P. F. Crapper
*Division of Land Use Research, CSIRO*
*Canberra City, A.C.T. 2601, Australia*

# Errors Incurred in Estimating an Area of Uniform Land Cover Using Landsat

A formula for the variance in estimating an area, due to including or excluding a pixel within the boundary of that area, is developed.

Abstract: *For many years the simplest means of determining the area of a region has been to count the number of points, of a randomly located overlay, enclosed by the boundary. This technique has much in common with the method of determining the area of a region identified by Landsat as being of uniform land cover. In both cases errors of commission or omission occur at the boundary depending on whether the perimeter cells are included or excluded. A formula has been developed for the Landsat pixel which gives the variance of this area estimate. Relative errors of 1 percent have been found for areas of 132 ha, 5 percent for areas of 15 ha, and 10 percent for areas of 6 ha.*

## Introduction

The technique of determining the area of a region by counting the squares or the points enclosed by the boundary has been used for many years. These techniques have been used in disciplines ranging from engineering, where they have been used to calculate forces and moments, to forestry. The fundamental difference between the method of counting squares and the method of counting points is that with points there is always an integer number whereas with squares fractional values are included. The apparent loss of accuracy in point counting techniques is compensated for by treating the count as a form of statistical sampling and, repeating the measuring process several times, using random locations of the point overlay, and averaging the results. Instruments have been developed which allow automatic counting and marking of counted points. Frolov and Maling (1969) have performed a theoretical error analysis of the point counting method. Bonner (1975) has determined the error of area estimates, when a point grid is used, by performing repeated overlays and deriving empirical relationships from the experimental results. Photographic techniques have recently been developed for superimposing the points on the original (see Mathews and Mason, 1979). If the point grid can be overlaid a number of times, then an accurate area estimate can be obtained. However, there are situations in which a single fixed overlay occurs, and so significant errors can occur in such an area evaluation. One such situation is when Landsat scans an area of uniform land cover.

The theory of the point counting technique is also relevant to automated geographic data processing where a grid cell is either counted or not counted in an area evaluation. In automated geographic data capture a raster scanning device scans an area of land, divides the land into regular grid cells, and records one or more spectral signals for each grid cell. The signals recorded can be used at a later time to help determine the type and area of a particular class of land cover. The important feature is that the grid cells must be included or excluded in their entirety as there is no information at the sub grid cell level.

The Landsat series of satellites have been extensively used for determining ground cover and monitoring temporal change. These satellites record spectral information in four bands, two in the visible region and two in the near infra-red region of the electromagnetic spectrum. The effective size of each grid cell (or pixel) is 57.10 by 79.06 m (or 0.45 ha) on the ground. The reflectance in the east-west direction is spatially integrated over 79 m

Photogrammetric Engineering and Remote Sensing,
Vol. 46, No. 10, October 1980, pp. 1295-1301.

but with a sampling interval of 57.1 m. Therefore, to allow a common north-south boundary for adjacent pixels in the same scan line, the pixel is treated as being 57.10 m wide. Thus, each pixel in a scan line contains information derived from the two adjacent pixels in that line (for this reason the pixel is often stated to be about 80 by 80 m). It should also be remembered that atmospheric effects can further reduce the resolution of the scanner, although the reduction is difficult to quantify.

The spectral information can be used to allocate (or classify) a pixel to a land-cover category. The classification process can be supervised (delineation of spectral and textural classes by man followed by machine searching for similar classes) or unsupervised. In each case, it is hoped that the derived spectral and textural signatures of distinctive land covers do not overlap. The area of a particular type of land cover is then determined by counting the number of pixels whose signatures fall within predetermined limits and multiplying by the area of a pixel. This technique has been used extensively for several years (see Bauer *et al.*, 1979). In this paper an error variance for this area estimate is determined.

Similar errors arise when a land-cover map is being coded to grid cells (called the coding problem). The coding of biophysical data has become increasingly widespread. These data are stored in a computer in an array format and used as the input to land suitability, planning, or management models (see Miller and Carter, 1979).

