

The Use of Regression Estimation
With LANDSAT and Probability Ground Sample Data

by

R. S. Sigman, G. A. Hanuschak, M. E. Craig,
P. W. Cook, and M. Cardenas

I. INTRODUCTION

The Economics, Statistics, and Cooperatives Service (ESCS) of the U.S. Department of Agriculture is presently conducting research in possible uses of LANDSAT satellite data in agricultural surveys. This research is in the following areas:

1. improvement of crop-hectare estimates for multi-county areas, such as Crop Reporting Districts and states,
2. development of small-area crop-hectare estimates for individual counties, and
3. photo-interpretive use of LANDSAT imagery in developing area sampling frames.

This paper briefly describes ESCS's statistical methodology and discusses some recent applications in using LANDSAT data to improve crop-hectare estimates for multi-county areas. ESCS's research in developing small-area estimates from LANDSAT data is discussed in another paper at this conference [1]. Hanuschak and Morrissey [2] describe ESCS's use of LANDSAT imagery in developing area sampling frames.

II. DATA SOURCES

A. GROUND-SURVEY DATA

As a part of its operational program, ESCS conducts in late May an annual nationwide agricultural survey called the June Enumerative Survey (JES). The

THESE DOCUMENTS SONT EN PARTIELLEMENT
REPRODUITS EN FRANCAIS DANS LE
DOCUMENT D'IDENTIFICATION
DE LA COMMISSION INTERNATIONALE
DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE
ET DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE
ET DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE

LES DOCUMENTS SONT EN PARTIELLEMENT
REPRODUITS EN FRANCAIS DANS LE
DOCUMENT D'IDENTIFICATION
DE LA COMMISSION INTERNATIONALE
DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE
ET DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE
ET DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE

LES DOCUMENTS SONT EN PARTIELLEMENT
REPRODUITS EN FRANCAIS DANS LE
DOCUMENT D'IDENTIFICATION
DE LA COMMISSION INTERNATIONALE
DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE
ET DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE
ET DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE
ET DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE

LES DOCUMENTS SONT EN PARTIELLEMENT
REPRODUITS EN FRANCAIS DANS LE
DOCUMENT D'IDENTIFICATION
DE LA COMMISSION INTERNATIONALE
DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE
ET DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE
ET DE LA MER MEDITERRANEE
ET DE LA MER NOIRE
ET DE LA MER EGEE



hectare area of the earth's surface. The MSS measures the amount of radiant energy reflected and/or emitted from the earth's surface in various regions (bands) of the electromagnetic spectrum. The LANDSAT II and LANDSAT III MSS's have four and five bands, respectively.

The individual .4 hectare MSS resolution areas, referred to as pixels, are arrayed along east-west running rows within the 185 kilometer wide north-to-south pass of the LANDSAT satellite. A given point on the earth's surface is imaged once every eighteen days by the same LANDSAT satellite and once every nine days by either one of two satellites. Satellite passes which are adjacent on the earth's surface are at least one day apart with respect to their dates of imagery.

III. STATISTICAL METHODOLOGY

ESCS's approach for using LANDSAT data is to use it as an auxiliary variable with data acquired from operational ground surveys [3]. The information from these surveys is actually used twice in the ESCS procedure for computing LANDSAT-based crop-hectare estimates. The ground-survey data is used (1) as "ground-truth" for developing a set of discriminant functions for the LANDSAT data, and (2) as the primary survey variable for estimating crop-hectare.

A. DIRECT EXPANSION ESTIMATION (GROUND DATA ONLY)

The estimation procedure presented here is for a given state. National totals are then obtained by appropriately combining state totals.

Let $h = 1, 2, \dots, L$ be L land-use strata. Within each stratum, the total area is divided into N_h area-frame units from which a simple random sample of n_h units is drawn. Using only JES data for the L strata, an estimate of total hectares of a particular crop (corn, for example) can be computed by direct expansion as follows:

1. The first part of the document discusses the importance of maintaining accurate records of all transactions. It emphasizes that proper record-keeping is essential for the integrity of the financial system and for the ability to detect and prevent fraud. The document also notes that records should be kept for a sufficient period of time to allow for a thorough review if necessary.

2. The second part of the document outlines the specific requirements for record-keeping. It states that all transactions must be recorded in a clear and concise manner, and that the records must be accessible to authorized personnel at all times. The document also requires that records be kept in a secure and confidential manner, and that they be protected from unauthorized access or disclosure.

