87-05

DRAFT CLASSIFIER STUDY

NEDS Paper

I. INTRODUCTION

NED JONES · 1987

The USDA's National Agricultural Statistics Service (NASS) has been using Landsat data as an auxiliary in conjunction with data, the June Enumerative Survey (JES) to produce Domestic Crop and Land Cover (DCLC) estimates since 1978. During this period, 13 of 15 State-level DCLC corn estimates were less than the JES and 13 of 16 State-level DCLC soybeans estimates and 14 of 18 State-level DCLC winter wheat estimates were less than their comparable JES estimates. This has led us to question whether the DCLC estimates are biased, because we assume the JES estimates are unbiased. Details of the DCLC, JES data series are given in Appendix A (Tables A.1 and A.2).

Some have suggested that the JES over expands. The reason behind this theory is the average actual sample segment size is larger than the target size. A segment is the sampling unit in the JES. So when we expand the "large" segment size by the expected number of frame units associated with the target segment size the resulting estimates are too large. Based on the results in Table 1 the JES does not over expand because of "large" segment size. Table 1 compares the Census Land Area, the 1984-85 JES direct-expansion estimates of total land planimetered acres and total land reported acres and the DCLC total land estimates. If we have a problem with segment size in the selected JES segments the DCLC total land estimates would correct for the "large" segment size. According to Cochran a regression estimator similar to the DCLC estimator would correct for the proposed segment size problem (1). When the DCLC regression estimate of total land reported was compared with the JES direct-expansion estimate of total land reported a consisted downward bias was not present.

Another possible explanation of the difference in level of the JES and DCLC

estimates was that the DCLC corrected for reporting error such as unreported waste in a corn field. If we look at the JES direct-expansion estimate and the DCLC regression estimates we can see why the regression estimate is not correcting for reporting errors. First of all, within each stratum and analysis district the estimators are composed of the following:

N= total number of frame units in the population

$\bar{y}$ = segment mean acres of crop Z reported

$\bar{x}$ = sample segment mean number of pixels classified to crop Z

$\bar{X}$ = population segment mean number of pixels classified to crop Z

Y= population estimate of total acres reported of crop Z

b= the regression slope when y_i segment reported acres of crop Z was regressed on x_i, segment number of pixels classified to crop Z.

The JES direct-expansion estimate is as follows:

$$Y_{JES} = N \bar{y}$$

The DCLC regression estimates is as follows:

$$Y_{DCLC} = N (\bar{y} + b (\bar{X} - \bar{x})$$

If we look at the difference between the JES direct-expansion estimate and the DCLC regression estimate we discover the following:

$$Y_{JES} - Y_{DCLC} = N \bar{y} - N \bar{y} + b (\bar{X} - \bar{x})$$
$$= - N b (\bar{X} - \bar{x})$$

We notice that the $\bar{y}$ which contained the reporting error disappeared from our expression. So the difference between the two estimates could not be due to reporting error.

The only factors left which could be responsible for the difference $Y_{JES} - Y_{DCLC}$ are N, b, $\bar{X}$, and $\bar{x}$. The N does not appear to be the problem because a difference due to N would be indicated in Table 1. If the N was incorrect the DCLC total land estimates would be consisted off.

So the

Bias = N (E (b ($\bar{X}$ -

If we let

$= E(\bar{x}) - \bar{X}$

Bias $= N E(b(E(x) - -\bar{x})$

$= N(-E(b(\bar{x} - E(\bar{x} + b)$

$= -N(\text{cov } b, \bar{x}) + E(b)$

The -cov (b, $\bar{x}$) is the bias of the regression estimator according to Cochran (2) and represents a contribution from the quadratic component of the regression of y or $\bar{x}$. If a sample plot of y_i against x_i appears approximately linear there should be little risk of major bias in Y_{DCLC} due to -cov (b, $\bar{x}$). Since plots of the x_i, counts of the pixels classified to crop Z against y_i farmer reported acres of crop Z are approximately linear the cov (b, $\bar{x}$) does not appear to be the source of the bias.

The second term $E(b) = E(b)(E(\bar{x}) - \bar{X})$

In the theory we would expect the second term to be zero because the $E(\bar{x}) = X$; however, this must not be the case because the $Y_{JES} - Y_{DCLC}$ is usually greater than zero.

