General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

Produced by the NASA Center for Aerospace Information (CASI)

E82-10385

7. IT-E2-04246

5. NAS9-15476

NASA-CR-167622

AgRISTARS

A Joint Program for Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing

6. January 1982

Inventory Technology Development

Technical Report

"Made available under NASA sponsorship in the interest of early and wide dissemination of Earth Resources Survey Program information and without liability for any use made thereof."

ESTIMATING ACREAGE BY DOUBLE SAMPLING USING LANDSAT DATA

3. F. Pont, H. Horwitz and R. Kauth

(E82-10385) ESTIMATING ACREAGE BY DOUBLE SAMPLING USING LANDSAT DATA Technical Report, 15 Nov. 1980 - 14 Nov. 1981 (Environmental Research Inst. of Michigan) 36 p HC A03/MF A01 CSCL 02C G3/43 N82-32807 Unclas 00385

4. ERIM
ENVIRONMENTAL RESEARCH INSTITUTE OF MICHIGAN
ANN ARBOR, MICHIGAN

UCB
SPACE SCIENCES LABORATORY
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA

PRECEDING PAGE BLANK NOT FILMED

IT-E2-04246
NAS9-15476

TECHNICAL REPORT

ESTIMATING ACREAGE BY DOUBLE SAMPLING USING LANDSAT DATA

by

F. Pont, H. Horwitz and R. Kauth

Environmental Research Institute of Michigan
P.O. Box 8618
Ann Arbor, Michigan 48107

January 1982

PRECEDING PAGE BLANK NOT FILMED

PREFACE

The Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing Program, AgRISTARS, is a six-year program of research, development, evaluation, and application of aerospace remote sensing for agricultural resources, which began in Fiscal Year 1980. This program is a cooperative effort of the National Aeronautics and Space Administration, the U.S. Departments of Agriculture, Commerce, and the Interior, and the U.S. Agency for International Development. AgRISTARS consists of eight individual projects.

The work reported herein was sponsored by the Inventory Technology Development (ITD) Project under the auspices of the National Aeronautics and Space Administration, NASA. Dr. Jon D. Erickson, NASA Johnson Space Center, was the NASA Manager of the ITD Project and Mr. Lewis C. Wade was the Technical Coordinator for the reported effort.

The Environmental Research Institute of Michigan and the Space Sciences Laboratory of the University of California at Berkeley comprised a consortium having responsibility for development of corn/soybeans area estimation procedures applicable to South America within both the Supporting Research and Inventory Technology Development Projects of AgRISTARS.

This reported research was performed within the Environmental Research Institute of Michigan's Infrared and Optics Division, headed by Richard R. Legault, a Vice-President of ERIM, under the technical direction of Robert Horvath, Program Manager, and Richard C. Cicone, Task Leader.

TABLE OF CONTENTS

Section		Page
1	INTRODUCTION	1
2	DOUBLE SAMPLING OPTIMIZATION	3
3	EXAMPLE APPLICATIONS	9
4	SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS	29
	4.1 THE PERFECT PROCEDURE COMBINED WITH EXISTING LANDSAT PROCEDURES	29
	4.2 EXISTING LANDSAT PROCEDURES COMBINED	29
	4.3 GENERAL	30
	REFERENCES	31
	DISTRIBUTION LIST	33

PRECEDING PAGE BLANK NOT FILMED

LIST OF FIGURES

Figure		Page
1	GEOMETRY FOR THE CASE r = 2 AND s = 10	7
2	SCATTERPLOT OF STAGE 2 CORN VS. STAGE 1 CORN ESTIMATE	14
3	SCATTERPLOT OF STAGE 2 SOYBEANS VS. STAGE 1 SOYBEANS ESTIMATE	15
4	GEOMETRY FOR STAGE 2 WITH STAGE 1 (400 hrs analyst time, 35 computer hours)	17
5	GEOMETRY FOR STAGE 2 WITH STAGE 1 (320 hrs analyst time, 30 computer hours)	18
6	SCATTERPLOT OF GROUND TRUTH CORN VS. STAGE 2 CORN ESTIMATE	20
7	SCATTERPLOT OF GROUND TRUTH SOYBEANS VS. STAGE 2 SOYBEANS ESTIMATE	21
8	SCATTERPLOT OF GROUND TRUTH CORN VS. STAGE 1 CORN ESTIMATE	23
9	SCATTERPLOT OF GROUND TRUTH SOYBEANS VS. STAGE 1 SOYBEANS ESTIMATE	24
10	RELATIVE CORN VARIANCE AS A FUNCTION OF COST OF G.T. ESTIMATES	26
11	RELATIVE SOYBEAN VARIANCE AS A FUNCTION OF COST OF G.T. ESTIMATES	27

PRECEDING PAGE BLANK NOT FILMED

LIST OF TABLES

Table		Page
1	ANALYST HOURS AND COMPUTER HOURS FOR STAGE 1 AND STAGE 2	12

PRECEDING PAGE BLANK NOT FILMED

1
INTRODUCTION

In crop inventory applications, as in many forms of survey sampling, there may be two, nominally competing, techniques of measurement available, each with its associated variance, bias and cost per sample. If it is necessary to choose one or the other technique how should the choice be made? If the techniques both have an acceptably small bias the answer is well known [1]; choose the technique with smaller cost-variance product.

More often it is not necessary to choose strictly among measurement techniques; rather it is possible to make some of both kinds of measurements and mix the results to obtain an overall lower variance at the same total cost, even when one of the techniques has an unacceptable bias. Consider a low cost, biased, high variance technique and a high cost, (nearly) unbiased low variance technique whose results on the same samples are well correlated. We can view the high cost technique as a method of calibration of the low cost technique. The calibration is performed by double sampling wherein the bulk of the samples will be measured inexpensively, and a certain subset of samples are measured by both techniques. The entire set of measurements is then used to make a regression estimate which is unbiased with respect to the more expensive measurement technique and lower variance (than either technique used separately) for a given total cost. The conditions for which this is true are again given by Cochran [1]. The answer (the number of double and single samples allocated) is obtained by minimizing the variance of the estimator subject to a fixed total cost. Such situations are most likely to arise in practice if the competing techniques in question

[1] Cochran, page 341, formulas 12.64 and 12.65.

share some substantial portion of their overhead costs in common, e.g., if the more expensive technique is a more extensive or thorough application of the lower cost technique.

The USDA's Domestic Crop/Land Cover Project [5] utilizes double sampling techniques to adjust a Landsat-based estimate over a large region by the use of an estimated regression relationship between the Landsat-based and ground survey-based estimates over a subset of the region.

The application discussed in this report centers around several Landsat-based techniques for estimating crop acreages, namely: a fictional perfect procedure, a relatively expensive analyst-intensive use of Landsat data, and a less expensive but closely related method of using Landsat data. However, the application studied in this report is of more general interest than described above in two significant ways:

a) The quantity to be estimated is multivariate, i.e., the acreages of two or more crops (in particular, corn and soybeans) simultaneously,

b) The cost constraints are more general, consisting of limitations on two or more types of resources (analysts and computers) as well as total cost.

In this more general situation one must define a suitable objective function to minimize (replacing the variance) subject to the (more elaborate) constraint set.

In Section 2 of this report, the double sampling optimization problem with multiple constraints and vector valued estimates is considered and a solution algorithm described. In Section 3, the several competing techniques are discussed, a scenario for joint use of the techniques is described, and a data base consisting of joint samples of perfect estimates (simulated using ground truth) and the actual estimates of the existing techniques is used to apply the analysis to three cases of double sampling. Section 4 contains the summary and conclusions of the study.

2

DOUBLE SAMPLING OPTIMIZATION

In double sampling with a regression estimate the variance, V, of the estimator, $\hat{y}$, is approximately (see [1] page 343):

$$V(n,n') = \frac{S_y^2(1-\rho^2)}{n} + \frac{\rho^2 S_y^2}{n'} - \frac{S_y^2}{N}$$

where

S_y^2	Variance of the population
ρ	Correlation coefficient between primary and auxiliary variables
n'	Size of auxiliary sample (i.e., the less expensive measurement)
n	Size of primary sample
N	Size of population

This formula assumes that $\frac{1}{n}$ is negligible. Suppose the cost of a primary observation is c and an auxiliary observation is c', then the total cost C is

$$C = cn + c'n'.$$

Cochran [1] shows under what conditions double sampling is superior to a single sample with no regression adjustment.

In our application we have more than one constraint. For example, total computer time for processing is limited. Further, AI (Analyst-Interpreter) time is also bounded. These constraints are linear and may be described by

$$A \begin{bmatrix} n \\ n' \end{bmatrix} \leqq b$$

with

$$A = [a_{ij}] \quad \text{an } r \times 2 \text{ matrix}$$

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

and r the number of these constraints. We note here that the entries of A and b are positive.

We have one further constraint. In order to get an adequate estimate of ρ, we require that

$$n' \geqq s$$

Because of the nature of double sampling, we have $n \leq n'$.

