SAMPLING SYSTEM FOR WHEAT (Triticum aestivum L) AREA ESTIMATION USING DIGITAL LANDSAT MSS DATA AND AERIAL PHOTOGRAPHS

Maurício A. Moreira, Sherry C. Chen, Getulio T. Batista

CNPq - Instituto de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12200 - São José dos Campos - SP - Brasil
Comission VII

ABSTRACT

A procedure to estimate wheat (Triticum aestivum L) area using sampling technique based on aerial photographs and digital LANDSAT MSS data was developed. Aerial photographs covering 720 km² were visually analyzed. Computer classification of LANDSAT MSS data acquired on Sept. 2, 1979 was done using unsupervised and supervised algorithms and classification results were spatially filtered using a post-processing technique. To estimate wheat area, a regression approach was applied using different sample sizes and various sampling units (10, 20, 30, 40 and 60 km²). Based on four decision criteria proposed in this study, it was concluded that: (a) as the size of sampling unit decreased, the percentage of sampled area required to obtain similar estimation performance also decreased; (b) the lowest percentage of the area sampled for wheat estimation with relatively high precision and accuracy through regression estimation was 13.90% using 10 km² as the sampling unit; and (c) wheat area estimation using only aerial photographs was less precise and accurate than those obtained by regression estimation.

1. INTRODUCTION

Brazil is considered in the world scenario as an agricultural country. Nevertheless, the internal consumption of wheat is far greater than the country's production. A considerable amount of wheat has to be imported every year and currently wheat is the second largest (after petroleum) import commodity of Brazil. Considering the great pressure that this commodity is causing to the commercial balance of the country, it becomes evident the need to seek for accurate methods to evaluate wheat production in Brazil in order to provide means for better trade actions.

Any system to estimate crop production requires the estimate of two parameters: yield (kg/ha) and area (ha). Several models have been proposed for wheat yield estimates based on agrometeorological variables, simulation of plant growth, and spectral information of crop conditions. With respect to crop areal estimate, LANDSAT MSS data have been demonstrated to be a very useful tool considering that the data are multispectral, synoptic, repetitive and of global coverage.

The great amount of information provided by land remote sensing systems makes it necessary to use computers in order to extract most efficiently the information required for accurate crop production estimate. In fact, several authors have stated that LANDSAT MSS data analysis through computer-aided techniques has been proved to be effective, fast, and of great potential for crop identification in different regions (Bauer and Cipra, 1973; Economy et al., 1974; Dietrich et al., 1975). In spite of efficiency and potentiality of computer techniques to analyze LANDSAT MSS digital data, in the case of wheat which occupies large regions in Brazil, several variables such as growth stage, field size, soil types, topography, crop

density, scene composition, crop management technique, etc. may affect areal extent estimation accuracy. Significant errors may occur if training statistics obtained in a small area are used to classify a large area due to eventual nonrepresentativeness of the spectral response of the training scene. On the other hand, to get the training statistics from a complete coverage of the entire scene can be very costly and time-consuming.

Wigton and Bormann (1977) mentioned that the use of sampling is efficient for crop area assessment, especially for regions where a complete survey is not economically indicated.

Several studies (Thomas and Hay, 1977; MacDonald and Hall, 1978; Hanuschack et al., 1979; and Graig et al., 1979), utilizing sampling systems and regression method for crop areal estimates have been described.

The objective of this study was to establish a sampling system based on aerial photography and LANDSAT MSS data for estimating the wheat area in a test site of 720 km² in Southern Brazil.

Several criteria are proposed to determine the optimal sizes of sampling unit (segment) and the sample size (number of segments) for the study area. The efficiency of the statistical method utilized was also compared to that obtained from a direct expansion procedure.

2. STUDY AREA AND DATA ACQUISITION

The test site of Cruz Alta (720 km² approximately) was selected in one of the major wheat production areas of the Rio Grande do Sul State. This area represents the technological level of the cropping practices utilized in Southern Brazil (Figure 1). In this region, wheat may be planted in April or May and harvested in October or November, depending on climatic conditions. Figure 2 shows the wheat calendar for the crop year of 1979 with planting occurring mostly in late May. Pasture is also a dominant vegetation class in the test site intercalating with small proportion of natural forest and gallery forest. Natural grassland predominates in the pasture category, and at the time of this study most of the pasture did not show a high amount of green biomass, except those in the depression or humid area.

2.1 - AIRCRAFT DATA

On September 2, 1979, a cloud-free day, color infrared (CIR) aerial photographs (1:20,000) with spectral response from 400 to 900 nm were taken using a photogrammetric camera (23 x 23 cm format). The aerial photographs were visually interpreted and served as reference data for evaluating the results of the sampling procedure studied. Also visual interpretation results of some sampled aerial photographs were used for wheat area estimate in the direct expansion procedure.