3. The third part of the document discusses the role of internal controls in maintaining accurate records. It notes that internal controls are essential for ensuring the accuracy and reliability of the financial records. The document also emphasizes that internal controls should be designed to prevent errors and to detect and prevent fraud. Finally, the document states that internal controls should be regularly reviewed and updated to ensure their effectiveness.

4. The fourth part of the document discusses the importance of training and education in maintaining accurate records. It notes that all personnel who are involved in the financial system should receive appropriate training and education. The document also emphasizes that training and education should be ongoing and should cover all aspects of the financial system, including record-keeping, internal controls, and fraud prevention.

5. The fifth part of the document discusses the importance of monitoring and reporting. It notes that all transactions should be monitored and reported to the appropriate authorities. The document also emphasizes that monitoring and reporting should be done in a timely and accurate manner, and that any suspicious activity should be reported immediately.

6. The sixth part of the document discusses the importance of documentation. It notes that all transactions should be properly documented, and that the documentation should be clear and concise. The document also emphasizes that documentation should be kept in a secure and confidential manner, and that it should be accessible to authorized personnel at all times.

7. The seventh part of the document discusses the importance of communication. It notes that all personnel who are involved in the financial system should be kept informed of any changes or updates to the financial system. The document also emphasizes that communication should be done in a clear and concise manner, and that it should be done in a timely and accurate manner.

8. The eighth part of the document discusses the importance of compliance. It notes that all transactions must be conducted in accordance with applicable laws and regulations. The document also emphasizes that compliance should be a top priority for all personnel who are involved in the financial system, and that any violations should be reported immediately.

11
12
13
14
15
16
17
18
19
20

21
22
23
24
25
26
27
28
29
30

31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200

201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300

301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400

401
402
403
404
405
406
407
408
409
410

411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500

1. The first part of the document discusses the importance of maintaining accurate records of all transactions. This is essential for ensuring the integrity of the financial statements and for providing a clear audit trail. The records should be kept up-to-date and should be easily accessible to all relevant parties.

2. The second part of the document outlines the various methods used to collect and analyze data. These methods include interviews, surveys, and focus groups. Each method has its own strengths and weaknesses, and it is important to choose the most appropriate method for the specific research objectives. The data collected should be analyzed carefully to identify any trends or patterns that may be significant.

3. The third part of the document describes the results of the research. The findings indicate that there is a strong correlation between the variables studied. This suggests that the factors being investigated are closely related and may have a significant impact on the outcome. The results should be interpreted in the context of the research objectives and the existing literature.

4. The fourth part of the document discusses the implications of the research. The findings have several practical implications for the industry and for policy makers. It is important to consider these implications carefully and to develop strategies that address the issues identified. Further research is needed to explore these issues in more detail and to test the findings in a wider context.

5. The fifth part of the document concludes the research and provides a summary of the key findings. The research has shown that there is a clear relationship between the variables studied and that this relationship has important implications for the industry and for policy makers. The findings should be used to inform decision-making and to guide the development of strategies and policies. The research also highlights the need for further investigation in this area.

$$\text{and } \bar{y} = \left(\sum_{h=1}^L N_h \bar{y}_h \right) / N.$$

The approximate variance of the combined regression estimator and the expression for \hat{b}_c are given in Cochran [4, pp 202-203].

When a LANDSAT pass does not cover the entire state on one date, it is necessary to partition the state into analysis areas which are wholly contained within the individual passes. The estimation procedure described above is carried out in each analysis area, and then analysis-area-level estimates as well as variances are combined to the state level by treating the analysis areas as post-strata.

The relative efficiency of the regression estimator compared to the direct expansion estimator will be defined as the ratio of the respective variances:

$$\text{R.E.} = v(\hat{Y}_{DE}) / v(\hat{Y}_R). \quad (3)$$

The auxiliary variables described above, i.e.

$$x_{hj} = \sum_k c(z_{hjk}) \text{ and } X_{hj} = \sum_k c(Z_{hik}) \quad (4)$$

where the variable z_{hjk} (Z_{hik}) is the signature of the k^{th} pixel of the j^{th} sample unit (i^{th} area-frame unit) in the h^{th} stratum and the function $c(z)$ is 1 if signature z is classified as the crop of interest and 0 otherwise. These auxiliary variables are probably not optimum in the sense of producing the estimate of Y with smallest possible variance. Alternate approaches which are being investigated are

1. using a multiple regression estimator, where the set of auxiliary variables includes not only the quantities in equation (4) but also the classification results into cover types other than the crop of interest (discussed in [5]); and

2. changing $c(z)$ in equation (4) to the posterior probability that a pixel with signature z is from the crop of interest. The posterior probability

function can be estimated by approximating it with a linear combination of basis functions with the coefficients estimated by least squares (suggested by Fuller [6]) or by assuming a logistic form for the posterior probability and then estimating unknown parameters by maximum likelihood.