The errors arise in the perimeter cells, where a mixed signature occurs, and, depending on the tolerances of the classifier and of that particular spectral class and the type of classifier used, this pixel may or may not be included in the region of uniform land cover. For regions in which the region-to-pixel ratio is comparatively small this error can be quite significant. To make matters worse, these perimeter cells with mixed spectral signatures (which have been called "mixels", see Jupp *et al.*, 1979) are sometimes not classified with the land cover on either side of the boundary but erroneously mislabelled as a separate class of land cover. There are many more perimeter cells than one would intuitively expect. Jupp *et al.* found that, for a mapping exercise in the Batemans Bay area of New South Wales, over 50 percent of pixels were mixels. For example, in Figure 1 an enclosed comparatively regular region is shown overlaid by a square grid. For this case if the grid cell area is 0.45 ha (as for Landsat) then the region area is 40 ha, i.e., the grid cell is 1.13 percent of total area. However, the number of perimeter cells is 45 percent of the total number of cells (where a perimeter cell is defined as a grid cell through which a section of the boundary passes).

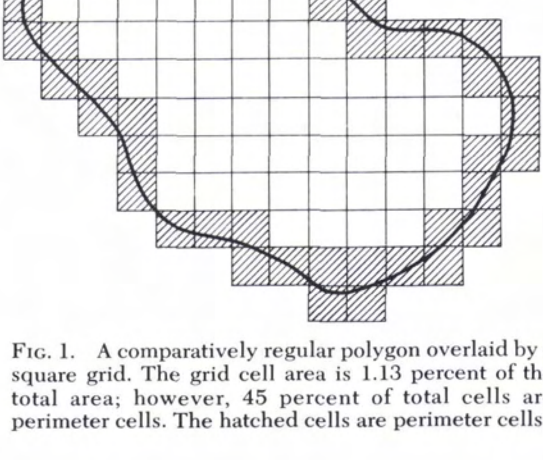

FIG. 1. A comparatively regular polygon overlaid by a square grid. The grid cell area is 1.13 percent of the total area; however, 45 percent of total cells are perimeter cells. The hatched cells are perimeter cells.

### Random Lines and Rectangles

The error, which occurs in the estimation of the area of a region with uniform land cover, will be confined to the perimeter pixels. Following Goodchild and Moy (1976), the number of perimeter pixels is estimated as a function of the total area. If $A$ and $L$ are the exact area and perimeter of the region, respectively, then

$$L = 2K_1 \sqrt{\pi A} \qquad (1)$$

where $K_1$ is the shape factor. The shape factor is a measure of how contorted is the region (see Crapper, 1980 for a discussion of the parameter). A circle which has the minimum perimeter for a given area has $K_1 = 1$ whereas for a square $K_1 = 1.128$. Higher values of $K_1$ occur for regions having longer perimeters for a given area.

Now when a grid with rectangular cells is overlaid on the region, the total perimeter, $L$, is given by

$$L = \sum_{i=1}^{i=N_b} l_i,$$

where $l_i$ is the length of the i^th^ boundary pixel and $N_b$ is the number of boundary pixels. The average length of perimeter per boundary pixel $\bar{l}$ is given by

$$E\,(l_i) = \bar{l} = \frac{L}{N_b} = K_2\bar{L} \qquad (2)$$

where $K_2$ is a measure of the average within pixel distortion and $\bar{L}$ is the average length of a straight line laid across a rectangle. Thus, the average number of perimeter pixels, $N_b$, is given by

$$N_b = \frac{L}{\bar{l}} = \frac{2\,K_1\,\sqrt{\pi \mathrm{A}}}{K_2\bar{L}}\,. \qquad (3)$$

Of the $N_b$ perimeter pixels, on average half will be recorded as belonging to the region under consideration and half will be recorded as belonging to adjacent regions. In the former case an error of commission occurs in which the region area will be overestimated and in the later case an error of omission occurs in which the region area will be underestimated. However, in both cases the error is the smaller section of the pixel which has been divided by the boundary and so for any one pixel the maximum error is half a pixel.