II. OVERVIEW

The DCLC procedure combines farmer reported data from the JES with Landsat

data as a auxiliary variable to produce regression estimate. A detailed explanation of the JES can be found in "Area Sampling in Agriculture", E. E. HOUSEMAN (3) and "Scope and Methods of the Statistical Reporting Service" (4). A detailed explanation of the DCLC methodology can be found in "Landsat Large-Area Estimates for Land Cover", May, Holko, and Jones (5). The following is an outline of the procedure and the result of each step in processing an analysis district and the associated segments. An analysis district is an area of land used for analysis covered by a single data in the same path.

1) Using aerial photograph and 7.5' maps the segments are registered to a map base.

   Result: We know the latitude and longitude of every point in the segment and it is in a computer readable form.

2) Digitize county maps which identify land use strata boundaries.

   Result: We know the latitude and longitude of strata boundaries and it is in a computer readable form.

3) Build a file of counts of the number of total potential segments in each strata within each county.

   Result: We have the number of potential segments (frame units) by strata in each county in computer readable form.

4) Ground data for the sampled area segment is collected and edited (JES and Field Level Edit).

Result: We have identified the location and cover type in each field i.e., corn, waste, double-crop, winter wheat-soybeans, etc., and the farmer reported number of acres in each field and put it in a computer readable form. (A field is a continuous area of land devoted to one use.)

5) Using the aerial photography of the segment used in enumeration digitize the segment field boundaries and label fields.

Result: We know the location of fields and field boundaries within each segment in a computer readable form.

6) Acquire Landsat data and register to a map base.

Result: We know the latitude and longitude of each pixel in the Landsat scene and visa versa the Landsat row and column of every possible latitude and longitude associated with the scene. The data are in a computer readable form.

7) Combine (Pack) the Landsat data and the JES ground data into files by cover type, i.e. all corn in one file and also a file of all cover types together.

Results: We have several files of pixels from the segments each pixel labeled by segment and field, one file for each different cover type is i.e., a file of corn pixels, a file of wheat pixels, and a file of waste pixels ect.

We also have a file of all pixels from all the segments within each segment with each pixel labeled by segment field and cover type.

7) Each file of pixels associated with a use is run through one of two clustering algorithm. If their are more than 250 pixels in the file, the CLASSY clustering algorithm (6) is used. CLASSY is a maximum likelihood algorithm which returns clusters of pixels which have multivariate normal distribution. If the file has more than 75 points, but less than 250, we use the ordinary clustering algorithm a deviation of the ISODAT algorithm of Ball and Hall (7). The mean vector and covariance matrix are calculated for the resulting clusters.

Result: The pixels associated with each cover type are divided into groups of pixels with similar reflectance. Since we know the mean and variance of each cluster we can estimate the distribution associated with each cluster.

8) We edit the distributions resulting from clustering throwing out distributions with less than 75 pixels. The remaining distributions are each assigned to a category. The means and covariance of each category are used to develop a maximum likelihood classifier.

Result: A mechanism is developed that can classify pixels to their associated cover type based on their reflectance.

9) All of the pixels from the segments are classified to cover type using the maximum likelihood classifier developed from the category means and covariances.

Result: We know the number of pixels in each segment classified to each cover type.

10) Estimate the parameters associated with the regression of farmer reported acres of crop Z per segment on the number of pixels classified to crop Z per segment.

Result: We know the slope and intercept of the regression equation farmer reported acre on classified number of pixels of crop Z

mean number of pixels per segment classified to cover

mean number of reported acres of crop Z.

n mean number of segments in the stratum in the analysis district

Which allow us to develop the estimator

$$Y_D = N (\bar{y} + b (X - \bar{x}))$$

with variance

where the coefficient of determination is

11) Classify and aggregate all the pixels associated with the analysis district.

Result: We know the total number of pixels classified to each cover type within the analysis district.

12) Calculate the mean number of pixels per segment classified to crop Z in the analysis district

Use this population mean in the the regression equation to estimate Y_D the farmer reported acres of crop Z in the analysis district.

Result: Estimate of farmer reported acres in the strata where sufficient Landsat coverage and/or segment numbers are available to perform regression estimates.

13) Accumulate analysis district regression estimates and direct expansion estimates were not available.