2.2 - LANDSAT DATA

Single date LANDSAT-3 MSS digital data acquired on September 4, 1979 were used for this study. The path/row annotation of these data is (220/32).

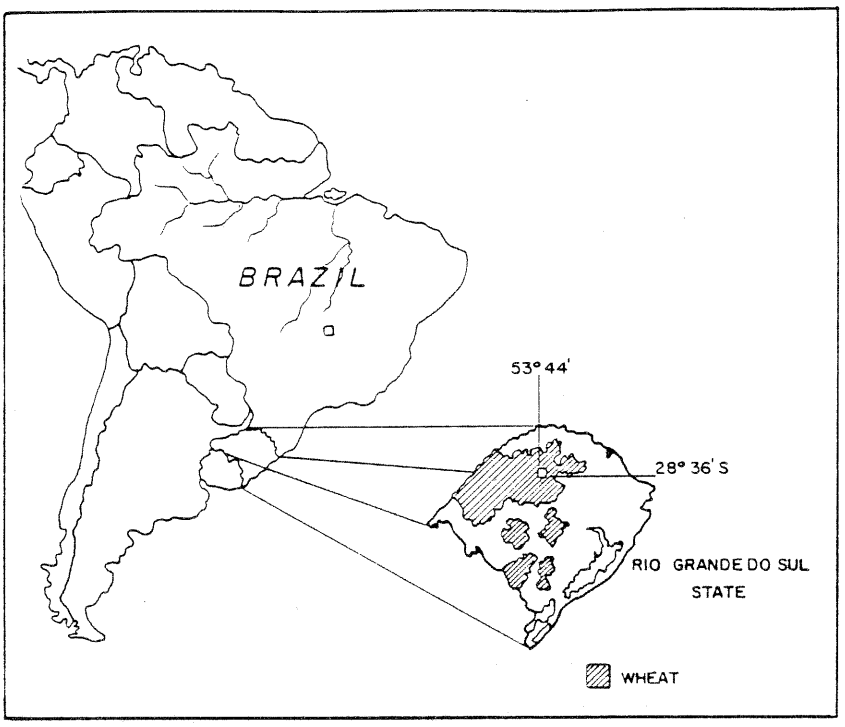

Figure 1 - Location of the study area.

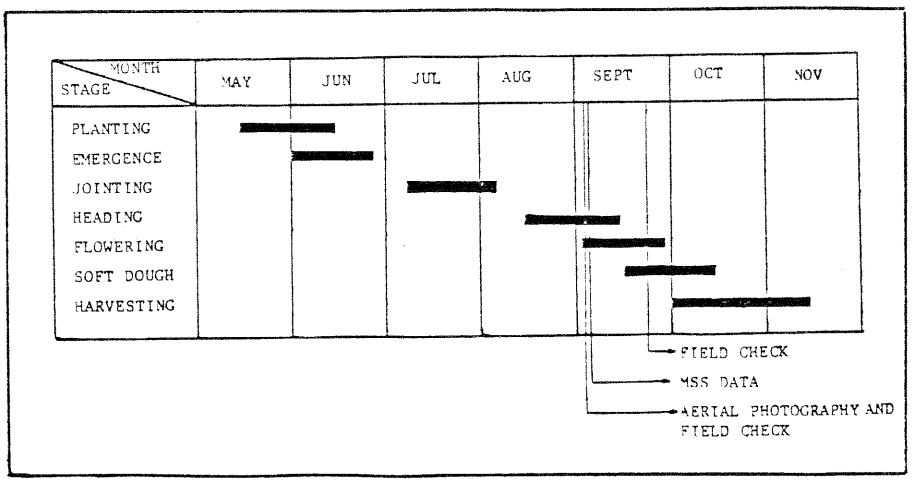

Figure 2 - Crop calendar of wheat for Rio Grande do Sul State in 1979.

3. CLASSIFICATION PROCEDURE

The digital analyses of LANDSAT MSS data were performed by a hybrid (unsupervised and supervised) procedure. Initially, all four spectral bands were classified using a clustering algorithm called K-means (Hartigan, 1975), to separate the spectral information of the test site into several single mode classes. Then pure pixels (located in the center of fields) were selected and the classification statistics (mean vectors and covariance matrices) were acquired for each spectral class which were used later in a maximum likelihood classifier implemented in INPE's (Institute for Space Research) Image-100 System. For improving classification results, a post-classification spatial filtering procedure was used. This procedure consists

of examing sequentially matrices of three-by-three pixels. There are two threshold values to be set by the analyst. The first threshold value is the number of times (weight) that the analyst wants the central pixel to be considered in the calculation of class frequency. After the frequency for all classes has been assessed for all pixels of the three-by-three matrix, the highest class frequency is compared to the second threshold value. If the second threshold value is smaller than the highest class frequency, then the assignment of the class of the central pixel will be substituted by the class which has the highest frequency; otherwise it remains unchanged. Previous study done in spatial filtering by Moreira et al. (1982) indicated that this procedure improved significantly the classification results and that the best setting of the threshold values was (2,2).