If we set

$$k = \frac{S_y^2(1-\rho^2)}{n}$$

$$k' = \rho^2 S_y^2$$

Then the variance $V(n,n')$ is

$$V(n,n') = \frac{k}{n} + \frac{k'}{n'} - \frac{S_y^2}{N}$$

and finding the optimal (n,n') for V is equivalent to finding the optimal (n,n') for $f(n,n')$ where

$$f(n,n') = V(n,n') + \frac{S_y^2}{N} = \frac{k}{n} + \frac{k'}{n'}$$

ORIGINAL PAGE IS
OF POOR QUALITY

Thus the problem of finding the optimal (n,n') may be formulated as

$$\min_{n,n'} f(n,n')$$

subject to the constraints:

$$A \begin{bmatrix} n \\ n' \end{bmatrix} \leqq b$$

$$n' \geqq s$$

$$n \leqq n'$$

$$n,n' \text{ positive integers}$$

We now give a procedure for finding the optimal pair (n,n'). Let $n_o = s$ and find the largest value of n' for which (n_o,n') is feasible, call it n'_o.

$$\nu_i(o) = \frac{b_i - n_o a_{i1}}{a_{i2}} \qquad 1 \leqq i \leqq r$$

$$n'_o = \left[\min_i \nu_i(o)\right]$$

where [x] denotes the greatest integer in x. There are no feasible solutions if $n'_o < n_o$. If $n'_o \geq n_o$, set $n_1 = n_o + 1$ and find the largest value of n' for which (n_1,n') is feasible, call it n'_1.

$$\nu_i(1) = \frac{b_i - n_1 a_{i1}}{a_{i2}}$$

$$n'_1 = \left[\min_i \nu_i(1)\right]$$

If $n'_1 \geq n_1$, set $n_2 = n_1 + 1$ and find the largest value of n' for which (n_2,n') is feasible, call it n'_2.

ORIGINAL PAGE IS
OF POOR QUALITY

$$\nu_1(2) = \frac{b_i - n_2 a_{i1}}{a_{i2}}$$

$$n_2' = \left[\min_i \nu_i(2)\right]$$

and so on. Because of the nature of the set of feasible solutions, we will reach a stage J such that (n_J, n_J') is feasible, but there are no feasible solutions with $n > n_J$.

Then the optimum (n,n') is given by

$$(n,n') = (n_{j_o}, n_{j_o}') \qquad 0 \leq j_o \leq J$$

for which

$$f(n_{j_o}, n_{j_o}') \leq f(n_j, n_j') \qquad 0 \leq j \leq J$$

Figure 1 illustrates the geometry for a case r=2 and s=10. The set of feasible solutions contains all the integer points in the quadrilateral $Q_1Q_2Q_3Q_4$. The points designated by crosses are the (n_j, n_j').

If we have double sampling with estimates, say $\hat{y}_1$ and $\hat{y}_2$, of two different crops, but with n and n' the same, we would like to optimize the pair (n,n') for some function of the variances $V_1(\hat{y}_1)$ and $V_2(\hat{y}_2)$. Two such functions are

$$V = V_1 + V_2$$

and

$$V = \max V_1, V_2$$

Either one of these objective functions is a monotone decreasing function of each of the variables n and n', so that the algorithm described will work with the obvious modification in computing the values of the objective function at the points

$$(n_j, n_j').$$

ORIGINAL PAGE IS
OF POOR QUALITY

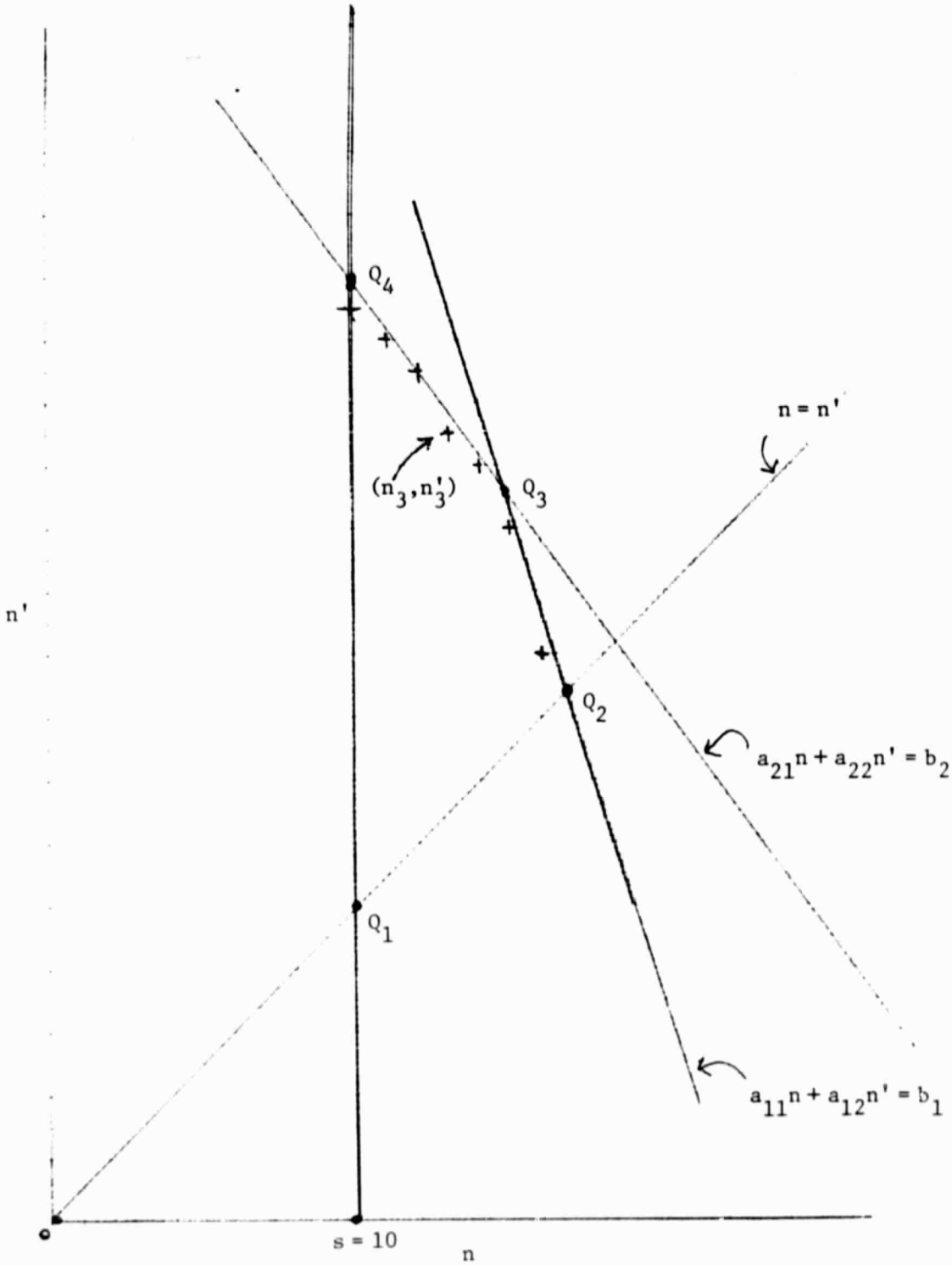

FIGURE 1. GEOMETRY FOR THE CASE $r = 2$ AND $s = 10$

PRECEDING PAGE BLANK NOT FILMED

3
EXAMPLE APPLICATIONS

In this section we describe the examples of double sampling which were the subject of this study. These examples have a common context or scenario which will be described first. The specific measurement techniques will then be outlined and the data base available will be discussed. Finally the examples of double sampling will be given with their results.

Scenario

In the AgRISTARS project, under the corn and soybeans subproject, a baseline method of processing Landsat multitemporal data for acreage estimation of corn and soybeans has been developed [2]. This C/S Baseline is applied currently to LACIE size sample segments (117 lines x 196 pixels) drawn from corn and soybeans growing regions.

In AgRISTARS, for research purposes, ground truth is acquired on a large number of sample segments. This allows us to simulate a perfect Landsat-based technique. Briefly, the three techniques of measurement we are dealing with are

1) Perfect Estimate (most expense)
2) C/S Baseline Stage 2 Estimate (intermediate expense)
3) Abbreviated C/S Baseline Stage 1 Estimate (least expense)

Within the context of the scenario which follows three cases were studied: Stage 1 used jointly with Stage 2; Stage 1 used jointly with ground truth; and Stage 2 used jointly with ground truth.

These cases were examined within a specific set of assumptions and constraints (i.e., a scenario) which were chosen to be representative of a possible future operational system environment and consistent with a recently conducted "shakedown" exercise of the C/S Baseline procedure.

Page 8 blank

1) An estimation system manager has been given two weeks (i.e., ten 8-hour working days) to obtain an estimate.

2) The system has at its disposal five analysts, i.e., a maximum of 400 hours.

3) The system has at its disposal a maximum of 35 hours of computer time.

4) The costs of resources for processing are as given in Table 1.

5) The data for a sufficient number of segments (perhaps in full frame format) is already available and is not counted in the cost analysis.

The estimation system manager now has the problem of minimizing the variance of his estimates within his total resource constraints.