Result: State level accumulation estimates.

The primary factor to note is that we used the Landsat data from the JES segments and JES ground data twice. The first time we used it to develop a maximum likelihood classifier, i.e. developed a trainer. The second time we used the datum we classified the JES Landsat pixels with our maximum likelihood classifier that was develop on the same data set, then we estimated regression using the JES reported acres and the classified Landsat data.

The literature on discriminate analysis contains a great deal of discussion on how to estimate misclassification. Estimates of misclassification on the same data set used to develop the classifier are generally overlay optimistic simply because the same data are used for testing and development (8), (9), and (10).

II. CLASSIFIER OVERFITTING

We define classifier overfitting as the natural tendency of a maximum likelihood classifier to preform "better" on the data used to develop a classifier in our case the JES segment Landsat data, than it does on the Landsat data not in the JES segments. Basically, we believe there are two ways this can happen, as follows:

1) The JES segments do not contain a complete enough set of signatures so the classifier preforms well on the Landsat data used to develop it but is not representative of the data in the rest of the Landsat scene.

2) We do a very "good" job of fitting the classifier to the JES segment Landsat data; however, the high performance of the classifier is not duplicated in the rest of the Landsat scene.

We want the "best" possible classification we can get so that our regression estimator will have a small variance. As our classification accuracy increases the r^2 of the regression of y or x usually increases. The variance of the regression estimator is directly related to r^2 and the variance of the direct expansion. The relationship is

We can easily see that if our classification accuracy for Landsat data inside the JES segments over states the overall accuracy of the scene classification the sample estimation of variance is under estimated. In fact several papers have pointed out this

fact. The 1975 Illinois on crop acreage estimate recognized the problem that the r^2 could be biased upward (11). A study by Hung at Iowa State actually quantifies the degree to which the variance is underestimated when the same segment Landsat data are used for classifier development and regression parameter estimation (12). An evaluation of large area crop estimation using Landsat and JES data concluded the current practice of evaluating the classifier and developing the regression on the same data set used to train the classifier can lead to optimistic preformance estimates. Holko in a California study found that in nearly all cases that when the correlation from the independent procedure was less, it was significantly less (13). Based on these studiers the r^2 for DCLC procedure produces a optimistic estimate of variance. A direct relationship exists between r^2 and b the regression slope for a given sample. The relationship is as follows:

While this does not necessarily imply that b is biased if r^2 is biased, nevertheless; the relationship does exist.

A large simulation study conducted by Lundgren at Lockheed (14) looked at the bias and variance of the DCLC regression estimates. JES segment data and the accompanying Landsat data were simulated using a simulator developed on actual Missouri segment and Landsat data . Three of seven covers had significant biases in the number of pixels classified to the cover and in the regression estimates compared to the actual cover. These covers were pasture, corn, and waste. An analysis of the regression slopes showed five of seven covers' sample slope was significantly different from the population slope.

The mean difference for all covers was positive, ie sample slope larger than population slope. Lundgren suggested that the sample slope tended to be larger than the population slope. Chhikara and Hudson have shown that the relationship between reported acres X and classified number of pixels can be described in terms of and the proportion of incorrectly classified among those classified pixels into class 1 and O. An examination of their equation:

Y =

suggest why the sample slope are generally larger than the population slope. The term (1

which is analogous to the slope will tend to be larger whenever the sum of the two error terms is smaller. Lundgren argues that since the classifier is trained and thus optimized, i.e. classifier overfitting, on the same sample segments used in computing the regression equation, the classification errors on those segment will be less than the other segment in the population.

These concerns of bias in the DCLC estimates due to classifier overfitting have led us to the 1985 Remote Sensing Classifier Study.

Problem Statement

The Objectives of the Remote Sensing Classifier Study are as follow:

1) Determine if classifier overfitting is present in the DCLC procedures.

2) If classifier overfitting is present is it causing a unacceptable bias of significant proportions with DCLC estimates.

3) If the DCLC estimates are biased due to classifier overfitting what procedures can be implemented to reduce or eliminate the bias.