4. STATISTICAL ANALYSIS

The determination of a sampling system to estimate wheat area involves the selection of the statistical method, of the best size of the sampling unit (segment) and of the suitable sampling size (number of segments).

4.1 - SELECTION OF THE STATISTICAL METHOD

Initially a simple correlation analysis between wheat area estimations determined by both LANDSAT MSS data and aerial photos, varying the segment size, was performed in order to verify the effect of segment size on the correlation between the two sources of data. Cochran (1965) states that when the relationship between two variables is approximately linear and when the straight line representing this correlation does not pass through the origin of the axes, the regression estimate is more appropriate than the ratio estimate.

The greater the correlation between the variables the lower the variance of the estimation by regression procedure (Hanuschack et al., 1979).

The test site was divided into 72 basic segments of approximately 10 km² each. Besides the 10 km² segment, segments of 20, 30, 40 and 60 km² were also investigated by combining 2, 3, 4 and 6 basic segment units, respectively.

The wheat area of each segment through visual interpretation of aerial photos and digital analysis of LANDSAT MSS data were obtained using the following procedure:

1) An alphanumeric computer printout was obtained from the digital classification of the entire test site. By comparison of this printout with the map obtained by aerial photointerpretation the following areas (hectares) were manually evaluated: (1a) wheat area correctly classified; (1b) area of nonwheat classified as wheat (omission error); and (1c) wheat area not classified as wheat (omission error).

2) For each segment the wheat area estimated by LANDSAT MSS data was equal to the addition of the areas obtained in (1a) plus area obtained in (1b).

3) The area estimated of each segment through photointerpretation was determined by adding area obtained in (1a) and (1c).

For the five tested segment sizes (10, 20, 30, 40 and 60 km²) the correlation coefficients between the wheat area estimates obtained by LANDSAT MSS data and by aerial photos were significant and the regression lines representing these relationships did not pass through the origin of the axes. These results indicated that the regression procedure was suitable for this study and therefore was selected as the statistical method.

4.2 - DETERMINATION OF THE SEGMENT SIZE AND NUMBER OF SEGMENTS REQUIRED

For each segment size (10, 20, 30, 40 and 60 km²) 20 replications of different sampling size were extracted from the population using a simple random procedure without reposition. Figure 3 shows schematically the procedure used.

