansen, M. and Townshend, J., 2000. Global continuous fields of vegetation characteristics: a linear mixture model applied to multi-year 8km AVHRR data. *International Journal of Remote Sensing* 21(6/7), pp. 1389–1414.
- Furby, S., 2002. Land cover change: Specification for remote sensing analysis. Technical Report 9, Australian Greenhouse Office.

Huang, C. and Townshend, J., 2003. A stepwise regression tree for nonlinear approximation: applications to estimating sub-pixel land cover. *International Journal of Remote Sensing* 23(1), pp. 75–90.

Huang, C., Yang, L., Wylie, B. and Homer, C. A., 2001. A strategy for estimating tree canopy density using landsat 7ETM+ and high resolution images over large areas. *Proceedings of the Third International Conference on Geospatial Information in Agriculture and Forestry*, Denver, Colorado.

Iverson, L., Cook, E. and Graham, R., 1989. A technique for extrapolating and validating forest cover across large regions: calibrating AVHRR data with TM data. *International Journal of Remote Sensing* 10(11), pp. 1805–1812.

R Development Core Team, 2006. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ray, T. and Murray, B., 1996. Nonlinear spectral mixing in desert vegetation. *Remote Sensing of Environment* 55(1), pp. 59–64.

Roberts, D., Smith, M. and Adams, J., 1993. Green vegetation, nonphotosynthetic vegetation, and soils in AVIRIS data. *Remote Sensing of Environment* 44(2), pp. 255–269.

Xian, G., Zhu, Z., Hoppus, M. and Fleming, M., 2002. Application of decision-tree techniques to forest group and basal area mapping using satellite imagery and forest inventory data. *Proceedings of I/FIEOS Conference, Percora 15/Land Satellite Information IV/ISPRS Commission*.

Zhu, Z. and Evans, D., 1994. US forest types and predicted percent forest cover from AVHRR data. *Photogrammetric Engineering and Remote Sensing* 60(5), pp. 525–531.