Estimating Forest Area using Remote Sensing and Regression Estimator

MOHD HASMADI ISMAIL & KAMARUZAMAN JUSOFF
Department of Forest Production
Faculty of Forestry
Universiti Putra Malaysia
43400 UPM, Serdang, Selangor
MALAYSIA
http://www.upm.edu.my

Abstract: - Area estimates using remotely sensed data is an important subject that has been investigated around the world during the last decade. It plays an important role in the production of vegetation statistic when area frame sample design is used using regression estimator. This technique is used widely in estimation of crop area and yield. This work is carried out utilizing the same method but tested for the tropical forest in Malaysia. The estimates have been conducted using direct expansion from sample survey and regression estimator approaches. The latter result using regression of ground data and satellite data seem more reliable when training pixels are chosen at random subset of the area sampling frame. The regression analyses showed all the land cover class had a very high correlation (r^2 = 0.86 to 0.89). This method is not only practical with accurate estimation for this task but also does not have any additional time and cost implications.

Key-Words: - forest, estimation, remote sensing, regression

1 Introduction

Area estimation through remote sensing is often used for classification and production of crops statistics. This effort was demonstrated in the past by Gonzales-Alonso *et al*.(1991), Gallego and Delince (1993), Gonzales-Alonso and Cuevas (1993), and Ferencz *et al*.(2004). Integration of ground data and classification of remote sensing data is shows a greatest operational feasibility and economical interest that contribute for the benefits of the global society. However the emphasis of their research is on the agricultural crop or vegetation. Thus, this work is carried out for different type of land cover–forest in Malaysian context. In order to develop forest management strategies, surveying the forest resources and monitoring the forest area for harvesting or affected by logging operation is essential.

Forest resource maps must be completed and kept up to date to be useful and effective in forest development and management. Several methods have been used to establish data and information about forest area for this purpose. Examples of studies investigating the use of remote sensing data for forestry applications in Malaysia were described by Kamaruzaman and Souza, 1997; Mohd Hasmadi and Kamaruzaman, 1999; Khali, 2001. Most methods involve use of aerial photographs, Landsat and SPOT satellite data. Although optical data such as Landsat and SPOT have a great limitation with respect to spatial resolution, spectral characteristics, and cloud cover, it is still viable due to cost effectiveness and ease of understanding. On the other hand, Hyppa et al. (2000) claimed that optical remote sensing images still include more information for forest survey.

The collection of accurate and timely statistic about the area of forest in Malaysia is great importance to the Forestry Department, logging company and other interested parties such as Department of Environment and Natural Resources. However, according to Deppe (1994), estimations of forest areas by classical approaches using digital classification alone usually suffer from mis-classified pixels of satellite images, although it has no sampling errors. Meanwhile, estimation by ground observation has probably suffered from high sampling errors (Taylor *et al*., 1997). Thus, the relationship between the area estimate from ground survey and image classification results can be combined in order to improve area estimates (Cochran, 1977). Example of the study by European Commission on crop inventories in 1997 revealed that the error of sampling survey could be minimized by assistance of remote sensing data. The main objective of this study was to estimate the areal extent of forest resources in Sungai Tekai Forest Reserve, Peninsular Malaysia and investigate methodology used to carry out forest area estimation with the aid of remotely sensed data.

2 Methodology

2.1 Study area

The study area is a forest reserve situated in the north east of Pahang state in Peninsular Malaysia. The surface area is mainly covered by virgin forest with some of them is logged over forest, bare land, mix agricultural crop and water bodies. The geographical limit coordinates are latitude 04°10´N - 04°30´N and longitudes 103°03´E - 103°30´E, covering an area of approximately 10,000 hectares (Figure 1). The forest area is composed of mixed virgin hill forest, high in species diversity with predominance of *Shorea* species such as Meranti Seraya (*S.curtisii*) and Meranti Rambai Daun (*S.acuminata*). The elevation is mostly over 600 m above sea level. The slope gradient of the study area is undulating with steep rugged slopes ranging from 10^0 to 80^0. The annual precipitation is about 210 cm with a fairly high tropical climate with a mean temperatures ranging from 20^0C -31 ^{0}C. The precipitation occurs mainly in two seasons: April to May and November to December. The relative humidity is high, ranging from 62.3 to 97.0% with a daily mean of 85.7%.