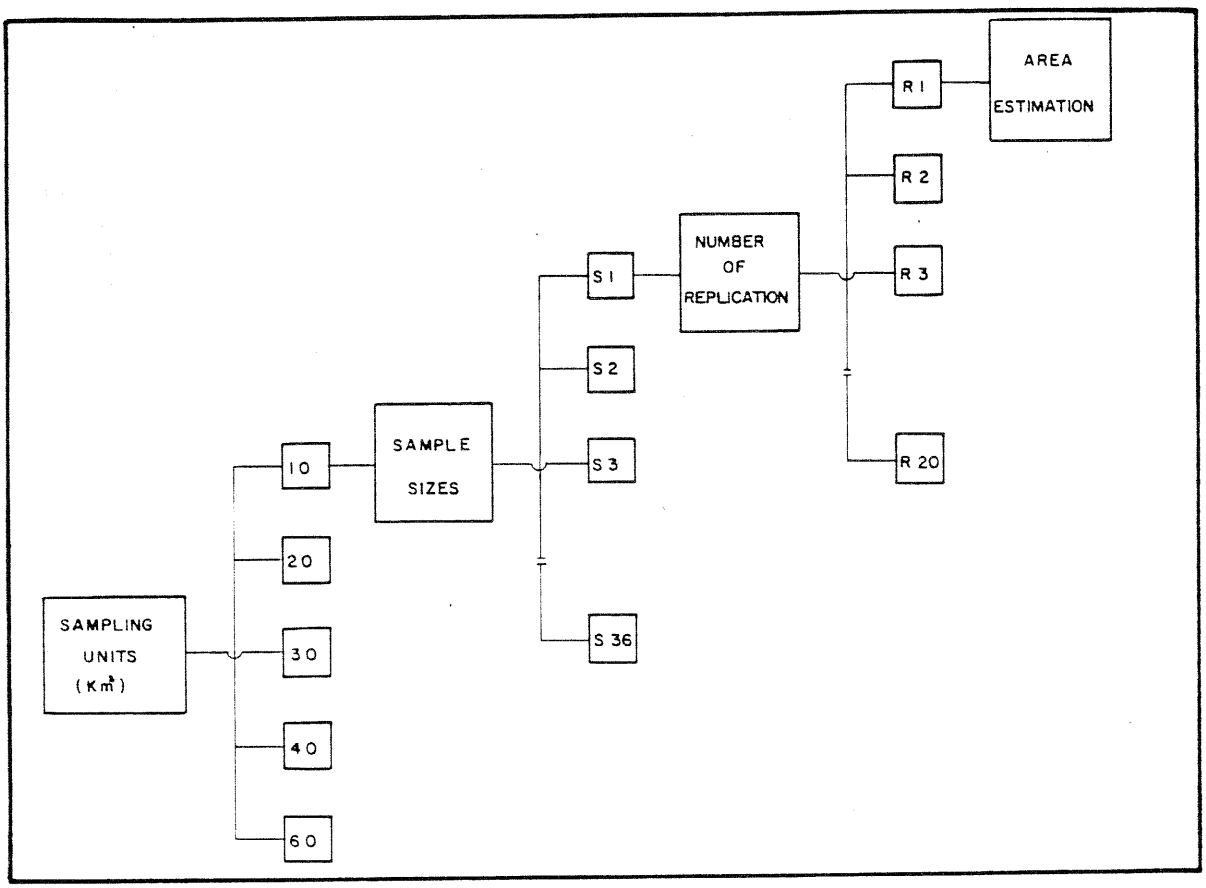

Figure 3 - Sampling procedure.

Four accuracy and precision criteria were established for the selection of the number of segments (i.e. sampling size) required for the regression estimate procedure:

Criterion 1 - The relative difference of the estimated area by regression should not be greater than ± 7.53%, which was the error found by using computer-aided classification of LANDSAT MSS data.

Criterion 2 - The mean of the population of the difference between area estimated through the sampling system and that obtained from aerial photos should not be significantly different from zero ($\alpha = 0.05$).

Criterion 3 - The estimate obtained through the sampling system should not be biased.

Criterion 4 - The coefficient of variation should not be greater than 5%.

Based on criterion 1, the minimum sampling size for each segment was determined. Then, it was examined if the other criteria were satisfied sequentially. Criterion 2 was tested using t-test at $\alpha = 0.05$. Criterion 3 was based on the sign test with $\alpha = 0.05$. Finally the coefficient of variation was determined in order to verify if the sampling procedure (i.e. selected segment and sampling sizes) satisfy criterion 4. The minimum sampling size for each segment had to satisfy simultaneously all four criteria.

4.3 - EFFICIENCY OF THE REGRESSION ESTIMATE

After the segment and sampling sizes have been determined,the relative efficiency (RE) of the estimate utilizing the regression method (both LANDSAT and aerial photos) over the estimate using only aerial photos (direct expansion) was determined by comparing their variances:

$$RE = V(\bar{Y}_{DE})/V(\bar{Y}_{RE}),$$

where $V(\bar{Y}_{DE})$ is the variance using direct expansion and $V(\bar{Y}_{RE})$ is the variance of the regression estimate calculated based on Cochran (1965).

5. RESULTS AND DISCUSSION

The relationships of wheat areas estimated by LANDSAT data and aerial photos for the several segment sizes investigated were presented in Figure 4. All the regression lines are highly significant and have nonzero interceptions.

Figure 5 shows the wheat area estimates obtained by the regression approach for the five segment sizes studied when varying the sampling size. For each sampling size 20 replications were made. Based on Criterion 1 proposed in this study, the sampling sizes which resulted in a relative difference greater than $|7.53\%|$ were excluded. Thus, based on this criterion, at least 10, 8, 8, 6, and 5 segments were required for the segment sizes of 10, 20, 30, 40 and 60 km^2, respectively. These sampling sizes correspond to 13.9%, 22.2%, 33.3%, 33.3%, and 41.7% of the entire test site,respectively.

Results of t-tests show that criterion 2 was satisfied for the sampling sizes selected by the first criterion associated with the segment sizes of 10, 20, 30, and 40 km^2. However, the estimates using 5 segments of 60 km^2 did not meet this criterion. Therefore six sampling units were tested and used as the minimum sampling size for the 60 km^2 segment. As a result, proportion of area sampled increased from 41.7% to 50.0% for the 60 km^2 segment. Sign tests showed that the minimum sampling size for each segment selected by criteria 1 and 2 were not biased and that their CVs were all

smaller than 5% as required by criterion 4. Table 1 summarizes the results obtained for the five sampling unit sizes tested.