C. PIXEL CLASSIFICATION

The pixel classifier is a set of discriminant functions corresponding one-to-one with a set of classification categories. Each discriminant function consists of the category's likelihood multiplied by the category's prior probability. If the prior probabilities used are correct for the population of pixels being classified, then the resulting set of discriminant functions, called a Bayes classifier, minimizes the over-all probability of misclassifying a pixel.

In crop-hectarage estimation, however, the objective is to minimize the variance of resulting hectarage estimates. Since minimizing the over-all probability of misclassification does not necessarily achieve this objective, optimum hectarage estimation may require the use of prior probabilities different from the optimum Bayes set. (Strictly speaking, there is only one correct set of prior probabilities for a given geographical region, i.e. the actual probabilities of occurrence for the various cover types. Using "different prior probabilities" actually means using different weighting factors for the category likelihoods in computing the category discriminant functions.) We have investigated two types of "prior probabilities": equal probabilities and probabilities proportional to direct-expanded hectarage, i.e. the \hat{Y}_{DE} 's. The results of this investigation are discussed in the next section.

Since the type of ground cover in every JES field is known as a result of JES enumeration, the pixels lying inside JES fields are of known cover type. These pixels, called field-interior pixels, determine the cover types for which

classification categories are created. In addition, pixels are selected from rivers, lakes, and ponds to determine classification categories for surface water.

The field-interior pixels for a given cover type are extracted from the LANDSAT data, and the corresponding signatures are clustered in MSS measurement space. A classification category is then associated with each cluster which has more than some specified number of pixels (usually 100 pixels).

Category likelihoods are computed by assuming that the signatures in a given category follow a multivariate normal distribution. Thus, the calculation of category discriminant functions involves the estimation by category of signature means and covariances and prior probabilities. Once this has been done, all the JES segment-interior pixels (field-boundary pixels included) can be classified and the sample coefficient of determination

$$r_h^2 = \frac{\sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h) (x_{hj} - \bar{x}_h)]^2}{\sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2 \sum_{j=1}^{n_h} (x_{hj} - \bar{x}_h)^2}$$

calculated. In small samples, however, r_h^2 can have a large positive bias as an estimate of R_h^2 because much of the same data is used to both develop the sample discriminant functions and to compute r_h^2 . Less biased estimates for R_h^2 can be obtained by many of the same methods used to estimate error rates in discriminant analysis; e.g., jackknifing, sample partition, etc. We have found, however, that in moderate size samples, e.g. $n_h = 84$, that the difference between r_h^2 and a jackknifed estimate of R_h^2 is acceptably small so as to not warrant the additional labor involved in performing the jackknife calculations [7, 8].

IV. RECENT APPLICATIONS

ESCS has applied the methodology described above in a number of different areas in the U.S. over the past several years. Major demonstration efforts have been conducted in Illinois, Kansas, and Kings County, California. All of these studies have been performed in a purely research mode, and except for the 1977 study effort in Kings County, California, none of these demonstration projects have produced timely crop hectarage estimates. Also, this methodology is not yet demonstrably cost effective. In 1978, however, ESCS expects to complete LANDSAT crop-hectarage estimates in time for input to USDA final season estimates for Iowa.

A. 1975 ILLINOIS STUDY [7, 8]

1975 LANDSAT data for the entire state of Illinois was used to estimate crop hectarages for Illinois spring-seeded crops at county and multi-county levels. Requiring three LANDSAT passes to completely image the state, the dates of imagery of the analyzed LANDSAT data ranged from July 16 to September 7. On account of the different dates of analyzed LANDSAT data, the state was partitioned into six analysis areas. The distribution of the 300 Illinois JES segments into the six areas ranged from 30 to 84 segments per analysis region.

The separate form of the regression estimator was used in Illinois. Cover types for which classification categories were created were corn, soybeans, alfalfa, other hays, permanent pasture, wheat stubble, oats and oat stubble, dense woodlands, water, and other non-agricultural land (called waste). Only for corn, soybeans, water, and waste, however, did the use of LANDSAT data result in significant increases in precision (relative to using JES data alone) of analysis-area crop-hectarage estimates. For the analysis-area estimates, the regression estimate relative efficiencies for corn ranged from 1.3 to 6.3; for soybeans, from 1.1 to 5.8.