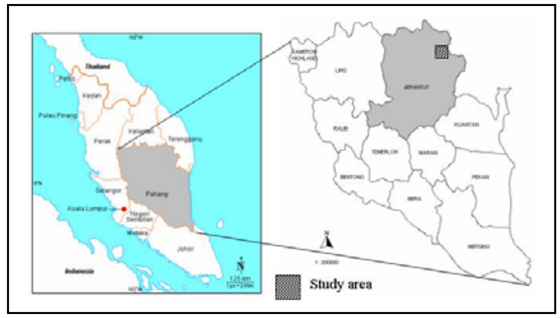

Figure 1: Location of study area

2.1.1 Classification and ground survey

The classification schemes used in this study were used to allow the analysis of satellite imagery and the purpose of the classification scheme is to provide primary formation about forest land cover and other non forest features such as rivers and the existing forest road system. Consequently, the classification scheme complies with the local classification for forestry purposes. The classification scheme is shown in Table 1.

Table 1: Land cover classification scheme.

No.	Main Class	Description
01.	Primary forest	Medium to large crown. High-density canopy cover >50%. This class remaining of the natural forest formation (*virgin forest*) and had no intervention.
02.	Logged over forest	Sparse /medium crown. Low-density canopy cover (<10-50%). This class refers to the area in which harvesting operations have taken place under the Malaysian selective management system (SMS).
03.	Agricultural	Sparse fragmented

	crop/mixed horticulture	(forest fraction 10-70%). This class is includes the area with self-plantation of small trees by villagers and *orang asli* (aborigines). This includes fruit trees for self-consumption.
04.	Water bodies	This class covers area by the main river which crosses the study area and also the reservoir
05.	Bare land	Refers to areas of exposed soil with very little or without vegetation coverage including the forest road network, forest camp and logyard area.

The ground survey was conducted in March 2003 (3 weeks). In this study, the numbers of sample segments adopted is only 97 instead of 100 due to cloud problems in three of them. The 97 sample segments were distributed unaligned systematically random over the 100 square km frame area, and represents a sampling frequency of 5.59 percent of the 100 square km area (Figure 2). A set of these samples was chosen at random using the MS EXCELL random generator. The sample segment adopted was 240m by 240m (5.76 ha) within each 1km by 1km block and a total of four observation sites (sub-sample) were made in every sample segment. Observation was made in the 60m by 60m area in the four corners of the sample segment. This size was chosen because it is adequate to carry out field survey and appropriate to enclose a land cover variation in the test site using Landsat TM. A photographic image of the segments were enlarged to 1:10 000 scale. Enumerators from the Pahang Foresty Department carried out annotation work of the field numbers, location and land cover type in the sample units. The image sample units were transparent overlay with transparency film to draw field boundaries within the sample units.

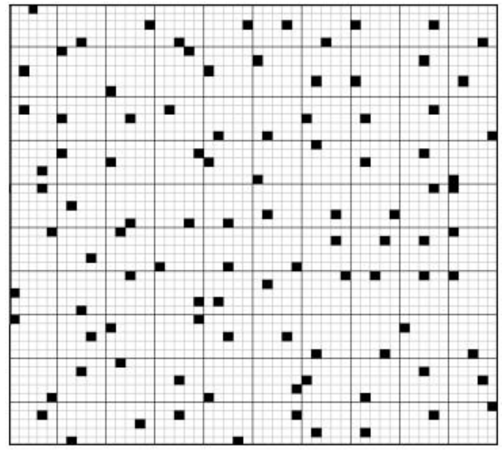

Figure 2: The area frame design illustrating the distribution of the sample segment (filled black) using unaligned systematic random sampling over the study area (10 km by 10 km).

2.1.2 Image analysis phase

Landsat TM data was acquired on 08 May 2001 also took into account image quality and a cloud cover of less than 5%. Selected Landsat TM image geometrically corrected using image-to-image approach and registered to UTM coordinates. The GCPs were taken from the previous Landsat TM image (master data) for the same area. Pure pixels of land cover were determined as training pixels, spectral signatures of the land cover were evaluated, and classification is performed using maximum likelihood. In the classification, the Battacharrya Distance was used to examine the quality of training sites and class signatures. This panel contains all the available information about signature and class information for each class. The value shows in Battacharrya Distance is a value between 0 and 2, where 0 indicates complete overlap between the signature of two classes and 2 indicates a complete separation between two classes (PCI, 1997). The larger the separability values achieved, the better final classification result. Statistically, it resulted average of the signature separability of 1.88, minimum separability of 1.588 and maximum separability of 1.97. As a result, this pre-clustering shows a significant improvement of the classification. After classification was made, we have two different images from each segment as a sample: the proportion of segments from ground survey and proportion of pixels from classified imagery. The regression method then used in order to test the reliability of regression estimators.