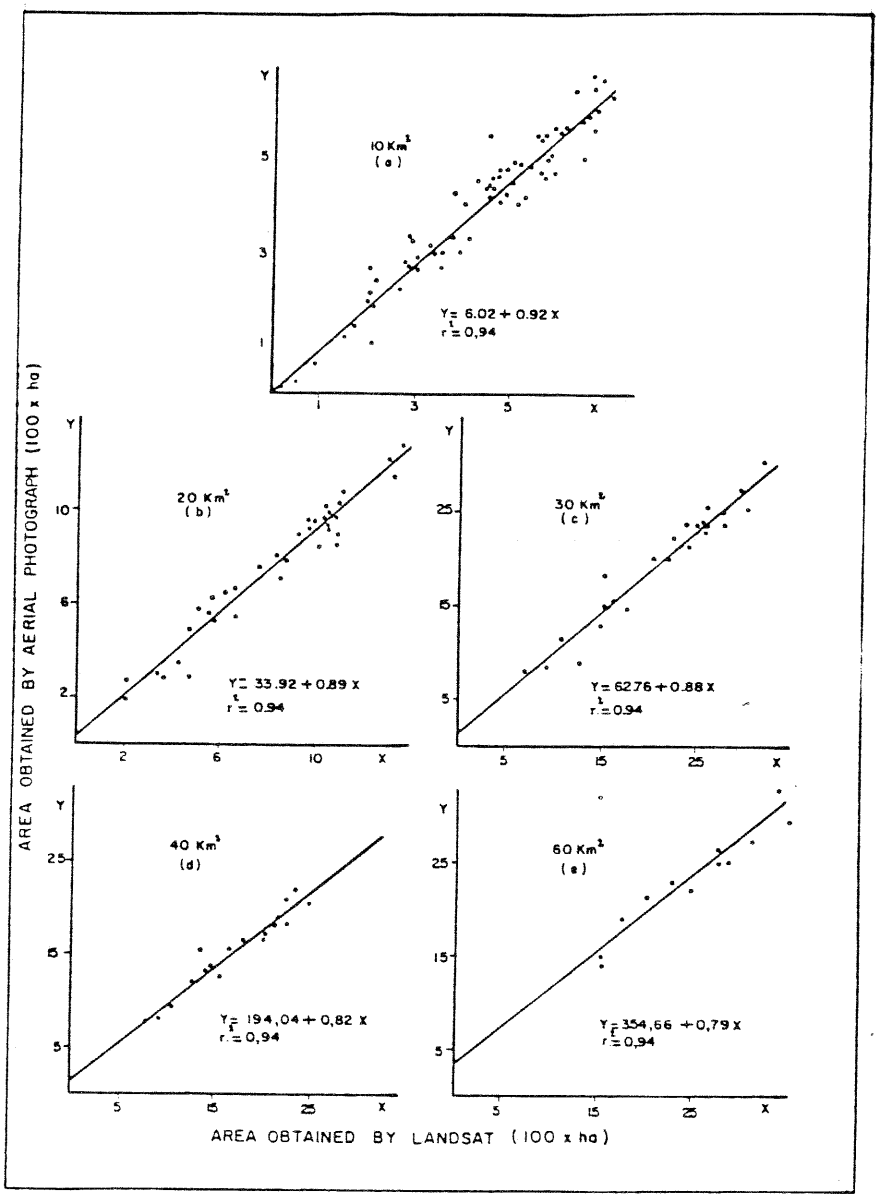

Figure 4 - Relationship of area estimates obtained from LANDSAT data and aerial photographs for different segment sizes (10, 20, 30, 40 and 60 km²).

Generally speaking, all five sampling schemes investigated in this study provided very accurate and unbiased results for wheat area estimates. This can be verified by examining the low relative difference (RD) values or the root mean square errors (RMSE) in Table 1. Also, coefficients of variation for the 20 replications of wheat area estimates for each sampling procedure were very low.