Organization of the Study

The 1985 Remote Sensing Classifier Study was conducted in Missouri and Iowa. Figure 1 indicates the study area in each of the States. In each study area we had three replications of JES segments. Each replicate was an independent JES sample. Non-agricultural strata were not considered. Strata break-down for the samples were as follows:

STATE

Replication	Missouri Strata					Iowa Strata		
	10	20	30	35	Total	11	12	Total
A	8	17	37	6	68	14	20	34
B	12	18	29	7	66	15	18	33
C	12	16	32	2	62	13	17	30
Total	32	51	98	15	196	42	55	97

Replication A of the Classifier Study was the Operational JES, so the data was collected in the usual time frame of late May and first week in June. After the JES edit the normal DCLC field level edit was completed. Tracts that contained intentions field, small grains observed or refusals or are in the objective yield survey, are included in a follow up survey in late July or early August. During the follow-up survey enumerators visited the tracts in question to verify the crop covers in each field. If a discrepancy from the cover indicated on the JES was found the enumerator records the correct cover and the field is updated on DCLC field level records.

The Iowa and Missouri State Statistical Offices (SSO) collected the data for replication B and C in late July and early August about the same time as the corn and soybeans objective yield surveys. The schedule used to collect the data is included in appendix A. The primary data collected was field use, field acres, planted acres, and harvested acres. After data collected was completed the additional replicates were edited using the DCLC field level edit. No JES edit was conducted on these segments because the DCLC edit was sufficient for the purposes of the classifier study. Basically, the DCLC edit is a JES tract level edit of the field data, plus a field level edit of the data for consistency. During the field level edit a one to one comparison of the questionnaire and photo is done to insure consistency.

The Missouri SSO digitized the additional Missouri segments while the additoinal Iowa segments were handled by the Remote Sensing Applications Section Staff using the video digitization procedures.

The Landsat data acquired for the Iowa analysis district was from July 25, 1985. A multi-temporial data set from July 3, 1985, and September 5, 1985, was created for the Missouri analysis district. Procedures were identical in each State after selection of the Landsat data. The Remote Sensing Branches PEDITOR software was used in signature development and some parameter estimator.

Tables 2 and 3 are a summary of the data available from signature development in the Iowa and Missouri analysis districts. If at least 75 pixels were available a signature was developed for the cover.

Table 2. Number of Pixels for Signatures Development by Replication Classifier Study Missouri, 1985.

Cover	Rep-A	Rep-B	Rep-C
	PIXELS		
Alfalfa	36	157	90
Corn	1061	1485	1011
Farmstead	23	24	28
Waste Land	1020	839	424
Winterwheat	376	195	663
Sorghum	510	515	896
Permanent Pasture	2158	2438	2202
Oats	51	95	58
Barley	62		10
Soybeans	2992	3052	2719
Dense Woodland	2679	3244	3058
Other Crops	47		43
Other Hay	584	916	1125
Cropland Pasture	148	55	
Idle Cropland	204	249	428
Other	106	24	81

Table 2. Number of Pixels Available for Signatures Development by Replication Classifier Study Iowa, 1985.

Cover	Rep-A	Rep-B	Rep-C
Alfalfa	55	29	33
Corn	7664	7103	8031
Farmstead	93	109	117
Waste Land	180	141	408
Permanent Pasture	347	309	102
Oats	206	376	339
Soybeans	6107	6930	6099
Dense Woodland	23	118	76
Cropland Pasture	5	29	14
Idle Cropland	9	2	67

A trainer was developed from each replicate of segments so that we have an A-trainer, B-trainer, and a C-trainer for each analysis district. Then each replicate of segment was classified with each trainer. So each replicate ended up being classified with three different trainers. The following Table summarizes the classification of the replicates.

TRAINER

Replicate	A	B	C
A	AA	AB	AC
B	BA	BB	BC
C	CA	CB	CC

Two letter denotes each combination of replicate and trainer. The first letter represent the replicate, the second represent the trainer. So AA represent replicate A trainer A, BC represents replicate B trainer C, and so forth. The dependently trained data sets were replicates of segment that were classified with the trainer developed from the replicated being classified. So the dependents data sets were AA, BB, and CC. The independent data sets were replicate that classified with a trainer which was different from the replicated being classified. So the data sets with independent classification were AB, AC, BA, BC, CA, and CB.