Measurement Techniques

We now discuss briefly the nature of the Stage 1 and Stage 2 estimates. The C/S Baseline procedure first applies screening, preprocessing and spectral feature extraction procedures to each pass over each segment to remove the effects of noise, solar zenith angle and varying haze over the scenes and to compress the spectral information into two feature channels ("Brightness" and "Greenness") [3]. An agricultural resource analyst then examines the data to establish a crop calendar for the segment and an automatic procedure is run to place each pixel into a temporal pattern class (TPC), by virtue of the times when that pixel's Greenness value is above a bare soil reference line. Temporal pattern classes are mapped into one of 5 crop groups: pasture, spring crop, summer crop, non-agriculture and unknown. Within the summer crop group the analyst then establishes a discriminant line in Brightness-Greenness space at one particular critical acquisition. Pixels above this line are labeled soybeans, pixels below this line are labeled corn. At this point all

pixels have been given a tentative (Stage 1) label from one of six classes and can be aggregated into a corn proportion estimate and a soybeans proportion estimate for the segment. This constitutes the abbreviated C/S Baseline measurement technique, i.e., the Stage 1 estimate.

The next step in the C/S Baseline procedure is to apply a multitemporal, multispectral spatial processing procedure which builds field-like structures (pseudo-fields) in the data by grouping together pixels which are alike both spectrally and spatially. The spatial mean (x,y coordinate value) of each pseudo-field identifies its location; the temporal-spectral mean is regarded as a feature describing the pseudo-field and constitutes a further feature compression. The pseudo-fields fall into two broad categories, those which have one or more interior pixels and those which do not. Those which do not are identified as "small" and are deemed not useful for analyst labeling.

Next, all pseudo-fields are grouped into as many as 40 spectral strata through an unsupervised clustering process, constrained by the condition that all spectral means also fall into the same crop groups and subgroups previously defined. All pixels within each pseudo-field are labeled into the same crop group or subgroup as the pseudo-field it belongs to.

To make a refined estimate of proportion, a sample of 100 pseudo-fields is drawn from the strata (using the Midzuno sampling technique) [4] and labeled by the resource analyst. Typically the analyst may relabel 8 to 12 out of the 100 he examines. The labels now form the basis for a stratified area estimate of the proportion of the segment that is corn and the proportion that is soybeans. This estimate is the Stage 2 estimate and is the final output of the C/S Baseline Procedure.

The Stage 2 estimate represents a considerable increment in analyst and computer time over the Stage 1 estimate, as shown in Table 1. This table shows that the incremental cost of a Stage 2 estimate is almost four times the cost of a Stage 1 estimate.

TABLE 1. ANALYST HOURS AND COMPUTER HOURS FOR STAGE 1 AND STAGE 2. The Stage 2 hours shown are the increment above Stage 1.

	Analyst Hours	Computer Hours
Stage 1 Estimate	2	.25
Stage 2 Estimate	8	.5

Data Base and Examples

The data base available for this study was a set of 39 multi-pass Landsat segments from Iowa, Illinois and Indiana which had good acquisition histories during the 1978-79 growing season, and for which ground truth was acquired. The C/S Baseline was applied to these segments and both a Stage 1 and a Stage 2 estimate was obtained for each. From the array of ground truth and Stage 1 and Stage 2 estimates all of the inputs called for by the analysis of Section 2 above were estimated and results (i.e., optimum sample sizes) were obtained for three cases: Stage 1 used jointly with Stage 2, Stage 1 used jointly with Ground Truth and Stage 2 used jointly with Ground Truth.

Results

Stage 2 With Stage 1

Denote the Stage 2 corn and soybeans estimates as y_c and y_s, and the Stage 1 corn and soybeans estimates as x_c and x_s. The sample correlation matrix of

$$\begin{pmatrix} y_c \\ x_c \\ y_s \\ x_s \end{pmatrix}$$

was

$$\begin{bmatrix} 1.00 & .79 & .34 & .26 \\ - & 1.00 & .15 & .11 \\ - & - & 1.00 & .90 \\ - & - & - & 1.00 \end{bmatrix}.$$

Figures 2 and 3 give scatterplots of Stage 2 vs. Stage 1 estimates for corn and soybeans, respectively. In this case and the subsequent case the multiple R was not significantly larger than the simple correlations so only simple regression was used.