2.1.3 Area estimation formula

Area estimation from sample survey data can be calculated by measuring the area of the cover class of fixed size known as sample units (Taylor and Eva, 1992; Taylor et al., 1997). The estimates were computed as a proportion rather than absolute area because the errors resulting from drawing or digitizing and map scale can be minimized. The unbiased estimate of land cover proportion of land area covered by class c is given by the equation:

$$\bar{y}_c = \frac{1}{n}\sum_{i=1}^{n} y_i$$

with variance

$$\text{v}\,Var(\bar{y}_c) = (1-\tfrac{n}{N})\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y}_c)^2$$

where: y_i is the proportion of segment i covered by class c; N = total number of segments in the region, n = number of segments in the sample. The proportion of the study region sampled $(n/N)$is referred to as the sample fraction. When this is less than 5%, the correction factor for a finite population $(1-\tfrac{n}{N})$ can be omitted from the above formula (Cochran, 1977). The estimate of the class area is:

$$\hat{Z}_c = D\,\bar{y}_c$$

with variance: $Var(\hat{Z}_c) = D^2 Var(\bar{y}_c)$

where D is the area of the region.

The standard error or accuracy of $\hat{Z}_c$ is estimated by calculating the 95% confidence interval as follows:

$$\hat{Z}_c \pm 1.96\sqrt{\text{var}(\hat{Z}_c)}$$

The equivalent proportions determined from digital classified image (p_i) were calculated to establish pairs of observations that could be plotted to establish relationships. The relationship between the two data sets of information is determined by linear regression of y on p given by:

$$y = \bar{y} + b(p - \bar{p})$$

where, $\bar{y}$ and $\bar{p}$ are the sample mean values and b is the slope of the regression line.

Then, the classified images from the 100 sample segments were extracted to produce the necessary data sets for regression estimator. For each class, a linear regression was applied to correlate the image sample segments with the equivalent class of sample segments acquired from sample survey.

The population estimate, $\bar{p}_{pop}$ of the digitally classified land cover proportion, for each class for the whole area of interest can be calculated by:

$$\bar{p}_{pop} = \frac{\text{Total pixel area classified as class c}}{\text{Region area}}$$

The value is then used in the regression equation to produce the correction for the sample estimate.In the case of the satellite data, $\bar{p}$ can also be estimated from the classified imagery by calculating the proportion of pixels classified the land cover class in the 240m by 240m cells. This is called population estimates, $\bar{p}_{pop}$. Population estimate is the proportion of pixels classified as the cover types in the entire study area. Then, this value is used in the regression equation to produce a correction for the sample estimate of the mean cover types proportion per unit area, $\bar{y}$, and is known as the regression estimate, $\bar{y}_{reg}$, and as given by:

$$\bar{y}_{reg} = \bar{y} + b(\bar{p}_{pop} - \bar{p})$$

For large random sample where n>50 the variance is approximately as given by:

$$Var(\bar{y}_{reg}) = \frac{1}{n} Var(y)(1 - r_{py}^2)$$

where r_{py}^2 is the coefficient of determination.

The estimate of cover types area in the study region and its variance are then:

$$\hat{Z}_{reg} = D\bar{y}_{reg}$$

and $$Var(\hat{Z}_{reg}) = D^2 Var(\bar{y}_{reg})$$

The standard error S.E.(Z_{reg}) and the 95% confidence interval (C.I.$_{95\%}$) are given by:

$$S.E.(Z_{reg}) = \sqrt{Var\ (Z_{reg})}$$

$$C.I._{95\%} = Z_{reg} \pm 1.96\ S.E.\ (Z_{reg})$$

3 Results and discussion

3.1 Ground survey result

The results of the ground survey should provide approximate proportions of each land cover type in the whole study area. Using the equation 3, the land cover area estimates were calculated. A summary of land cover area estimates by sample survey was presented in Table 2. The table includes standard error, coefficient of variation and 95% confidence interval. It can be seen that area estimation results are as follows; Primary Forest–6762.87 ha, Logged Over Forest-2783ha, Bareland–329.89 ha, Agric.Crop/Mix. Hort.–134.02 ha. and Water Bodies–164.95 ha., respectively. The class area proportions from the ground survey and the class area proportion from the classified satellite imagery were then regressed, and presented in the next section