The error variance for one pixel is equal to the mean square of the smaller area obtained when a random straight line is laid across a pixel ($\overline{A^2}$) and hence the overall error variance $\sigma^2$ is given by

$$\sigma^2 = N_b\overline{A^2} \tag{4}$$

The error variance is thus dependent on the number of pixels and the distortion parameters, $K_1$ and $K_2$, and is given by

$$\sigma^2 = \frac{2\,K_1\,\sqrt{\pi A}}{K_2}\,\frac{\overline{A^2}}{\bar{L}}\,. \tag{5}$$

## Problems Associated with Random Lines

In order to evaluate the error variance, it is thus necessary to calculate the average length of a random line laid across a rectangle and the mean square area under it. Initially this problem may appear simple; however, some reading in the field of geometric probability (see Kendall and Moran, 1963) would soon dispel this belief. The difficulty is that, for many apparently simple problems, there appears to be more than one 'correct' answer depending on how the randomness is defined.

One of the classic problems in the field is the Bertrand problem. The problem is to find the probability that a 'random chord' of a circle of unit radius has a length greater than $\sqrt{3}$, the side of an inscribed equilateral triangle. Kendall and Moran (1963) present three solutions to this problem, viz. ¼, ⅓, and ½, and conclude by saying that all solutions are correct but they relate to different problems depending on how the random chord was defined. Each of the three previous solutions can be defined by a different joint probability distribution. Thus, to define the distribution of a geometric element, one must first determine a system of coordinates which defines the element uniquely, and then define a probability distribution on the range of those coordinates.

There are some problems in which insufficient information is provided to allow one to determine a system of coordinates which uniquely defines the element (so called ill-posed problems). In such cases additional restrictive assumptions can be made to determine such a system of coordinates. An alternative approach is to seek solutions which are invariant under the transformations of translation, rotation, and reflection. Kendall and Moran (1963) conclude that, for the Bertrand problem, there is only one solution (½) which is invariant to the transformations of translation, rotation, and reflection.

Jaynes (1971) has re-examined the Bertrand problem in considerably more detail than Kendall and Moran (1963). Jaynes also found that, if one seeks solutions which are invariant under the transformation group, then unique solutions could be determined for apparently ill-posed problems. His invariance arguments and experiments verified the solution quoted by Kendall and Moran (1963).

The problem of the average length of a random line laid on a unit square is also ill-posed. Two separate analytical methods have been found in the literature for this problem giving two distinct answers. Only one of the methods, fortunately, was found to be invariant under transformations of translation, rotation, and reflection, and this is the method of Goodchild and Moy (1976). In order to check the result, a computer method was developed which was also invariant to the transformation group. In this method the random number generator was used to define a point on the line and the gradient. Using this method and a large number of trials, of which only a comparatively small number intersected the square, a satisfactory agreement with the Goodchild and Moy result was obtained. The method to be used in Appendix A for determining the average length of a random line laid on a rectangle is based on an extension of the Goodchild and Moy method.

## Determination of Error Variance for Landsat

The error variance $\sigma^2$ is found by substituting values for $\overline{A^2}$ and $\bar{L}$ (as calculated in Appendix A) into Equation 5 and is given by

$$\sigma^2 = 0.0848\frac{K_1}{K_2}\sqrt{A}\ (\text{ha}^2) \tag{6}$$

where $A$ is measured in hectares. Now the shape factor, $K_1$, of an area of land cover, deemed uniform by Landsat, depends on many factors including previous history and present vegetation, landform, soils, etc., and it would be impossible to theoretically derive an average value. Crapper (1980) has determined shape factors for 1605 regions defined by a polygonal data base on the south coast of New South Wales. These regions were defined on the basis of relative homogeneity in the spatial pattern of the biophysical properties. The regions covered a total area of 6000 km² and included significant variations in landform, vegetation, and soils. Crapper found that the shape factor varied with landform and varied slightly with area. Ignoring the variation with region area and averaging the values for different landforms, $K_1 = 1.82$. As the characteristics which delineate areas of uniform land cover, according to Landsat, and the above regions are similar and as $K_1$ has been assumed to be independent of area, it seems reasonable that the processes which determine the shape of the perimeter are common to both and hence the average shape factors are approximately equal.