We estimated regression parameter for corn over soybeans using each of the data sets which gave us nine sets of regression parameters for each crop. Three of the regression sets were from dependent data and six were from independent data. We refer to sets of regression parameters because regression were estimates for each strata. We also estimated a combined regression overall strata which is available in the PEDITOR software. The method for calculating the combined regression was given Cochran (15).

After the regression parameters were estimate we created the appropriate validation sets by estimating a y for each segment within each data set. So, for the A trainer regression parameter from AA, BA, and CA were used to estimate y for each segment in the data sets AA, BA, and CA. We also generated y's using the B trainer and C trainer regression parameter and data sets. The following table summarizes the validation data.

TRAINER

Replicate	A	B	C
A	Y_{AAA} Y_{ABA} Y_{ACA}	Y_{AAB} Y_{ABA} Y_{ABB}	Y_{AAC} Y_{ABC} Y_{ACC}
B	Y_{BAA} Y_{BBA} Y_{BCA}	Y_{BAB} Y_{BBB} Y_{BCB}	Y_{BAC} Y_{BBC} Y_{BCC}
C	Y_{CAA} Y_{CBA} Y_{CCA}	Y_{CAB} Y_{CBB} Y_{CBC}	Y_{CAC} Y_{CBC} Y_{CCC}

Each Y_{VRT} represent a set of y's for each segment from replicate V using regression parameters estimate from set R and classifier with the trainer from set T.

The entire analysis district was also classified with each of the three trainers from the replicates and X, the population mean number of pixels was calculated for each strata. Then population regression estimates for corn and soybeans were calculated for each X and associated regression parameter which use the same trainer. So we end up with 9 population estimates for each crop. There were three trainers and three estimates for each trainer.

Analysis

Corn and soybeans data from the 1985 Remote Sensing Classifier Study were analyzed as follows:

1) The mean x number of pixels per segment classified to the crop in segments used to develop the trainer was compared with the mean in segment not used (independently) to develop the trainer.

2) Regression equation slopes and intercepts were compared when the regression parameters were estimated using the same segments the trainer was developed on and when the regression parameters were estimated using segments data independent of the trainer.

3) Comparisons were made between segment regression estimate $\bar{y}$ when the regression parameters were estimated from the same set of segments used to develop the trainer and when the regression parameters were estimated from a set of segment independent of the trainer.

4) Comparisons were made between populations regression estimates, $\bar{Y}$ when the regression parameters were estimated from the same set of segment used to develop the trainer and when the regression parameters were estimated from a set of segment independent of the trainer.

5) The population mean, $\bar{X}$ number of pixels per segment classified to the crop was compared to the sample mean $\bar{x}$ when the same sample was used to develop the trainers and when the sample was independent of the trainer.

The first item we looked at was the mean acres of corn and soybeans reported per segment. Since the mean number of pixels classified to corn or soybeans should be related to the amount of corn or soybeans in the segment summarized in Table 4. Replicate A represents the JES. The means appear to be fairly variable from one replicate to the next; however, none of the mean for a crop were significantly different across replicates.

Table 4. Comparison of the Mean Reported Acres of Soybeans and Corn for each Replicate 1985 Remote Sensing Classifier Study Missouri and Iowa.

REPORTED ACRES

MISSOURI

	CORN		SOYBEANS	
REPLICATE	MEAN	STD. ERROR OF THE MEAN	MEAN	STD. ERROR OF THE MEAN
A (JES)	27.7	5.4	72.6	9.1
B	37.4	5.6	62.5	9.2
C	22.8	5.6	66.8	9.5

IOWA

	CORN		SOYBEANS	
REPLICATE	MEAN	STD. ERROR OF THE MEAN	MEAN	STD. ERROR OF THE MEAN
A (JES)	299.8	12.6	233.0	14.2
B	273.6	12.8	258.5	14.4
C	300.6	13.1	236.6	15.1

Since the data were so variable from our replicate to the next we used analysis of Covariance most of our analysis. We used reported acres as our concomtant variable. Analysis of Covariance served two useful purposes error control and to adjust treatment means.

The mean number of pixel classified to corn and to soybean for each replicate in each state were analyzed using a separate analysis of covariance for each trainer. Details of the analysis were summarized in (appendix 2a.)