ORIGINAL PAGE IS
OF POOR QUALITY

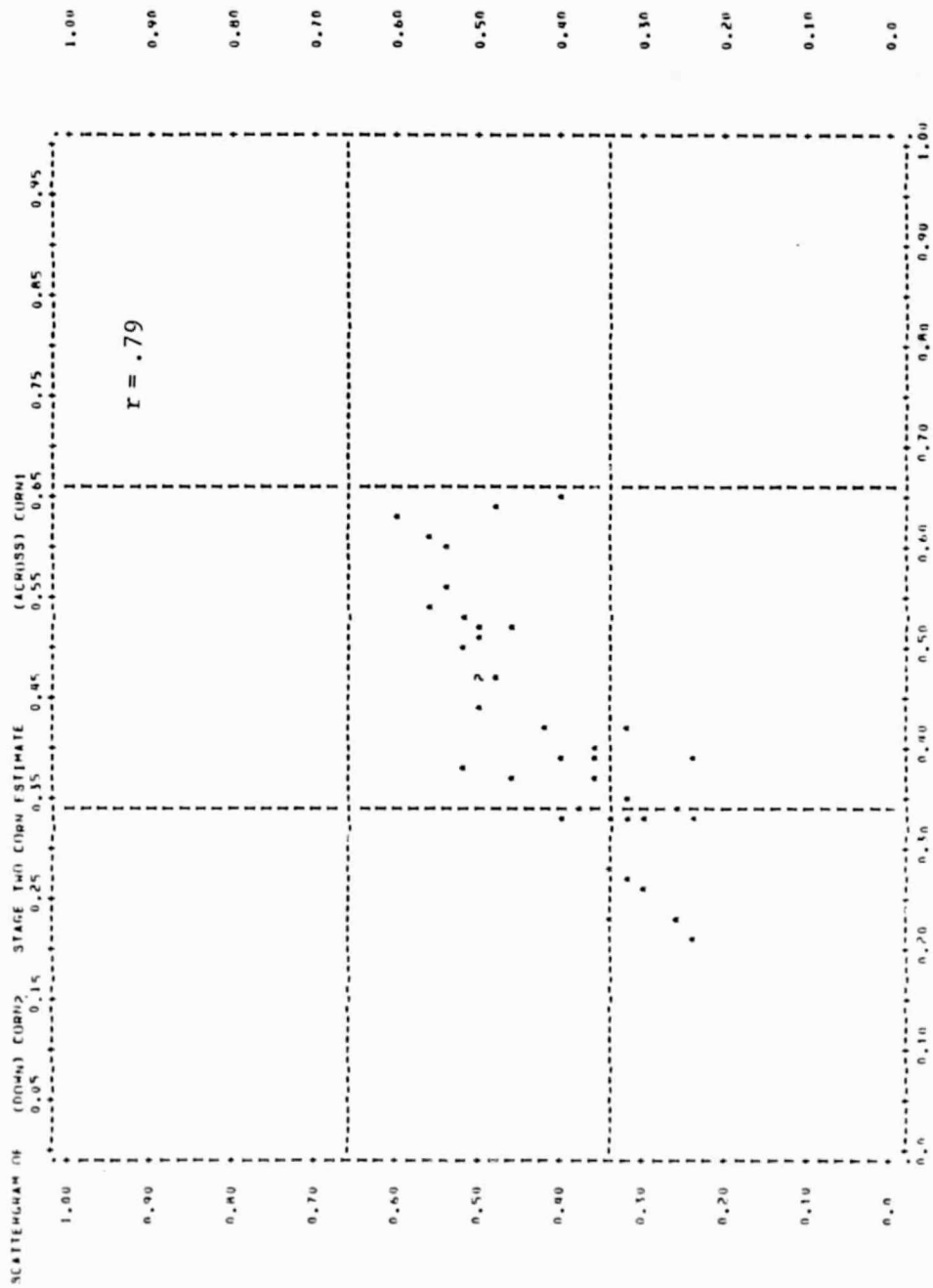

FIGURE 2. SCATTERPLOT OF STAGE 2 CORN VS. STAGE 1 CORN ESTIMATE

ORIGINAL PAGE IS
OF POOR QUALITY

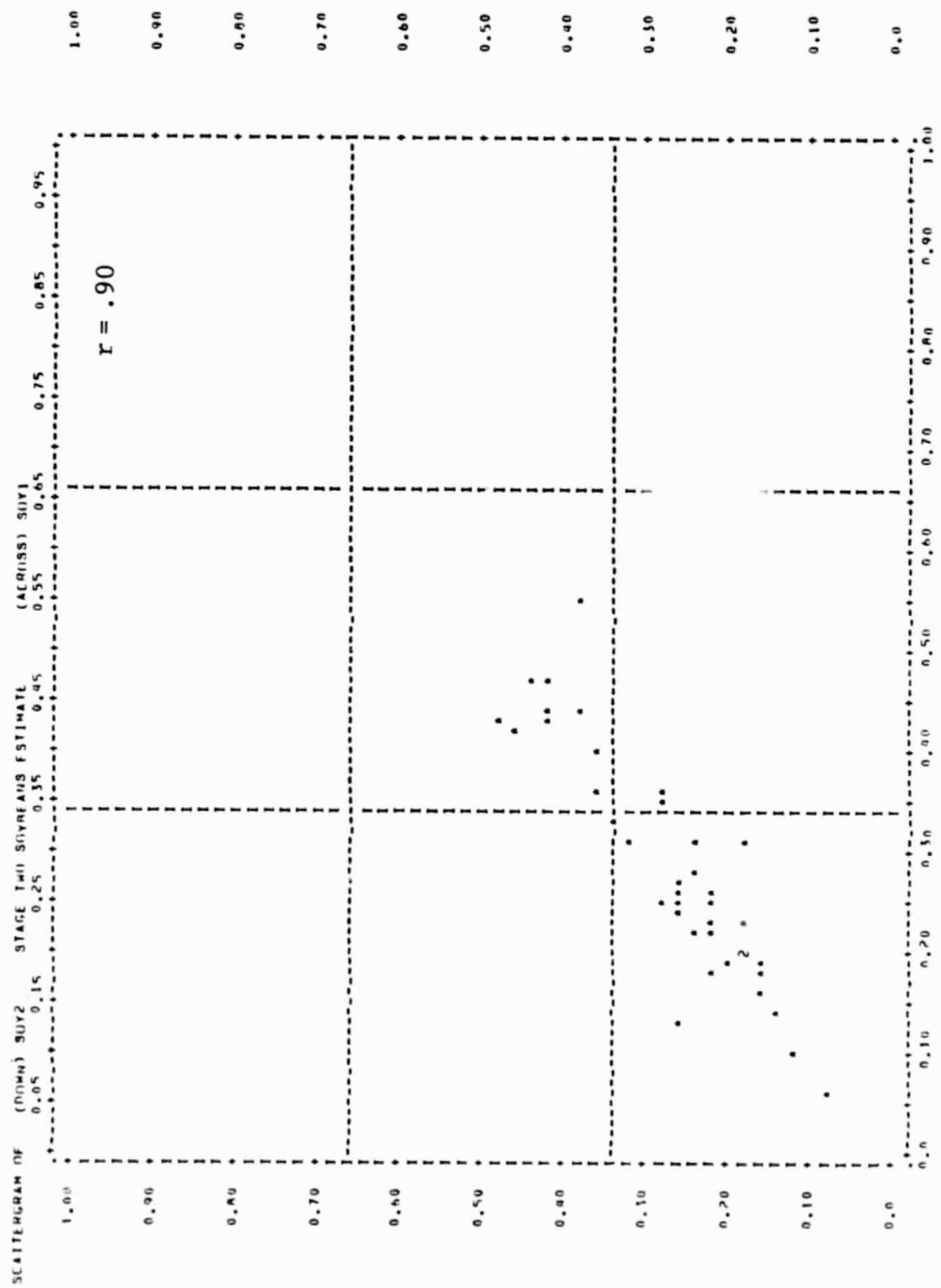

FIGURE 3. SCATTERPLOT OF STAGE 2 SOYBEANS VS. STAGE 1 SOYBEANS ESTIMATE

ORIGINAL PAGE IS
OF POOR QUALITY

In this example we assume that we have the analyst and computer constraints:

$$(1) \quad \begin{bmatrix} 2 & 8 \\ \frac{1}{4} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} n' \\ n \end{bmatrix} \leq \begin{bmatrix} 400 \\ 35 \end{bmatrix},$$

(2) $n' \geq n$, and

(3) $n \geq 10$

The first and third constraints were explained in the last section. The second constraint reflects the fact that a Stage 1 estimate is obtained automatically for every Stage 2 estimate. If we were not constrainted by (2) then ($n = 50$, $n' = 0$) would be the optimal allocation for corn. The constraints and feasible points are given in graph form in Figure 4.

The constraints were chosen so that the recently conducted baseline corn and soybeans procedure would be feasible. The baseline procedure currently replaces constraint (2) with (2') $n = n'$. The optimal allocation in this case is ($n = 40$, $n' = 40$). If (2') is replaced with (2) then point A in Figure 4 ($n = 30$, $n' = 80$) minimizes S_c^2, $S_c^2 + S_s^2$, and $\max(S_c^2, S_s^2)$. The precision relative to the baseline procedure is for S_c^2, $S_c^2 + S_s^2$ and $\max(S_c^2, S_s^2)$, 1.24, 1.38, and 1.39, respectively. Point B in Figure 4 ($n = 28$, $n' = 84$) minimizes S_s^2 with relative precision of 1.54. Since most users would be interested in both corn and soybeans error the overall optimal allocation would be 80 segments with Stage 1 estimate and a random sample of size 30 from the 80 should be selected for the segments with Stage 2 estimates. Figure 5 gives the constraint space if there are only 320 analyst and 30 computer hours which can be used to make a corn,soybeans estimate.

ORIGINAL PAGE IS
OF POOR QUALITY

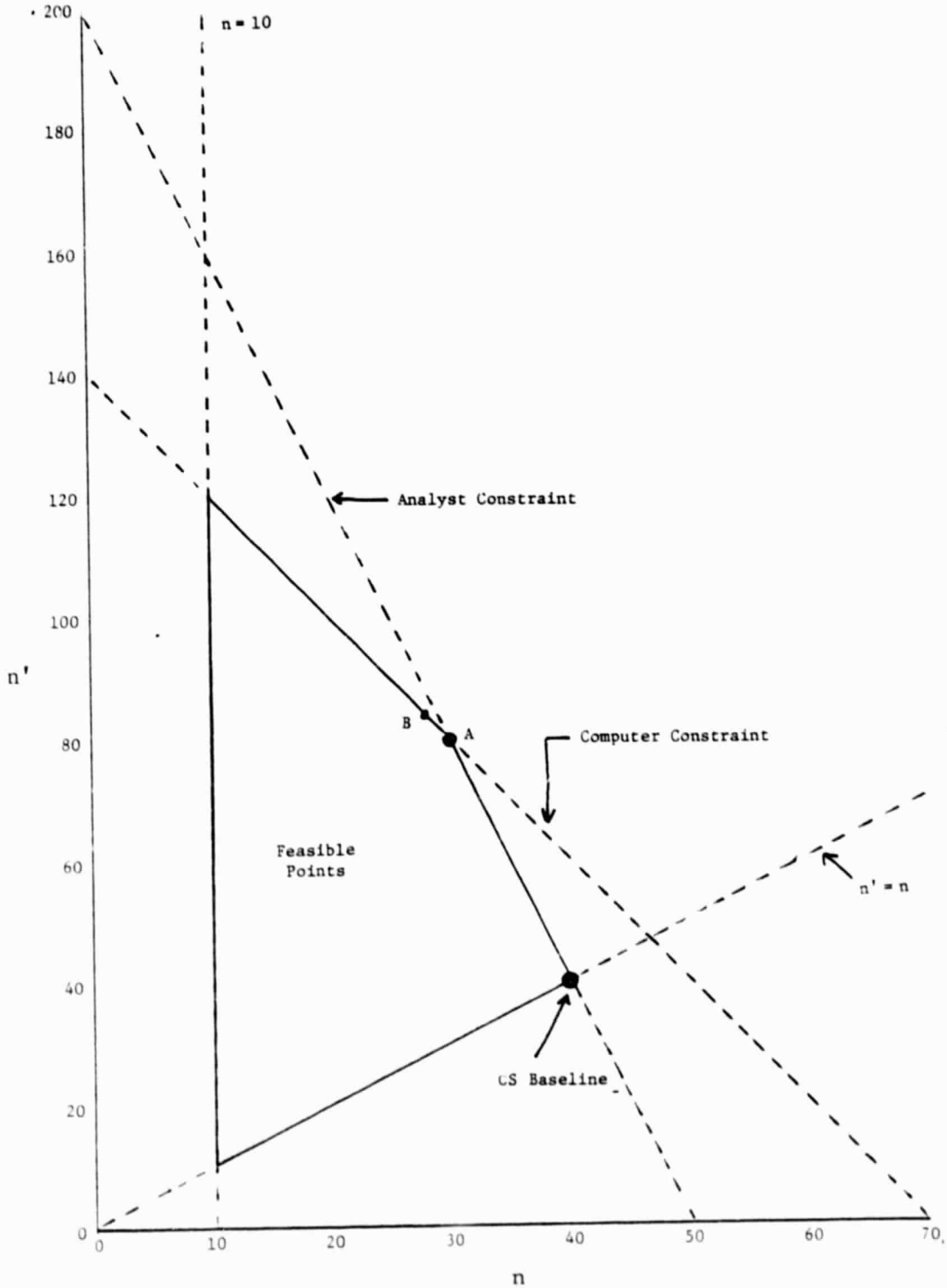

FIGURE 4. GEOMETRY FOR STAGE 2 WITH STAGE 1
(400 hrs analyst time, 35 computer hrs)