The within cell distortion parameter, $K_2$, is even more difficult to evaluate as it depends on the region to pixel area ratio as well as $K_1$. Also, the length of many biophysical boundaries depends on the scale at which they are measured. The finer the scale the greater the length. Mandelbrot (1977) has called those curves, which do not approach a limit as the scale is made finer, fractals. For our present purposes the scale of resolution is one pixel, and if the region-to-pixel ratio is comparatively large, then the segment of the perimeter within each pixel can be accurately represented by a straight line, i.e., $K_2 = 1$. For most land cover mapping exercises the region to pixel area ratio is quite high and so $K_2 \simeq 1$. Hence Equation 6 reduces to

$$\sigma^2 = 0.15 \sqrt{A} \text{ (ha}^2) \qquad (7)$$

where $A$ is in ha. The value of $A$, calculated by counting pixels, is an unbiased estimate of the area and so it can be used in Equation 7.

## Discussion and Conclusion

The relative error (sometimes called the standard error or S.E.) is given by

$$\frac{\sigma}{A} = 0.39\, A^{-3/4}.$$

This curve is displayed on Figure 2 from which it can be seen that for a relative error of 1 percent $A = 132$ ha, for a relative error of 5 percent $A = 15$ ha, and for a relative error of 10 percent $A = 6$ ha.

In the above calculation an average value has been used for the shape factor (i.e., $k_1 = 1.82$). There are, however, some situations in which reasonably accurate estimates can be made for the shape factor. One such situation is the shape of the areas used for wheat production. Typically, wheat fields in the United States are rectangular with aspect ratio 5 and average area 500 ha, in the U.S.S.R. and Australia wheat fields are typically square with areas 1000 ha and 150 ha, respectively, and in China wheat fields have irregular shape with typical areas of 2 ha. The shape factor can be calculated from Equation 1 and, after substituting into Equation 6, the relative error can be found. The shape factors, areas, and relative errors in area estimates are shown in Table 1. For smaller areas, as typified by Chinese wheat fields, the relative errors in area estimates are so large that great caution should be exercised when working with area estimates.

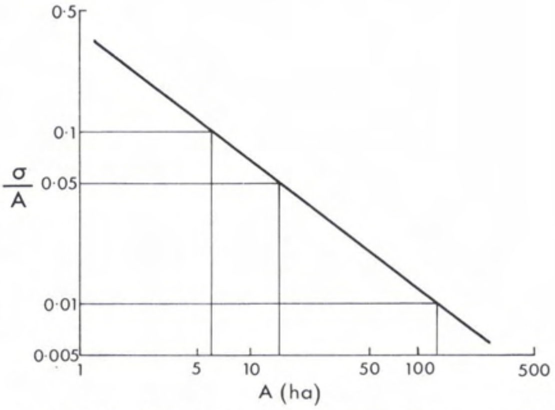

FIG. 2. The log of the relative error ($\frac{\sigma}{A}$) is plotted against the log of the area, $A$, in ha.

In conclusion, the Landsat series of satellites provides a convenient and comparatively accurate method for estimating areas of uniform land cover. However, when the areas are small or when the boundaries are highly contorted (e.g., lakes or coastal boundaries), considerable care must be exercised before using area estimates because the relative errors in the area estimate may be very high.

## Acknowledgments

The author acknowledges with gratitude many helpful discussions with Dr D. L. Jupp and helpful comments made by Dr A. A. Green and Dr F. R. Honey.

TABLE 1. THE SHAPE FACTOR, TYPICAL AREA, AND RELATIVE ERROR IN AREA ESTIMATE OF WHEAT FIELDS

| Country | Shape factor | Typical area (ha) | Relative error (%) |
|---|---|---|---|
| USA | 1.51 | 500 | 0.33 |
| USSR | 1.13 | 1000 | 0.17 |
| Australia | 1.13 | 150 | 0.72 |
| China | 1.82 | 2 | 23 |