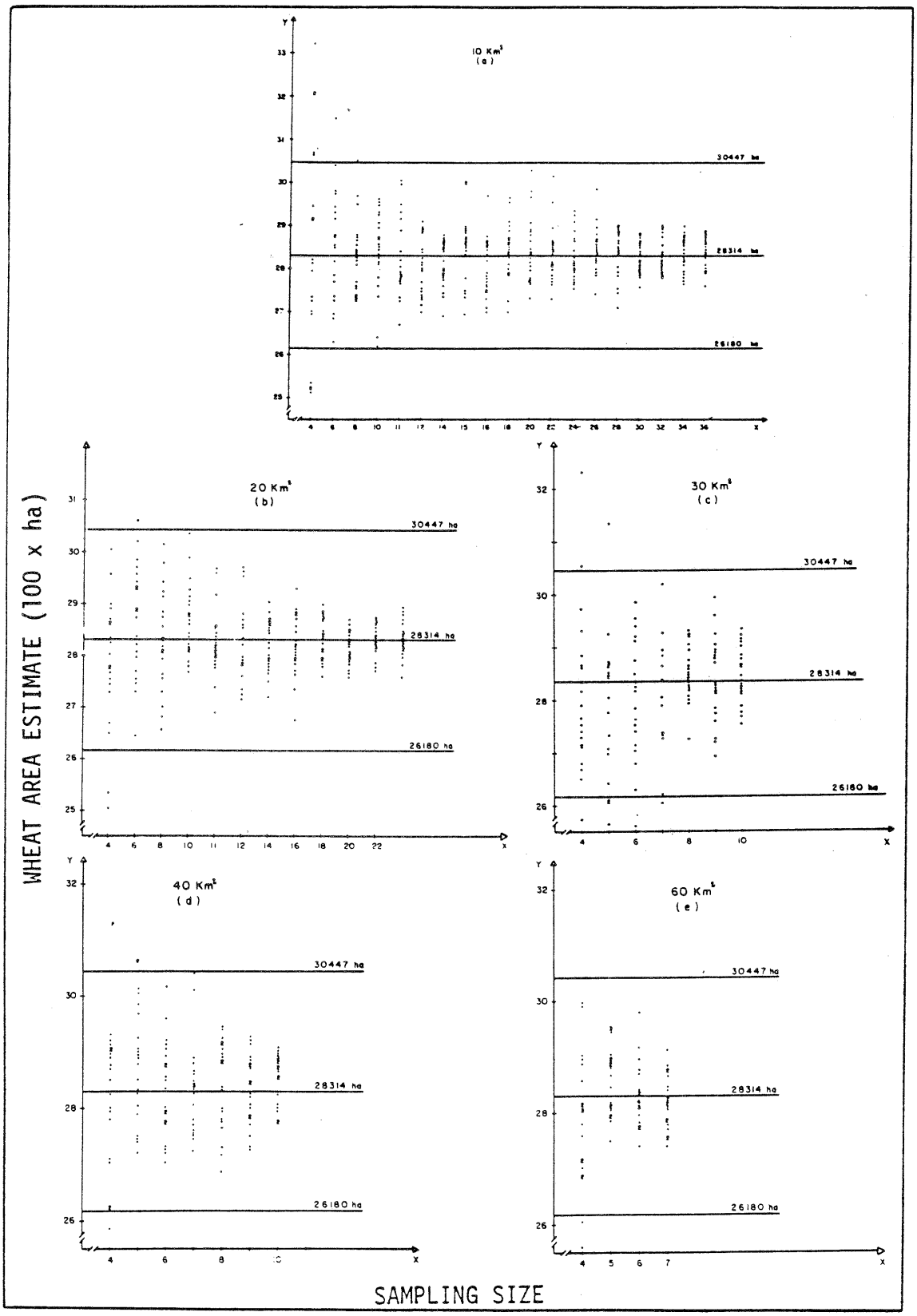

Figure 5 - Wheat area obtained by regression estimation, varying segment sizes (10, 20, 30, 40, and 60 km²) and sampling size.

TABLE 1

SUMMARY OF THE RESULTS OBTAINED FOR THE FIVE SAMPLING SCHEMES TESTED

SAMPLING SYSTEM			MEAN (ha)	STANDARD DEVIATION (ha)	COEFFICIENT OF VARIATION (%)	DIFFERENCE BETWEEN REGRESSION ESTIMATE AND REFERENCE DATA (28,314 ha)		R.M.S.E. (ha)
SEGMENT SIZE (km^2)	SAMPLING SIZE (n)	PERCENTAGE OF THE STUDY AREA SAMPLED (%)				AVERAGE (%)	MAXIMUM (ha)	
10	10	13.90	28,506.10	831.55	2.92	2.30	-2085.51	832.95
20	8	22.22	28,301.78	982.13	3.47	2.70	1882.82	957.34
30	8	33.33	28,502.71	491.17	1.72	1.40	-1045.82	526.20
40	6	33.33	28,371.68	833.33	2.90	2.50	1869.27	834.33
60	6	50.00	28,313.02	533.79	1.88	1.40	1477.36	533.79

The analysis of the results indicated that the most efficient sampling scheme was that of 10 segments of 10 km^2 each because it satisfied the accuracy criteria and required the minimum proportion of the test site to be sampled (13.9%).