Table 2: Summary land cover area proportions and area estimates by ground survey

No. of S.Sgmt 97	P. Forest	L.O. Forest	B. Land	A.C/M. Hort.
$\sum y_{ic}$	65.6	27.0	3.2	1.3
Proportion ($\bar{y}_c$)	0.6762	0.2783	0.0329	0.0134
Area (Z_c)[ha]	6762.87	2783.50	329.89	134.02
Std.Error [ha]	1676.43	289.47	128.75	36.76
C.V (%)	24.78	10.39	39.02	27.42
Z_c+1.96SE[ha]	10048.67	3350.86	582.24	206.06
Z_c-1.96SE [ha]	3477.06	2216.13	77.54	61.97

*S.Sgmt=Sample Segment, P.Forest=Primary Forest, L.O.Forest=Logged Over Forest, B.Land=Bareland, A.C/M.Hort.=Agricultural Crop/Mixed Horticulture, W.Bodies=Water bodies, Std.Error=Standard Error, C.V=Coefficient of Variation(Std.Error/Expanded area)*100, $\sum y_{ic}$ * 100*

3.2 Regression estimator result

To calculate the regression estimator, a total of 97 samples were examined and measured to determine their proportions. Generally a good relationship exists between the sample survey data and digital classified imagery for all classes, producing coefficient of determination (r^2) of more than 0.80. However the regression relationship for A.C./M.Hort and W.Bodies should be given special attention although their relationship was greater with r^2 of 0.96 and 0.89. The fitted lines for both classes were generated from only two numbers of area data due to the sample selection based on unaligned systematic random sampling, thus the sample area can be picked up from the sample segment which suffered from insufficient data. Because the main interest in the survey is to map and record forest cover area information, the other classes not emphasized. However, in all cases of the correlation, the regression relationship was highly significant with $p < 0.01$.

Area estimations of each land cover derived using regression estimator method are shown in Table 3. Area estimations using the regression estimator were compared with those derived by sample survey. The coefficient of variation decreased as a result of using the regression estimator method. For all land cover types the regression estimator method calculation appears to have a greater precision for all classes (C.V. range from 3.99% - 36.12%) than the sample survey method (C.V. range from 10.39% - 39.02%). The decrease could be noted in all classes ranging from 3.99% for Logged Over Forest to 36.12% for Bare Land. The statistical data from the regression estimator revealed and demonstrate the improvement in the accuracy of the area estimates by adjusting the estimate of the mean land cover area proportions, thus reducing the variance. The use of the regression estimator by mean produced more unbiased area estimates and showed more precise results.

Table 3: Area estimation by regression estimator method

Class	Area (Z_{reg}) [ha]	*b*	r^2	S.E [ha]	C.V (%)	C.I.95% [%]	C.I.95% [ha]
P.For	6743.42	0.063	0.86	1208.55	17.92	35.12	9112.17 4374.66
L.O. For	2775.16	0.027	0.89	110.80	3.99	7.82	2992.32 2557.99
B. L	155.89	0.000	0.86	56.32	36.12	70.8	266.27 45.50
A.C/ M. H	133.46	0.002	0.96	29.56	22.14	43.40	191.39 75.52
W.B	61.77	0.000	0.89	11.06	17.90	35.08	83.44 40.09

P.Forest=Primary Forest, L.O.Forest=Logged Over Forest, B.Land=Bareland, A.C/M.Hort.=Agricultural Crop/Mixed Horticulture, W.Bodies=Water Bodies,
C.V = S.E/Area(Z_{reg})*100