The covariate was the reported number of acres of the crop of interest for leach segment. The adjusted mean number of pixels classified to each cover were summarized in Tables 5 for Missouri and Table 6 for Iowa. The adjusted mean refer to the means after they were adjusted for the covariate, reported acres. The means with the dashed underlining were the mean from data set receiving a dependent classification ie. the same data classified was also used to develop the trainer. The primary comparison of interest in Tables 5 and 6 were mean within a trainer of the dependent replicate say the analysis of covariance was only applicable at the adjusted mean in most of these acres because slopes of the relationship between classified pixels varied from one source to the next. We will concentrate on these heterogenous slopes later when we look at regression slopes.

Table 5. Summary of the Adjusted Mean Number of Pixels Per Segment Classified to Soybeans and to Corn, Remote Sensing Classifier Study, Missouri, 1985.

	Trainer					
	A		B		C	
REPLICATE	MEAN	STD. ERR	MEAN	STD. ERR	MEAN	STD. ERR
CROP SOYBEANS						
A	92.0	5.2	85.9	5.2	84.2	5.2
B	89.5	5.3	84.9	5.3	81.4	5.3
C	90.3	5.2	87.7	5.2	87.7	5.2
GRAND MEAN	90.6	3.0	86.2	3.0	84.7	3.1
CROP CORN						
A	41.3	2.7	41.7	3.1	35.3*	2.9
B	37.8	2.8	38.2	3.2	32.5**	3.0
C	49.5*	2.8	46.4	3.2	46.0	3.0
GRAND MEAN	42.9	1.6	42.1	1.8	37.9	1.7

**Means for the independent replicate significantly different from the dependent replicate which received the same trainer. Significance level 0.01.

*Means different at the 0.05 level of significance.

Table 6. Summary of the Adjusted Mean Number of Pixels Per Segment Classified to Soybeans and to Corn, Remote Sensing Classifier Study, Iowa, 1985.

	Trainer					
	A		B		C	
REPLICATE	MEAN	STD. ERR	MEAN	STD. ERR	MEAN	STD. ERR
CROP SOYBEANS						
A	294.3	9.8	252.7	9.2	239.5	7.7
B	314	10	271.6	9.4	259	7.9
C	322	10.3	276.8	9.7	258.7	8.2
GRAND MEAN	309.6	5.8	266.6	5.4	252.1	4.6
CROP CORN						
A	357.9	8.8	414.2	8.8	417.6*	8.9
B	354.7	9.2	410.3	9.3	409.2	9.3
C	340.6	9.3	399.4	9.4	391.2	9.5
GRAND MEAN	351.4	5.3	408.3	5.3	406.6	5.3

**Means for the independent replicate were significantly different from the dependent replicate which received the same trainer. Significance level 0.01.

*Means different at the 0.05 level of significance.

Replicate A receiving the A Trainer versus the means of the independent replicate is replicate B and replicate C when the A Trainer was used. No significant differences within team between dependent and independent were in the Missouri soybean data. In the Missouri corn data three independent mean were significantly different from the dependent mean which received the same trainer. Each significant comparison involved the C replicate. With the C replicate having a larger mean in each comparison so the Landsat data for the C replicate behaved differently from the A&B replicates no matter what classifier was used. In this case, the difference in the classified number of pixels appear to be dependent on the data set being classified and not on the classifier being used. If the C replicate was the sample chosen in the operational JES, the classification of the C replicate would have provided a mean number of pixels classified to corn that was significantly different from a data set that was classified with either a dependent or independent trainer. In this case, the differences were due to the pixel reflectances in the C replicate and not whether the classifier was independent or dependent. In the Iowa data presented in table 6, there were no significant differences from one replicate to the next with-in trainer. We can find the same pattern for replicate A that was present in the Missouri corn data for replicate C. The only difference was that the adjusted means for replicate A were less than the replicate B&C adjusted means when the replicates were classified with the same trainer. Again, this points to a difference due to the pixel reflectance in replicate A. The Iowa corn data did not have as many significant differences as the Missouri corn data or the Iowa soybeans data; however, replicate A was always involved when there were significant differences. In an inspection of the Iowa corn data, we found that the A replicate had the larger mean within a trainer.