The results of the efficiency analysis shown on Table 2 indicated that the regression approach was substantially more efficient than direct expansion method. The relative efficiency values varied from 5.75 to 54.75. In other words, when LANDSAT MSS data were used as an auxiliary data along with aerial photos to estimate wheat area through an appropriate sampling procedure, a gain in precision from 5.75 to 54.75 times was obtained. The relative efficiency is highly associated with the coefficient of determination (r^2) estimated for each sample as shown in Figure 6. The greater the r^2, the smaller the variance by the regression procedure and consequently the greater the value of the relative efficiency.

TABLE 2

COMPARISON BETWEEN THE WHEAT AREA ESTIMATES OBTAINED BY REGRESSION AND DIRECT EXPANSION METHODS FOR 10 SAMPLING UNITS OF 10 KM^2 SEGMENT SIZE

REPLICATES	REGRESSION ESTIMATE		DIRECT EXPANSION		RELATIVE EFFICIENCY (RE)
	$\hat{Y}_R$ (ha)	$V(\hat{Y}_R)$	$\hat{Y}_{DE}$ (ha)	$V(\hat{Y}_{DE})$	
1	28,706.75	943,101.31	28,946.52	12,904,191.00	13.68
2	28,963.22	360,948.87	33,648.19	4,919,018.99	13.63
3	27,808.36	311,541.23	24,407.93	17,059,428.96	54.75
4	29,259.53	437,509.15	33,109.49	9,807,997.74	22.42
5	28,316.01	1,013,768.71	27,779.83	12,874,165.09	12.69
6	27,384.65	800,655.31	30,131.42	12,123,997.82	15.14
7	29,082.70	566,055.18	30,497.47	10,130,649.71	17.89
8	29,483.67	602,561.09	35,513.21	5,246,531.76	8,71
9	28,576.66	433,545.47	26,240.83	10,491,081.01	24.19
10	28,447.07	439,066.36	25,968.82	13,198,144.85	30.05
11	28,180.59	1,010,956.75	32,910.91	16,708,312.54	16.53
12	26,228.49	1,174,274.80	22,045.18	7,850,574.42	6.68
13	29,542.21	1,789,146.87	33,382.01	10,422,605.23	5.82
14	29,287.94	1,560,397.32	28,273.46	8,969,806.35	5.75
15	28,716.81	246,227.12	24,835.82	12,458,182.29	50.59
16	28,129.50	1,163,725.68	31,657.46	11,572,221.63	9.94
17	27,609.15	724,708.85	19,370.74	11,302,107.97	15.59
18	29,610.69	649,077.00	27,626.47	9,951,465.25	15.33
19	28,272.21	1,024,127.30	27,490.19	14,597,391.10	14.25
20	28,515.89	1,044,772.71	30,812.68	23,269,900.02	22.27

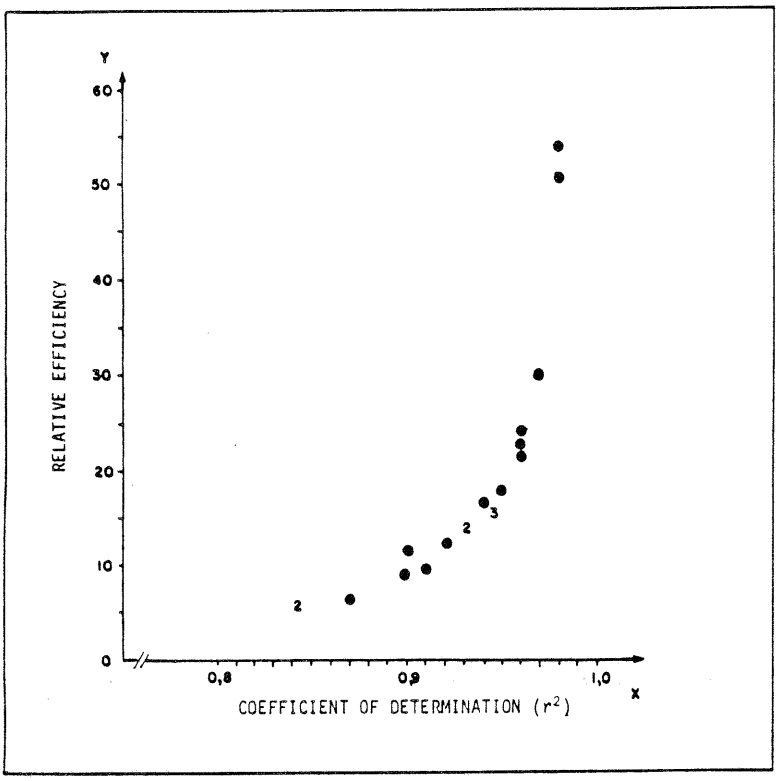

Figure 6 - Relationship of the relative efficiency and the coefficient of determination.