## Appendix A

### Evaluation of $\bar{L}$ and $\bar{A}^2$

Let us consider a random straight line laid across a rectangle. With no loss of generality, the rectangle can be considered fixed with the longer side horizontal, and one vertex can be treated as the reference vertex. If $a$ is the shorter side, $b$ the aspect ratio, $a \times b$ (subsequently written $ab$) the longer side, and $\alpha$ is the angle between the random line and the base, the lines being extended if required, then there is a uniform probability of $\alpha$ falling within the range $0 < \alpha < 180°$. By symmetry we need consider only the range $0 < \alpha < 90°$. Furthermore, this range can be subdivided into Part I, $0 < \alpha < \alpha_1$, and Part II, $\alpha_1 < \alpha < 90°$, where $\tan \alpha_1 = 1/b$. For convenience, Part II has been redefined to $0 < \alpha < \alpha_2$ with the longer side vertical, where $\alpha_2 = \pi/2 - \alpha_1$. The problem will be treated in these two parts and the results will be combined at the end.

PART I, $0 < \alpha < \alpha_1$

If $L$ is the length of the random line, $A$ is the area under it, and $R$ is the perpendicular distance from the line to the reference vertex, as shown in Figure A-1(a–d), then for $0 < R < ab \sin \alpha$

$$\begin{aligned} L_1 &= R \tan \alpha + R/\tan \alpha \quad \text{and} \\ A_1 &= R^2/(2 \sin \alpha \cos \alpha) \end{aligned}$$

and for $ab \sin \alpha < R < (ab \sin \alpha + a \cos \alpha)/2$
(by symmetry, we only need consider half the rectangle for the calculation of $\bar{\mathrm{L}}$ and only want to consider half the rectangle for the calculation of $\bar{\mathrm{A}}^2$).

$$\begin{aligned} L_2 &= ab/\cos \alpha \\ A_2 &= (ab/\cos \alpha)\,(R - ab \sin \alpha/2) \end{aligned}$$

where $\alpha_1 = \tan^{-1} 1/b$

Thus,

$$\bar{L}_1(\alpha) = \frac{2}{ab \sin \alpha + a \cos \alpha}\left(\int_0^{ab \sin \alpha} L_1 \, \mathrm{d}R + \int_{ab \sin \alpha}^{\frac{ab \sin \alpha + a \cos \alpha}{2}} L_2 \, \mathrm{d}R\right)$$

$$= \frac{ab}{b \sin \alpha + \cos \alpha}$$

$$\bar{L}_1 = \frac{1}{\alpha_1}\int_0^{\alpha_1} \bar{L}_1(\alpha)\, \mathrm{d}\alpha$$

$$= \frac{ab}{\alpha_1 \sqrt{1 + b^2}} \ln \frac{\tan \alpha_1}{\tan \alpha_1/2}$$

$$\overline{A_1^2}(\alpha) = \frac{2}{ab \sin \alpha + a \cos \alpha}\left(\int_0^{ab \sin \alpha} A_1^2 \, \mathrm{d}R + \int_{ab \sin \alpha}^{\frac{ab \sin \alpha + a \cos \alpha}{2}} A_2^2 \, \mathrm{d}R\right)$$

$$= \frac{a^4 b^2}{(b \sin \alpha + \cos \alpha)\, 60 \cos^2 \alpha}(b^3 \sin^3 \alpha + 5 \cos^3 \alpha)$$

$$\overline{A_1^2} = \frac{1}{\alpha_1}\int_0^{\alpha_1} \overline{A_1^2}(\alpha)\, \mathrm{d}\alpha$$

$$= \frac{a^4 b^2}{60\, \alpha_1}\left(b^2(\tan \alpha_1 - \alpha_1) + b \ln \cos \alpha_1 + \alpha_1 + \frac{4}{1 + b^2} (\alpha_1 + b \ln(\cos \alpha_1 + b \sin \alpha_1))\right)$$

PART II, $0 < \alpha < \alpha_2$

The meaning of the symbols is explained in Figure A-1(e–h). For $0 < R < a \sin \alpha$

$$L_3 = R \tan \alpha + \frac{R}{\tan \alpha}$$

$$A_3 = \frac{R^2}{2 \sin \alpha \cos \alpha}$$

and for $a \sin \alpha < R < \dfrac{ab \cos \alpha + a \sin \alpha}{2}$

$$L_4 = \frac{a}{\cos \alpha}$$

$$A_4 = \left(\frac{a}{\cos \alpha}\right)\left(R - \frac{a \sin \alpha}{2}\right)$$

where $\alpha_2 = \tan^{-1} b$.