ORIGINAL PAGE IS
OF POOR QUALITY

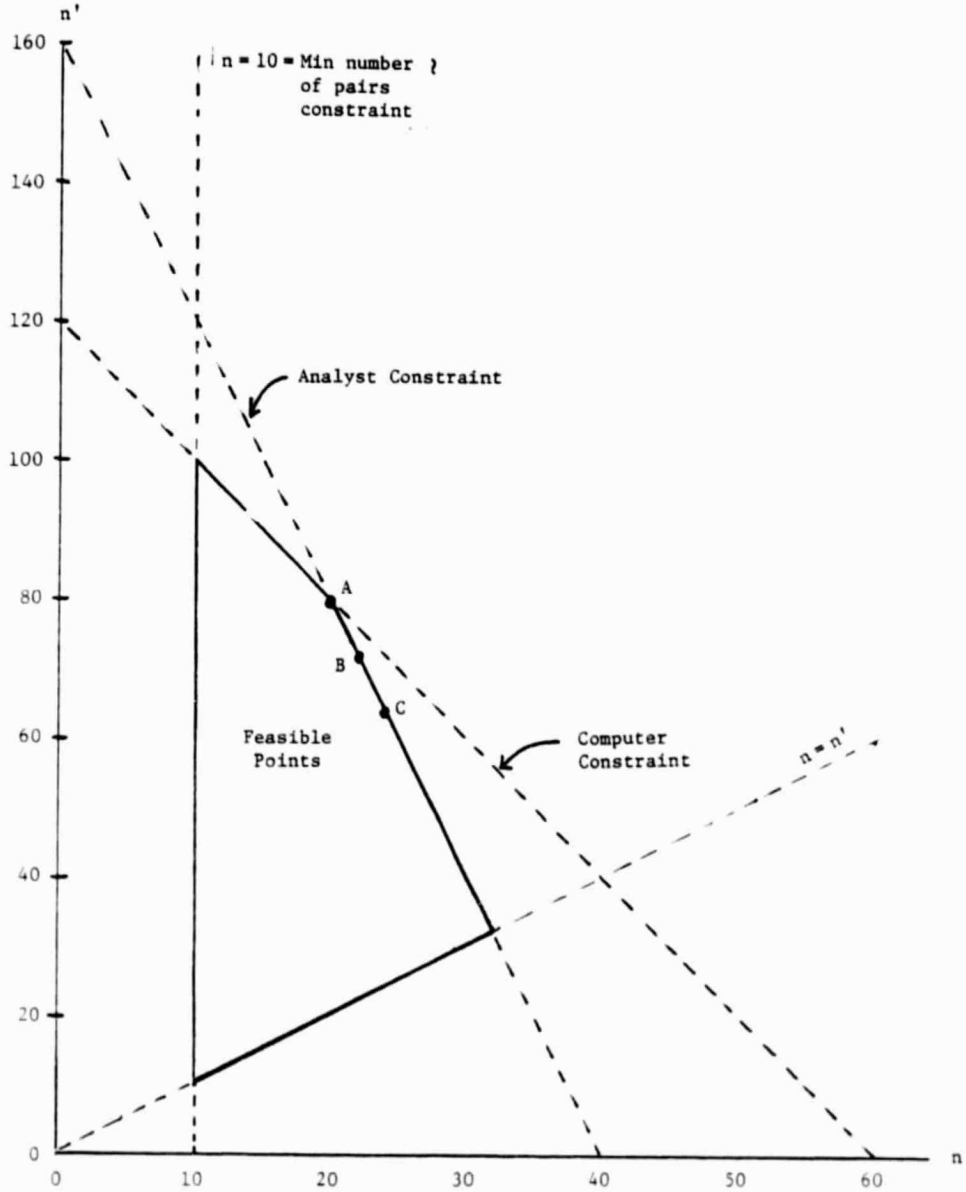

FIGURE 5. GEOMETRY FOR STAGE 2 WITH STAGE 1
(320 hrs analyst time, 30 computer hrs)

ORIGINAL PAGE IS
OF POOR QUALITY

Point A ($n = 20$, $n' = 80$) minimizes S_s^2,

Point B ($n = 22$, $n' = 72$) minimizes $S_c^2 + S_s^2$, and

Point C ($n = 24$, $n' = 64$) minimizes S_c^2 and max (S_c^2, S_s^2).

Perfect Estimates With Stage 2

In this example $\begin{pmatrix} y_c \\ y_s \end{pmatrix}$ denote the perfect (ground truth) estimates while $\begin{pmatrix} x_c \\ x_s \end{pmatrix}$ denote the Stage 2 estimates. We assume that these two estimation procedures do not compete with each other for resources. We therefore assume that the goal is to minimize the variance subject to a total cost constraint. The sample correlation of

$$\begin{pmatrix} y_c \\ x_c \\ y_s \\ x_s \end{pmatrix}$$

was

$$\begin{pmatrix} 1.00 & .89 & -.08 & .26 \\ - & 1.00 & .09 & .11 \\ - & - & 1.00 & .84 \\ - & - & - & 1.00 \end{pmatrix}.$$

Figures 6 and 7 give a scatterplot of ground truth vs. Stage 2 estimates. If the cost is $cn + c'n'$ then from Cochran [1] we obtain that double sampling gives a smaller variance if