As far as classifier over fitting was concerned we would expect the means for the data sets which were classified with a dependent trainer i.e., replicate A classified with trainer A to be different, most likely larger, from the means for data sets which were classified

with an independent trainer i.e., replicate B or C classified with trainer A. No evidence of classifier overfitting was detected in the analysis of covariance for the adjusted mean number of pixels classified to corn or to soybeans.

In addition to looking at the adjusted means within a trainer, we also compared the mean number of pixels classified to the crop of interest across trainers. Our comparisons were done using repeated measure analysis of variance with the number of pixels classified to the crop of interest by the A trainer, the B trainer, and the C trainer as the repeated measures over replicates A, B, and C. The data were analyzed to determine if the dependent classifications, i.e., trainer A classification of replicate A, were significantly different from the independent classification i.e., train A classification of replicate A and B. Details of the analysis were summarized in appendix 2b. The results of the analysis were summarized in table 7 for Missouri and in table 8 for Iowa. There were no significant differences between the dependent and independent classification in the Missouri soybean data.

Table 7. Summary of the Mean Number of Pixels Classified by a Dependent Trainer Compared with two Independent Trainers for Soybeans and for Corn in Missouri, 1985 Remote Sensing Classifier Study.

MEAN NUMBER OF PIXELS

CROP SOYBEANS

REPLICATE	DEPENDENT (TRAINER) dMEAN (STD-ERR)	INDEPENDENT (TRAINER) iMEAN1 (STD-ERR)	INDEPENDENT (TRAINER) iMEAN2 (STD-ERR)	DIFFERENCE dMEAN-iMEAN* (STD-ERR)	DIFFERENCE iMEAN1-iMEAN2 (STD-ERR)
	(A)	(B)	(C)		
A					
	(B)	(A)	(C)		
B					
	(C)	(A)	(B)		
C					
GRAND MEAN					
MEAN					

CROP CORN

REPLICATE	DEPENDENT	INDEPENDENT		DIFFERENCE
A				
	(A)	(B)	(C)	
B				
	(B)	(A)	(C)	
C				
	(C)	(A)	(B)	
GRAND MEAN				

* iMEAN = (iMEAN1 + iMEAN2)/2

Table 8. Summary of the Mean Number of Pixels Classified by a Dependent Trainer Compared with two Independent Trainers for Soybeans and for Corn in Iowa, 1985 Remote Sensing Classifier Study.

In the Missouri corn data, there were two significant differences between dependent and independent mean number of pixels classified. Both of these comparisons involve the C trainer but only when the C replicate was not being classified. So the C trainer was performing differently on the independent replicates A&B than the A&B trainers; however, the mean from the C trainer was significantly different from the other independent as well as the dependent mean. In one case, the B replicate the dependent trainer mean was significantly different from the mean of the 2 independent means; however, this was caused primarily by the C trainer classification mean. The effect of the C trainer classification was also seen when the grand dependent mean was compared with the independent mean. Again, the primary reason for this difference was the effect of the C trainer. This points to a difference due to different trainer not based on dependent-independent relationship data. The grand mean comparison of the dependent mean and the mean of the other two independent mean test, whether their is a bias due to the dependent-independent relationship of the trainer and classified data set.

In both the Missouri soybean and corn data the dependent means were not significantly different. The sensitivity ie. minimal detectable difference for detectable difference for the soybean data was approximately 5 pixels in 89 while the sensitivity for the corn data was approximately three pixels in 42.

In the Iowa soybeans data summarized in Table 8 all the comparisons between dependent and independent means which showed significant differences due to different trainer; however, the difference between the dependent mean and the average of the two dependent means for the grand mean was not significantly different. This comparison was the only one which compared dependent and independent mean

comparison of the dependent mean and the mean of the other 2 independent mean test the whether their is a bias due to the dependent-independent relationship of the trainer and mean number of pixels classified to the crop of interest classified data set. In both the Missouri soybean and corn data the dependent means were not significantly different. The sensitivity i.e., minimal detectable difference for the soybean data was approximately 5 pixels in 89 while the sensitivity for the corn data were approximately three pixels in 42.