4 Conclusions

Resource evaluation is not only considered in a statistical sense but also looks at the capability of the data to show trends and to discriminate between groups of classes of interest in forest management. The result from this study enables the use of remote sensing data and sample survey through the regression estimator technique to forest area in Malaysia. In conclusions it can be drawn that the use of this technique has an advantage. The technique results in an increase in the precision and reduces the error of variance and is an improvement, when compared to sample survey estimation. The coefficient of determination illustrated that for all land cover types, r^2 was high ranging from 0.86 for Primary Forest to 0.96 for Agricultural Crop/Mixed Horticultural. Meanwhile coefficient of variation shows a decrease in percent for all land cover types ranging from 3.99% for Logged Over Forest to 36.12% for Bare Land. The use of regression estimator from sample survey and classified image sample is generally reliable, acceptable and proven to be effective in forest area estimation in this study. The accuracy of the result largely depends on the quality of the data obtained from ground survey and image classification work. Although Landsat TM data used in this study has provided a broad definition of forest cover, such data is useful to the Forestry Department to evaluate and monitor the existing forest resource for further management and socio economic planning.

Acknowledgements
The authors are grateful to the Malaysian Centre for Remote Sensing (MACRES) and Pahang Forestry Department of Peninsular Malaysia for supplying data and supporting the field work for this study. To Dr. Graham Thomas and Mr. Tim Brewer at Cranfield University, UK thanks for giving an invaluable discussion in this work. The authors are also grateful for the comments provided by the referees that helped to enhance this article.

References:
[1] Cochran, W.G., 1977, *Sampling technique*. John Wiley and Son. 413p.
[2] Deppe, P. 1994. Application of remote sensing and GIS for management and planning forestry resources in southern Brazil. Ph.D (Thesis), Cranfield University,Silsoe,UK. 333p.
[3] Ference, C., Bognar, P., Lichtenberger, J. Hamar, Tarcs,I,G.,Timar,G.,Molnar,G.,Pasztor, R.S., Steinbach,P.,Szekel, Y.B., Ference, O.E., and Ferencez-Arkos, I., 2004, Crop yield estimation by remote sensing. *International Journal of Remote Sensing*, 25,4113-4149.
[4] Gallego,F.J., and Delince,J., 1993. Crop area

estimation through remote sensing: stability of the regression correction. *International Journal of Remote Sensing*, 14,3433-3445.
[5]Gonzales-Alonso,F., and Cuevas, J.M. 1993. Remote sensing and agricultural statistic: crop area estimation through regression estimators and confusion matrices. *International Journal of Remote Sensing*, 14,1215-1219.
[6] Gonzales-Alonso,F., Lopez-Soria,S., and Cuevas-Gozalo,J.M., 1991, Comparing two methodologies for crop area estimation in Spain using Landsat TM images and ground–gathered data, *Remote Sensing of Environment*,35,29-35.
[7] Hyypa,J., Hyypa,H., Inkinem,M., Engdahl,M., Linko,S., and Zhu,Y.H., 2000, Accuracy comparison of various remote sensing data sources in the retrieval of forest stand attributes. *Forest Ecology and Management*, 128, 109-120.
[8] Kamaruzaman,J., and Souza, G.D., 1997. Use of satellite remote sensing in Malaysia and its potential. *International Journal of Remote Sensing*, 18, 57-70.
[9] Khali,A.H., 2001, Remote sensing, GIS and GPS as a tool to support precision forestry practices in Malaysia. *Paper presented at the 22nd Asian Conference on Remote Sensing*, 5-9 November 2001, Singapore.5p.
[10] Mohd Hasmadi, I, and Kamaruzaman ,J., 1999, Use of satellite remote sensing in forest resource management in Malaysia. *Paper presented at Second Malaysian Remote Sensing and GIS Conference*, 16-18 March, 1999, ITM Resort & Convention Centre, Shah Alam, Selangor, Malaysia. 24p.
[11] PCI, 1997, PCI 7.0.1. Image analysis software, Help Menu. PCI, Toronto,Canada.
[12] Taylor,J.C., and Eva,H.D., 1993, Operational use of Remote Sensing for Estimating Crop Area in England. In: K.Hilton. Towards Operational Application. Proceeding of the 19th annual Conference of Remote Sensing Society, Chester College,UK.
[13] Taylor.J,C., Sannier,C., Delince,J., and Gallego,F.J., 1997, Regional Crop Inventories in Europe Assisted by Remote Sensing: 1988-1993. Synthesis Report of the MARS Project-Action 1.Joint Research Centre, European Commission-EUR 17319 EN.71p.