$$\frac{c}{c'} > \frac{\rho^2}{\left(1 - \sqrt{1-\rho^2}\right)^2}$$

ORIGINAL PAGE IS
OF POOR QUALITY

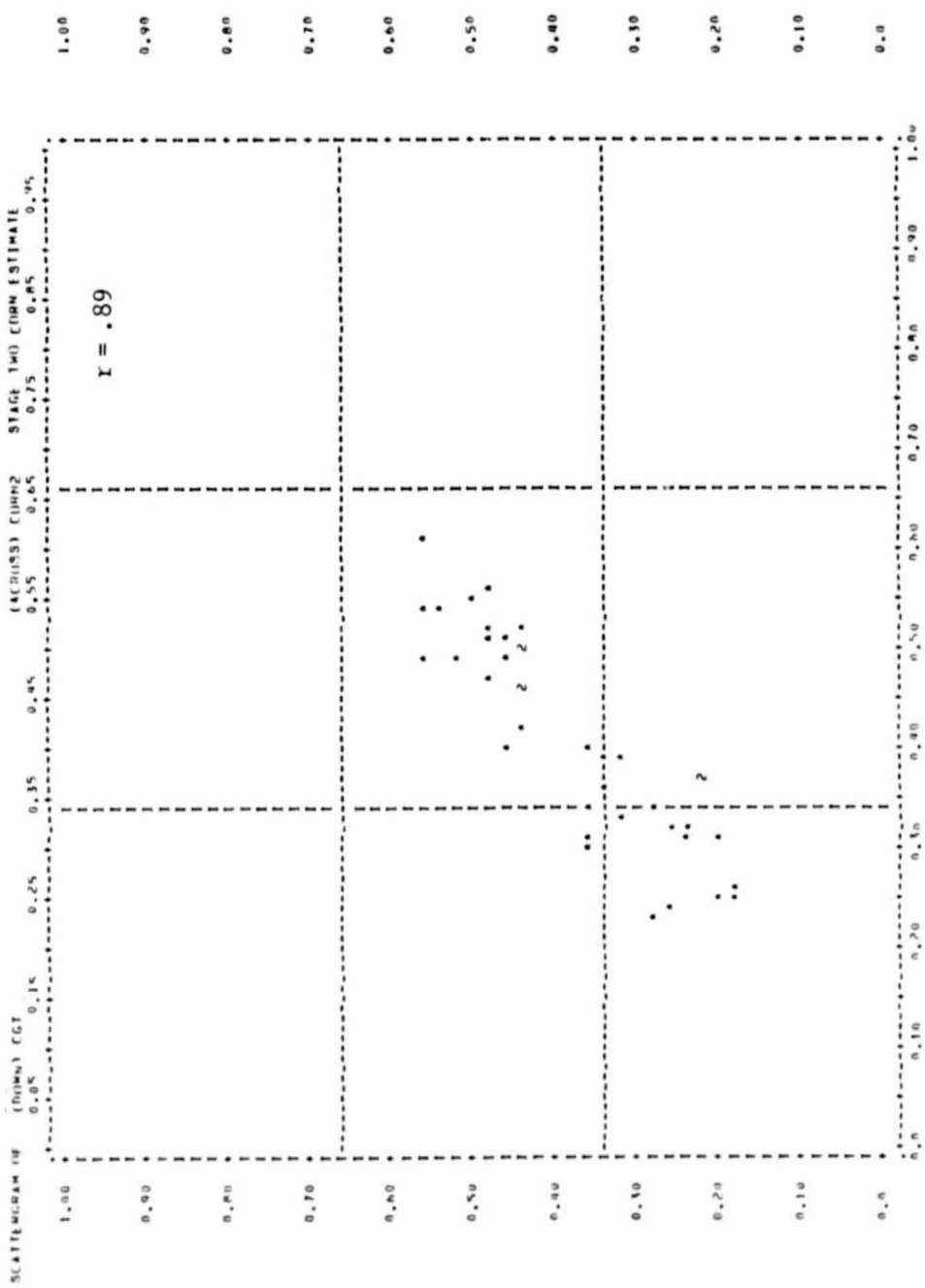

FIGURE 6. SCATTERPLOT OF GROUND TRUTH CORN VS. STAGE 2 CORN ESTIMATE

ORIGINAL PAGE IS
OF POOR QUALITY

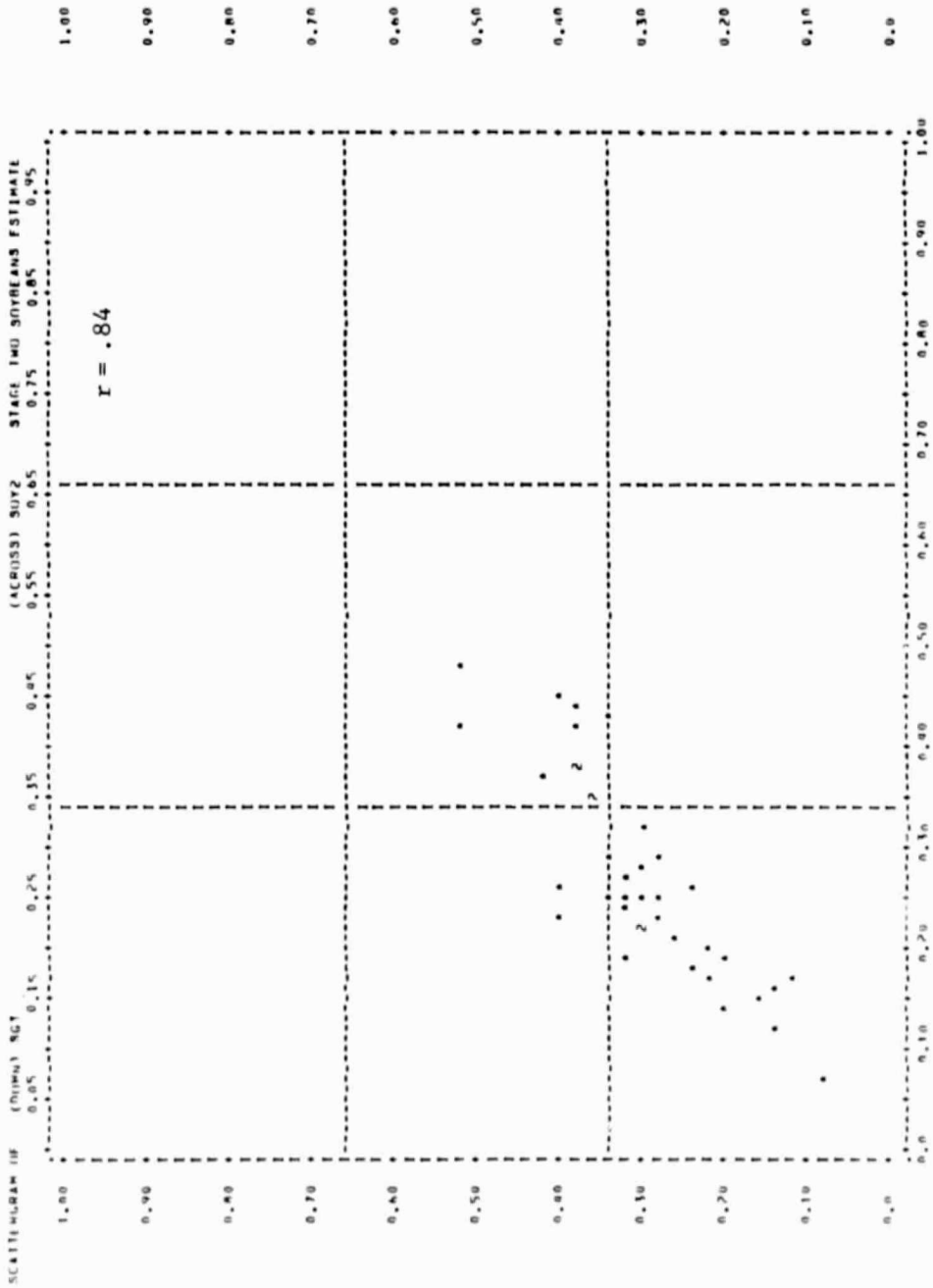

FIGURE 7. SCATTERPLOT OF GROUND TRUTH SOYBEANS VS. STAGE 2 SOYBEANS ESTIMATE

ORIGINAL PAGE IS
OF POOR QUALITY

For corn we obtain $\frac{c}{c'} > 2.68$ and for soybeans $\frac{c}{c'} > 3.37$. Thus, double sampling will be cost effective if the cost of the "perfect" estimates exceeds that of the Stage 2 estimates by a factor of about three or more.

Perfect Estimates With Stage 1

In the example we denote the Stage 1 corn/soybeans estimates as x_c, x_s. The nature of cost constr 'nts would be similar to the last example. The sample correlation matrix of

$$\begin{pmatrix} y_c \\ x_c \\ y_s \\ x_s \end{pmatrix}$$

was

$$\begin{pmatrix} 1.00 & .79 & -.08 & .21 \\ - & 1.00 & -.03 & .11 \\ - & - & 1.00 & .78 \\ - & - & - & 1.00 \end{pmatrix}$$

Figures 8 and 9 give scatterplots of ground truth vs. Stage 1 estimates. This implies that the double sampling gives a smaller variance if

$\frac{c}{c'} > 5.00$ for corn and

$\frac{c}{c'} > 4.34$ for soybeans.

In this case, the perfect estimate must be at least about 5 times more costly than the Stage 1 estimate in order that double sampling be cost effective. However, since the Stage 1 cost is only about one-fifth the Stage 2 cost (based upon analyst time, the more significant cost factor) this condition is more likely to be achieved than that identified for the Perfect/Stage 2 combination.