Table 2 also shows that estimates obtained by the regression approach were more accurate and consistent than estimates obtained by direct expansion. It can be observed in this table that the regression estimates varied from 26,228.5 ha to 29,610.7 ha, whereas the direct expansion estimates varied from 19,370.7 ha to 35,513.2 ha for the 20 replications using 10 segments of 10 km².

6. CONCLUSIONS AND RECOMMENDATIONS

Results of this study permit the following conclusion remarks and recommendations:

1) Digital classification of simple date LANDSAT MSS data provided reasonably accurate results for wheat area estimation (i.e. 7.53% overestimate) using a hybrid (unsupervised and supervised) training procedure and a Gaussian maximum likelihood classifier. However the application of computer-aided analysis without any appropriate statistical design can only provide a point estimate; the closeness of this estimate to its corresponding parameter can not be stated.

2) Area estimates from LANDSAT MSS data and aerial photos were highly correlated. Regression lines of these estimates did not intersect the origin (0,0) of the coordinate axes indicating that the regression estimate was appropriate for wheat area estimate in Southern Brazil.

3) As the segment size increases the proportion of the sampled area required to obtain results with the same accuracy and precision increases. The minimum proportion of the study area sampled to

estimate wheat area with high accuracy and precision employing the regression procedure was 13.9%, using 10 segments of 10 km².

4) Wheat area estimated by using sampled aerial photographs alone was less accurate and precise than that obtained by combining the auxiliary information provided by LANDSAT and sampled aerial photographs.

5) For large area with considerable variation in growth stage, field size, soil type, climate, etc. a stratification should be considered before applying the technique proposed in this study.

6) For an operational system, the use of real time aerial photographs could be eventually substituted by old aerial photos updated with field work in order to minimize the cost of obtaining accurate data of crop area in the sampled segments for regression estimate.

REFERENCES

Bauer, M.E.; J.E. Cipra. Identification of agricultural crops by computer processing of ERTS MSS data. Simposium on Significant Results obtained from the Earth Resources Technology Satellite - 1, 3., Washington, DC, 1973. Proceeding. Washington, DC, NASA, 1973, v.1,Sec.A, p. 205-212.

Cochran, W.C. Tecnicas de amostragem. Rio de Janeiro,Fundo de Cultura, 1965.

Dietrich, D.L.; R.F. Fries; D.D. Egbert. Agricultural inventory capabilities of machine processed LANDSAT digital data. In: NASA Earth Resources Survey Symposium; proceeding of a Symposium held in Houston,TX, June 9-12, 1975. Washington, DC, NASA, 1975, v. 1, Sec. A, p. 221-232.

Economy, R.; D.Goodenough; R. Ryerson; R. Towler. Classification accuracy of the IMAGE-100. In: Second Canadian Symposium on Remote Sensing, Guelph, ON, 1974. Proceedings. Ottawa, ON, Canadian Remote Sensing Society, v. 2, p. 277-287.

Graig, M.E.; R.S. Sigman; M. Cardenas. Area estimates by LANDSAT: Kansas 1976. Winter Wheat. In: 13th International Symposium on Remote Sensing of Environment, Ann Arbor, MI, 1979. Proceedings, Ann Arbor, ERIM, 1979, v. 3, p. 1727-1736.

Hanuschack, G.A.; R. Sigman; M.E. Graig; M. Ozgh; R.G. Luebbe; P.W. Cook; D.D. Klewend; C.E. Miller. Crop-area estimates from LANDSAT; transition from research and development to timely results. In: Fifth Annual Symposium on Machine Processing of Remotely Sensed Data, West Lafayette, IN, 1979. Proceedings. W. Lafayette, IN, Purdue Univ., 1979, p.86-96.

Hartigan, J.A. Clustering algorithms. John Wiley & Sons, New York, 1975.

MacDonald, R.B.; F.G. Hall LACIE: An experiment in global crop forecasting The LACIE Symposium, Houston, TX, Oct. 23-26, 1978. 32 p.

Moreira, M.A.; S.C. Chen; A.M. de Lima. Estudo do Método Uniformização de Temas (UNITOT) e análise da Correlação entre áreas estimadas utilizando dados do LANDSAT e fotografias aéreas. II Simpósio de Sensoriamento Remoto, Brasília, 10-14 maio, 1982.

Thomas, R.W.; C.M. Hay. Two phase sampling for wheat acreage estimation. In: Fourth Annual Symposium on Machine Processing of Remotely Sensed Data, West Lafayette, IN, 1977. Proceedings. West Lafayette, IN, Purdue University, 1977, p. 91-100.

Wigton, W.; P. Bormann. A guide to area sampling frame construction utilizing satellite imagery. In: Second International Training Course in Remote Sensing Applications for Agricultural; Crop Statistics and Agricultural Census, 2., Rome, 25 April - 13 May, 1977.