Thus,

$$\bar{L}_2(\alpha) = \frac{2}{ab \cos \alpha + a \sin \alpha}\left(\int_0^{a \sin \alpha} L_3 \, \mathrm{d}R + \int_{a \sin \alpha}^{\frac{ab \cos \alpha + a \sin \alpha}{2}} L_4 \, \mathrm{d}R\right)$$

$$= \frac{ab}{b \cos \alpha + \sin \alpha}$$

$$\bar{L}_2 = \frac{1}{\alpha_2} \int_0^{\alpha_2} \bar{L}_2(\alpha) \, \mathrm{d}a$$

$$= \frac{ab}{\alpha_2 \sqrt{1 + b^2}} \ln \frac{\tan \alpha_2}{\tan \alpha_2/2}$$

$$\overline{A_2^2}(\alpha) = \frac{2}{ab \cos \alpha + a \sin \alpha}\left(\int_0^{ab \sin \alpha} A_3^2 \, \mathrm{d}R + \int_{ab \sin \alpha}^{\frac{ab \cos \alpha + a \sin \alpha}{2}} A_4^2 \, \mathrm{d}R\right)$$

$$= \frac{a^4}{(b \cos \alpha + \sin \alpha)\, 60 \cos^2 \alpha}(5\, b^3 \cos^3 \alpha + \sin^3 \alpha)$$

$$\overline{A_2^2} = \frac{1}{\alpha_2} \int_0^{\alpha_2} \overline{A_2^2}(\alpha) \, \mathrm{d}\alpha$$

$$= \frac{a^4}{60\, \alpha_2}\Bigg(\tan \alpha_2 - \alpha_2 + b \ln \cos \alpha_2 + b^2 \alpha_2 + \frac{4}{1 + b^2}$$
$$(b\, \alpha_2 + \ln (\frac{b \cos \alpha_2 + \sin \alpha_2}{b}))\Bigg)$$

Now, as random lines over the range 0 to $\alpha_1$ and 0 to $\alpha_2$ (or $\alpha_1$ to $\pi/2$) are both equally likely, the average length and mean square area under the random line are as follows:

$$\bar{L} = \frac{2}{\pi}(\alpha_1 \bar{L}_1 + \alpha_2 \bar{L}_2)$$

$$= \frac{2\, ab}{\pi \sqrt{1 + b^2}} \ln \left(\frac{1}{\tan(\alpha_1/2) \cdot \tan(\alpha_2/2)}\right) \qquad \text{(A-1)}$$

$$\overline{A^2} = \frac{2}{\pi}(\alpha_1 \overline{A_1^2} + \alpha_2 \overline{A_2^2})$$

$$= \frac{a^4}{30 \pi}\Big(b^3 (1 + \ln \cos \alpha_1) + b\, (1 + \ln \cos \alpha_2) +$$

$$\frac{\alpha_1 (5b^2 - b^6) + \alpha_2 (5b^4 - 1)}{1 + b^2} + \frac{4\, b^3}{1 + b^2} \ln\, (4 \cos \alpha_1 \cos \alpha_2)\Bigg) \qquad \text{(A-2)}$$

For the simple case of random lines laid over a square ($b = 1$), Equations A-1 and A-2 reduce to

$$\begin{aligned} \bar{L} &= 0.7935\, a \text{ and} \\ \bar{A}^2 &= 0.0619\, a^4, \end{aligned}$$

which agree with the results of Goodchild and Moy (1976). For a Landsat pixel $a = 57.10$ (m) and $b = 1.3846$ and, hence,

$$\begin{aligned} \bar{L} &= 0.9288\, a \\ &= 53.036 \text{ (m) and} \\ \overline{A^2} &= 0.1194\, a^4 \\ &= 1.269 \times 10^6 \text{ (m}^4\text{)} \\ &= 1.269 \times 10^{-2} \text{ (ha}^2\text{)}. \end{aligned}$$