ORIGINAL PAGE IS
OF POOR QUALITY

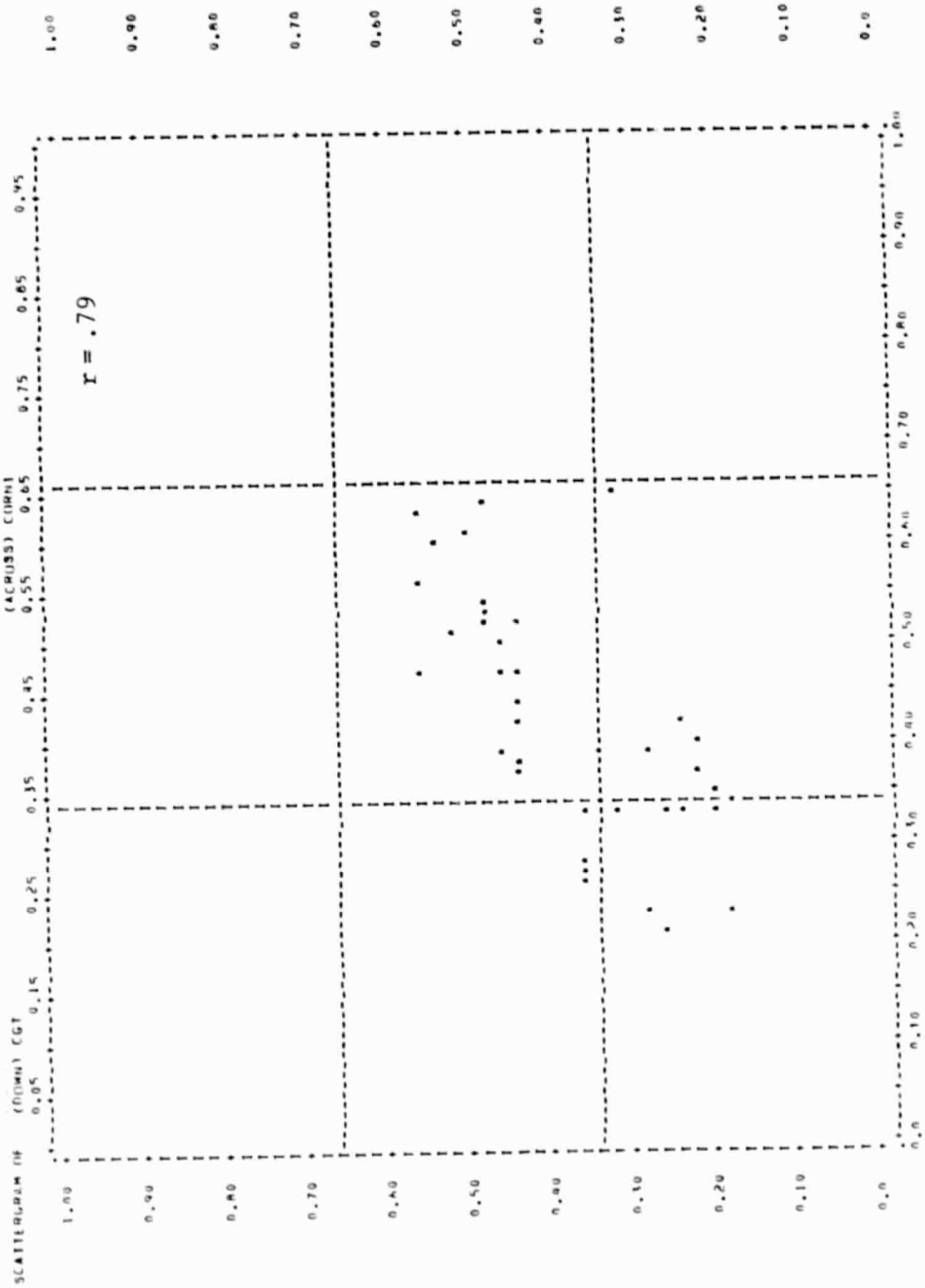

FIGURE 8. SCATTERPLOT OF GROUND TRUTH CORN VS. STAGE 1 CORN ESTIMATE

ORIGINAL PAGE IS
OF POOR QUALITY

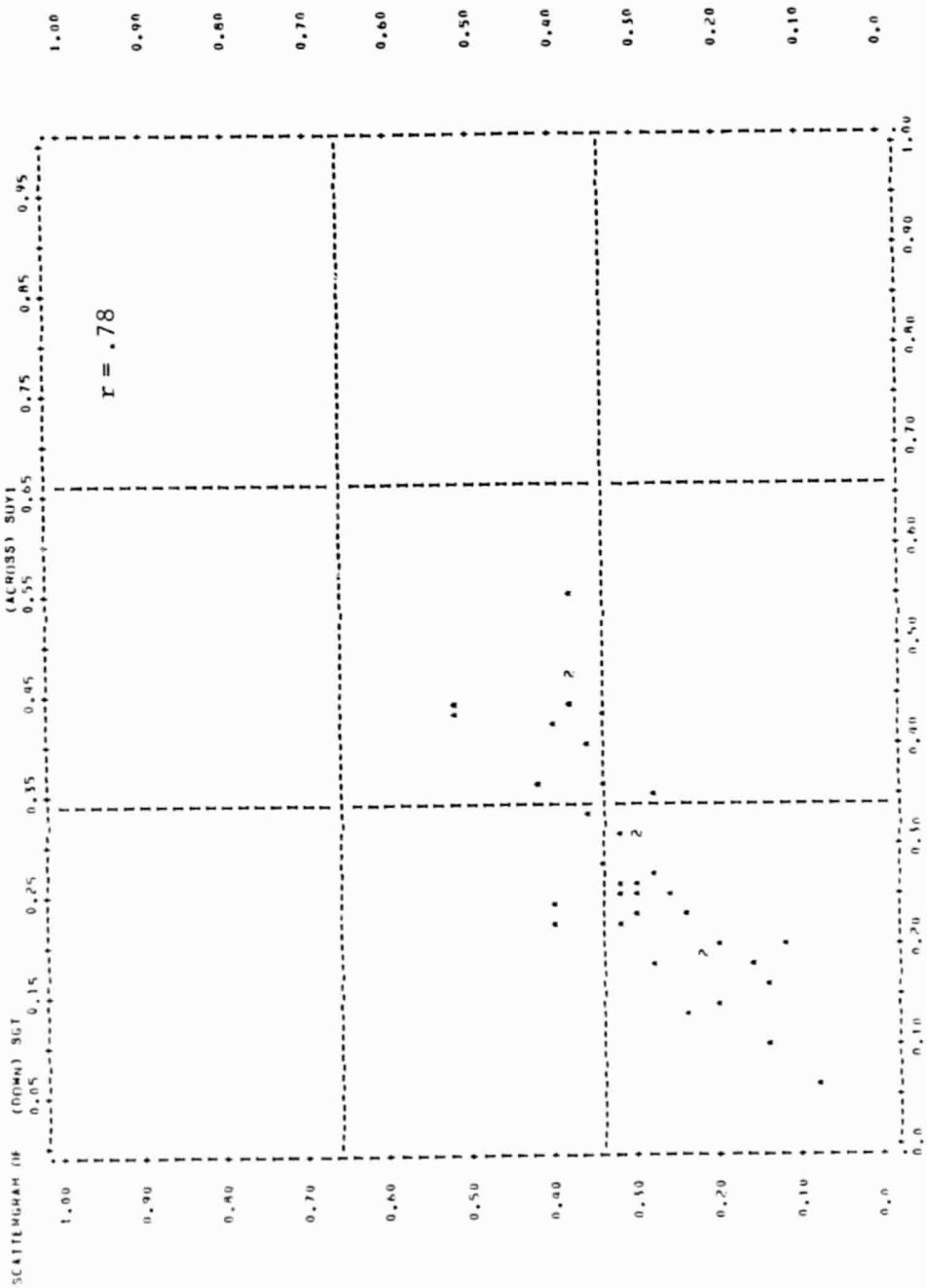

FIGURE 9. SCATTERPLOT OF GROUND TRUTH SOYBEANS VS. STAGE 1 SOYBEANS ESTIMATE

ORIGINAL PAGE IS
OF POOR QUALITY

Comparison of the Perfect/Stage 2 and the Perfect/Stage 1 Combination

Denote the cost of the perfect/Stage 2 estimate as $cn + c_2'n'$ and that of the perfect/Stage 1 estimate as $cn + c_1'n'$. Let ρ_2 and ρ_1 denote the correlations between the perfect and Stage 2 and between the perfect and Stage 1 estimates, respectively.

From Section 12.7 of Cochran [1] the best variance of the perfect/Stage 2 estimate is

$$V_2 = \frac{S_y^2\left(\sqrt{c\left(1-\rho_2^2\right)} + \sqrt{c_2'\rho_2^2}\right)^2}{C} - \frac{S_y^2}{N}$$

and the best variance of the perfect/Stage 1 estimate is

$$V_1 = \frac{S_y^2\left(\sqrt{c\left(1-\rho_1^2\right)} + \sqrt{c_1'\rho_1^2}\right)^2}{C} - \frac{S_y^2}{N} .$$

Thus $V_1 < V_2$ if

$$\left(\sqrt{c\left(1-\rho_1^2\right)} + \sqrt{c_1'\rho_1^2}\right)^2 < \left(\sqrt{c\left(1-\rho_2^2\right)} + \sqrt{c_2'\rho_2^2}\right)^2 .$$

Figure 10 gives plots of

$$\left(\sqrt{c\left(1-\hat{\rho}_1^2\right)} + \sqrt{c_1'\hat{\rho}_1^2}\right)^2 \quad \text{and} \quad \left(\sqrt{c\left(1-\hat{\rho}_2^2\right)} + \sqrt{c_2'\hat{\rho}_2^2}\right)^2$$

as a function of c for corn where $c_1' = 1$, $c_2' = 5$ and $\hat{\rho}_1$ and $\hat{\rho}_2$ are as in the last section. We note that $V_1 < V_2$ if $c \leq 58c_1$. Figure 11 gives the corresponding plots for soybeans. Here we note that $V_1 < V_2$ if $c \leq 174c_1$. Thus for moderate GT cost the perfect/Stage 1 combination will produce a lower variance estimate than the perfect/Stage 2 combination.