One of the major factors determining the ability of LANDSAT data to improve crop-hectarage estimates was the acquisition date of the LANDSAT imagery. Best results were obtained for August 3 and 4, when corn was nearly 100% silked. In the calculation of category discriminant functions, it was observed that using equal prior probabilities yielded more precise crop-hectarage estimates (compared to using probabilities proportional to direct expanded hectares) in most cases for corn and in some cases for soybeans.

B. 1976 KANSAS STUDY

The objective of this study was to estimate winter wheat hectarages for Kansas using 1976 LANDSAT data. In order to completely image the state, six LANDSAT passes are required. The easternmost pass, covering only four counties, was not analyzed because of insufficient JES data to estimate the required parameters. Also, the central pass was almost completely cloud covered during April, May, and June, causing loss of LANDSAT acquisitions for some major wheat-producing counties. Acquired from April 1 to May 6, usable LANDSAT data covered 87 of the 105 Kansas counties.

A 40% subsample of segments from the Kansas JES was used in the LANDSAT analysis. The number of segments in the subsample ranged from 11 to 35 per pass. The combined form of the regression estimator was used because of the small number of segments from the subsample within each stratum in a pass. Since only winter wheat estimates were of interest, classification categories were created only for wheat and 'other'. The 'other' cover type was a catch-all name for anything (crop, waste, pasture, etc) not labelled as winter wheat by the USDA enumerators.

Sample coefficients of determination between classification results and ground truth were high, ranging from .60 to .92. Relative efficiencies (with respect to the subsample) ranged from 3.1 to 13.0, with the exception of the

central pass. This pass was mostly cloud covered and analysis was done for only 7 counties using 11 segments. The resulting relative efficiency was slightly less than one.

C. 1977 CALIFORNIA STUDY

In both 1976 and 1977, crop-hectare estimates using LANDSAT data were calculated for Kings County, California. In 1977, timeliness of the estimates was a primary objective. This goal was successfully achieved: using July 7 LANDSAT data, the analysis was completed on August 15, 1977.

Kings County is several times larger in size than a typical Illinois or Kansas county. In 1977, sixty JES segments were allocated to the county. From these a random sub-sample of fifteen segments was selected for use in the LANDSAT study.

Major crops were cotton, barley, wheat, and alfalfa. For these crops all r_h^2 values exceeded 0.80 and regression estimator relative efficiencies (with respect to sub-sample direct expansion) ranged from 5.2 to 28.0.

V. REFERENCES

1. Cardenas, Manuel; Blanchard, Mark M.; Craig, Michael E.; "Small Area Estimators: County Crop Acreage Estimates Using LANDSAT Data," contributed paper, 1978 annual ASA meeting, San Diego, California.
2. Hanuschak, George A. and Morrissey, Kathleen M., "Pilot Study of the Potential Contributions of LANDSAT Data in the Construction of Area Sampling Frames," Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C., October 1977.
3. Von Steen, Donald H. and Wigton, William H., "Crop Identification and Acreage Measurement Utilizing LANDSAT Imagery," Statistical Reporting Service, United States Department of Agriculture, Washington, D.C., March 1976.
4. Cochran, William G., Sampling Techniques, (2nd Ed.), John Wiley & Sons, 1963.
5. Hanuschak, George A. and Cardenas, Manuel, "Multiple Regression Estimation Using Classified LANDSAT Data," Economics, Statistics, and Cooperative Service, U.S. Department of Agriculture, Washington, D.C., April 1978.

6. Fuller, Wayne, personal communication to William Wigton, December 1977.
7. Sigman, Richard S.; Gleason, Chapman P.; Hanuschak, George A.; and Starbuck, Robert A.; "Stratified Acreage Estimates in the Illinois Crop-Acreage Experiment," Proceedings of the 1977 Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana.
8. Gleason, Chapman; Starbuck, Robert R.; Sigman, Richard S.; Hanuschak, George A.; Craig, Michael E.; Cook, Paul W.; and Allen, Richard D.; "The Auxiliary Use of LANDSAT Data in Estimating Crop Acreages: Results of the 1975 Illinois Crop-Acreage Experiment," Statistical Reporting Service, U.S. Department of Agriculture, Washington D.C., October 1977.