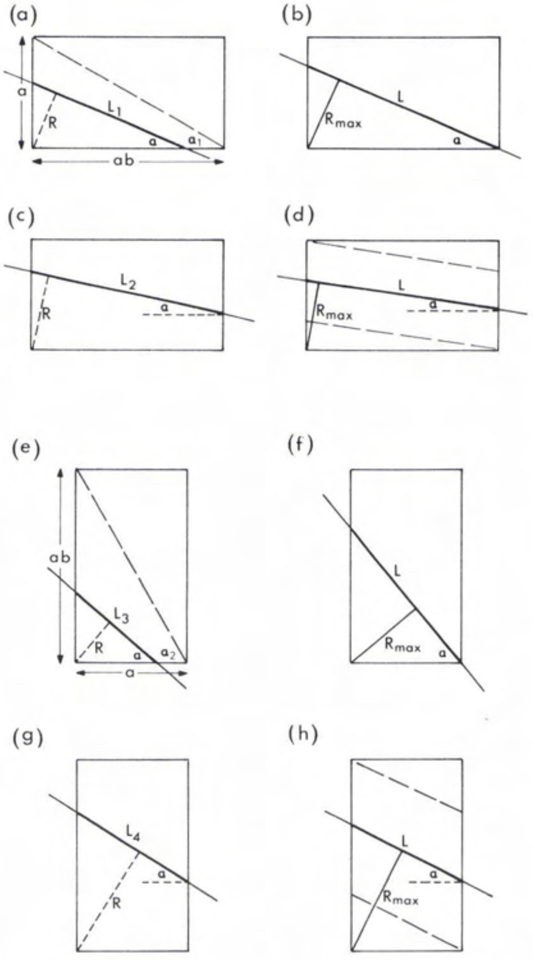

FIG. A-1. A rectangle showing the possible locations of random lines and defining the symbols used.

## REFERENCES

Bauer, M. E., J. E. Cipra, P. E. Anuta, and J. B. Etheridge, 1979. Identification and Area Estimation of Agricultural Crops by Computer Classification of LANDSAT MSS Data. *Remote Sensing of Environment* 8, pp. 77-92.

Bonner, G. M., 1975. The error of area estimates from dot grids. *Canadian Journal of Forest Research*, 5, pp. 10-17.

Crapper, P. F., 1980. The Geometric Properties of Regions with Homogeneous Biophysical Properties. Submitted to Australian Geographical Studies.

Frolov, Y. S., and H. D. Maling, 1969. The accuracy of area measurement by point counting techniques. *The Cartographic Journal*, 6, pp. 21-35.

Goodchild, M. F., and W. S. Moy, 1976. Estimation from grid data: the map as a stochastic process. *Proceedings of the Commission on Geographical Data Sensing and Processing*. Moscow, 1976. pp. 67-81.

Jaynes, E. T., 1971. The well-posed problem, in *Foundations of Statistical Inference*. Edited by V. P. Godancke and D. A. Sprott, Holt-Winston, Toronto. pp. 342-356.

Jupp, D. L., E. M. Adomeit, M. P. Austin, P. Furlonger and K. K. Mayo, 1979. The separability of Land Cover classes on the South Coast of N.S.W. *Proceedings of LANDSAT 79 Conference*. Sydney, May 1979.

Kendall, M. G., and P. A. P. Moran, 1963. *Geometrical Probability*. Griffin, London.

Mandelbrot, B. B., 1977. *Fractals: Form, Chance and Dimension*. W. H. Freeman, San Francisco.

Mathews, M. L., and G. N. Mason, 1979. High Intensity Dot Grids. *Photogrammetric Engineering and Remote Sensing*, 45, pp. 517-518.

Miller, W. F., and B. D. Carter, 1979. Rational land use decision-making: The Natchez State Park. *Remote Sensing of Environment*, 8, pp. 25-38.

(Received 5 November 1979; revised and accepted 11 May 1980)