ORIGINAL PAGE IS
OF POOR QUALITY

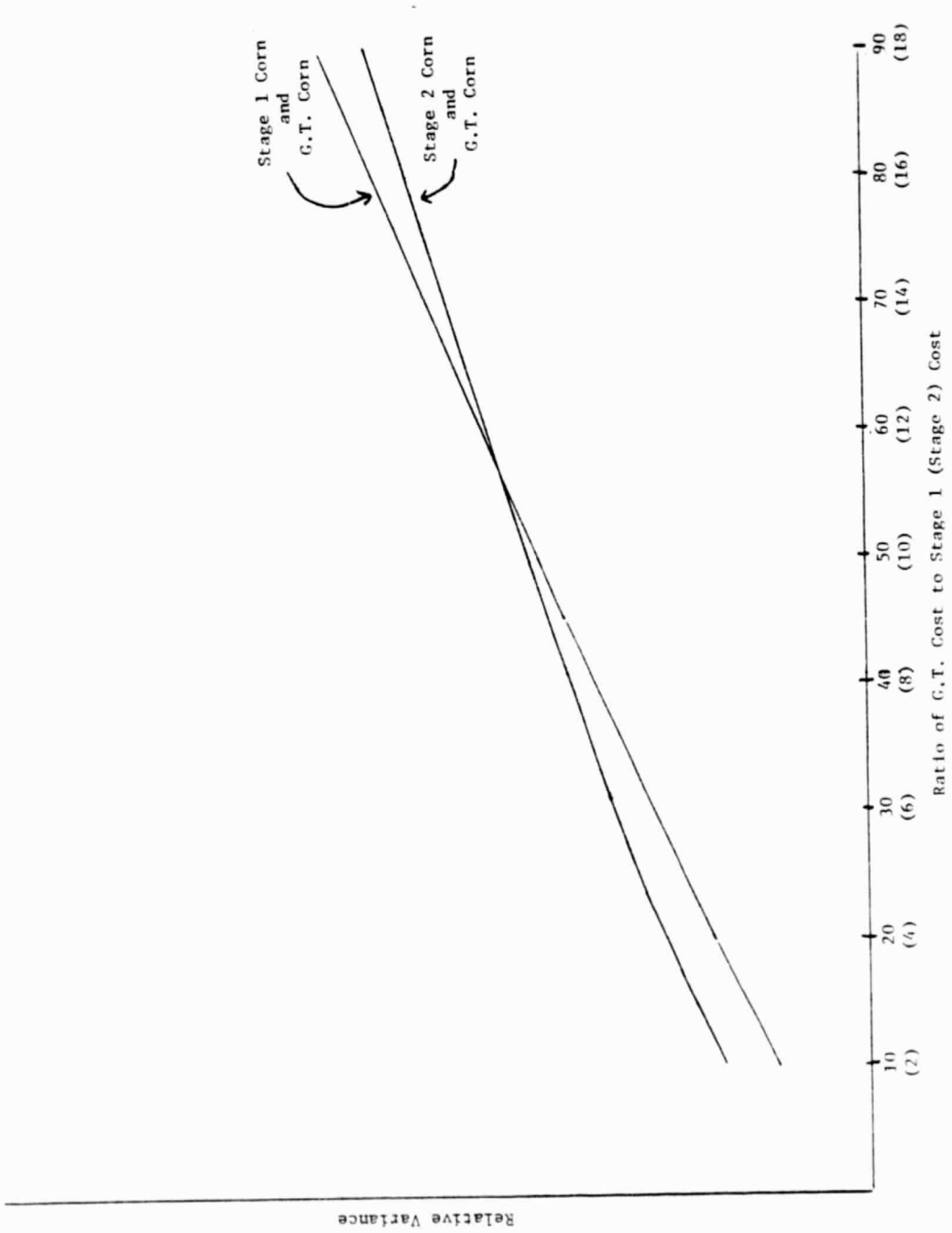

FIGURE 10. RELATIVE CORN VARIANCE AS A FUNCTION OF COST OF G.T. ESTIMATES

ORIGINAL PAGE IS
OF POOR QUALITY

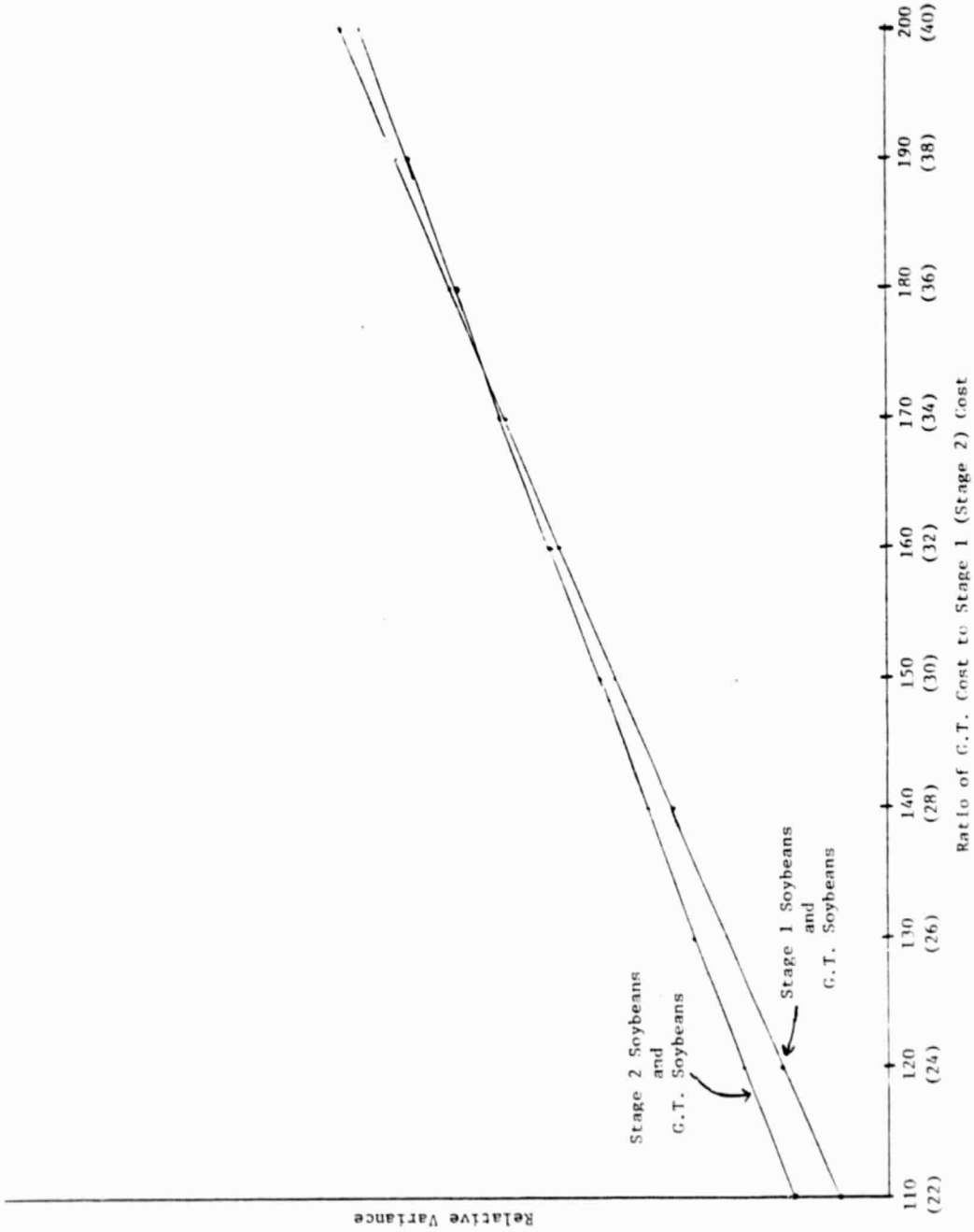

FIGURE 11. RELATIVE SOYBEAN VARIANCE AS A FUNCTION OF COST OF G.T. ESTIMATES

PRECEDING PAGE BLANK NOT FILMED

4

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

4.1 THE PERFECT PROCEDURE COMBINED WITH EXISTING LANDSAT PROCEDURES

The allocation of the number of perfect estimates, n, and the number of less costly estimates, n', can be solved by the use of classical double sampling techniques once the coefficients of the cost equation, $C = cn + c'n'$, are known and the correlations are known. These parameters allow one to choose between single sampling $n' = 0$ or double sampling $n' > n$. We do not know the cost of a perfect Landsat-based procedure, thus we cannot estimate c/c' in this case. However, the relative costs of the Stage 1 and Stage 2 corn and soybeans estimates allows us to conclude that, for a fixed cost, the variances of a perfect estimator combined with Stage 1 will likely be lower than the variances of that same estimator when combined with Landsat Stage 2. This result should be of interest to others who do have access to, and can determine costs for, "perfect" estimators (i.e., ground truth).

4.2 EXISTING LANDSAT PROCEDURES COMBINED

In the classical application of regression in double sampling we have the cost function and a constraint of the form

$$c(n,n') = cn + c'n' \leq C.$$

The single sampling point $(n,n') = \left(\left[\frac{C}{c}\right], 0\right)$, where $[\cdot]$ denotes the greatest integer function is a feasible point and is optimal if

$$\frac{c}{c'} < \frac{\left(1 + \sqrt{1 + \rho^2}\right)^2}{\rho^2}$$

The corn and soybeans baseline procedure automatically produces a Stage 1 estimate for every Stage 2 estimate, i.e., $n' \geq n$. Thus, the

single sampling point is not feasible in this case, and one must choose between $n' = n$ (the current procedure) or $n' > n$. In an operational situation, assuming that our 39 segment sample is reasonably representative, we expect a 25% to 50% decrease in variance for a fixed cost or a 25% to 50% decrease in cost for fixed variance by the use of double sampling.

4.3 GENERAL

The Stage 2 C/S Baseline estimate is representative of the state-of-the-art in the ability to make accurate measurements based entirely on Landsat data and intensive analyst interpretive activity. In areas of the world where ground truth is difficult to obtain, this intensive type of analysis is needed to provide the least possible bias in the resulting estimates. The Stage 1 estimate is only one of many possible less expensive uses of Landsat data, and its value in reducing costs when combined with the Stage 2 estimate encourages us to search for even less expensive alternatives to the Stage 1 estimate. One possibility is to automate the process of establishing a crop calendar for a segment, which currently is analyst intensive. Other possibilities involves various ratio estimators or other simple classifiers. In looking for cheaper alternatives, however, we note that the preprocessing which leads to the Stage 1 estimate (i.e., screening, haze and sun angle correction) is phenomenologically the correct kind of thing to do and has significant value in stabilizing the data which is important if high correlations between the Stage 2 estimate and any less expensive procedure are to be maintained.

When ground truth is available and is to be used to remove residual bias in the data, the Stage 1 estimate would be used rather than the Stage 2 estimate in a double sampling scheme since the lower cost of the Stage 1 estimate more than makes up for its smaller correlation to ground truth. Again, others who do have access to operational ground truth should be encouraged to look for even less expensive estimates which maintain their stability over large areas and a variety of conditions.

REFERENCES

1. Cochran, W. G., Sampling Techniques (3rd edition), J. Wiley & Sons, Inc., New York, 1977.

2. Roller, N., K. Johnson and J. Odenweller, Analyst Handbook for the Augmented U.S. Baseline Corn and Soybean Segment Classification Procedure (CS-1A), ERIM Report No. 152400-10-X (NASA Report No. FC-E1-00723), Environmental Research Institute of Michigan, Ann Arbor, Michigan, October 1981.

3. Kauth, R., and G. Thomas, "The Tasseled Cap -- A Graphic Description of the Spectral-Temporal Development of Agricultural Crops as Seen by Landsat", Proc. of 1976 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, W. Lafayette, Indiana, 1976.

4. Midzuno, H., "On the Sampling System with Probability Proportionate to the Sum of Sizes", Ann. Inst. Stat. Math., 2, pp. 99-108 (1951).

5. Wigton, W. H., and Huddleston, "A Land Use Information System Based on Statistical Inference", Twelfth International Symposium on Remote Sensing of Environment, Manila, Philippines, April 20-26, 1978.