In the Iowa soybeans data summarized in Table 8 the comparisons between dependent and independent means which showed significant differences were a result of differences due to different trainer; however, the difference between the dependent mean and the average of the two independent means for the grand mean was not significantly different. This comparison was the only one which did not have a trainer effect in it. The Iowa corn data were summarized in Table 8 and also many of the comparisons were significant, however these difference were due to trainer effects and whether the trainer and classified data were independent from each other. The comparison between the B trainer classification and the C trainer classification were never significant different from each other. No matter whether the B trainer was dependent and the C trainer was independent or the B trainer dependent and C trainer was independent or when both the B trainer and the C trainer were independent. On the other hand every comparison between the A trainer and B trainer showed a significant difference. The A trainer and the B trainer were also significantly different from each other. The grand mean comparisons between dependent and the independent which were significantly different from each other were also the results of the between trainer. The grand mean comparison between the dependent mean and the average of the tow independent means did not show a significant difference; furthermore this comparison was balanced, that is it represented or contained an equal number of trainer A - trainer B, trainer A - trainer C and trainer B trainer C differences with each trainer playing the role of dependent trainer and of independent

trainer an equal number of times.

Based on the Missouri and Iowa data sets the estimated mean number of pixels classified to the crop of interest does not appear to be a bias due to the dependence of the trainer and the classified data set. The one noticeable point in the comparison between the two data set was that there was less difference between the mean number of pixels classified to the crop of interest from trainer to trainer in the Missouri data set than in the Iowa data set. We speculate that the reason for this difference may have been one or more of the following:

A) The time of acquisition of the data set was different so crop development and atmoshpere condition were less uniform in the Iowa scene than in the Missouri scene

B) The Missouri trainers were developed from approximately 60 half square mile segments while the Iowa trainer used 30 square mile segments. The Missouri trainer was made up of more independent bundals of pixels representing more diverse signatures than the Iowa data set.

C) The multi-temporal 8 channel data set used in Missouri provided a more consistent classification than the uni-temporal 4-channel data did in Iowa.

While we have not analyzed any data to determine what could be causing these differences, the multi-temporal versus uni-temporal appears to be the most plausable explanation.

Regression Parameters

We examined regression parameters used to make area regression estimates. The comparisons we made were between parameter estimates from dependent data set is the data was classified with a trainer developed on the same data set, and parameter estimates from an independent data set. The independent data set was to represent the population relationship between reported acres and the number of pixels classified to the crop of interest. If the relationship between reported acres and the number of pixels classified to the crop of interest for the dependent data set was not significantly different from an identical relationship for the independent data set then there would be no bias attributable to the estimate regression parameters.

The regression parameter estimates were summarized Tables 9, 10, 11, and 12. The Missouri soybean data was the only data set which showed a significant difference between the parameter estimate on the dependent data set and on the independent data set. We contend that the Missouri soybean data represented the true relationship of the parameters estimated from the sample and the population relationship for classified number of pixels and reported acres. The slope estimates were analyzed based on the R^2 value for the regressions the Missouri soybean classifications explained the variation in the reported acres better than the Missouri corn or the Iowa corn or soybeans relationship. When we compared the Missouri corn and Iowa corn and soybean regression relationships from replicate to replicate Tables 10, 11, and 12 we found that the data sets themselves dominated the relationships. For example the Missouri corn regression slopes for strata 10 the C replicate always had the largest slop. So the replicate we used had more effect on the regression slope than whether the data set used to estimate the regression parameters was dependent or independent.

The slope estimate were analyzed using the signed-rank or parameter test. A summary of the analysis was presented in (Appendix 2c). The results showed that the estimated

slope from the dependent data was larger than the independent slope in 16 of 18 comparisons. Also we found that the intercept for the dependent data set regression was smaller than the intercept for the independent data in 15 of 18 comparisons.

The large slope would be supported by previous work which showed that the dependent data set regression had a large R^2 value than exited in the population. Since we know

$B^2 = R^2 \frac{S_Y^2}{S_X^2}$ the larger slope result

seems to be a reasonable conclusion.

Figure 5 summarizes the difference between the dependent and independent data sets relationship of reported acres to classified number of pixels. While these relationships were different and an important fact to remember is the mean number of pixels and the mean reported is a point on the regression line and earlier we showed that the mean number of pixels classified to the crop of interest was not different between the dependent data and the independent data sets. We also showed that the reported number of acres was not different from replicate to replicate. So the intersection of the dependent and the independent regression relationship was usually close to the mean number of pixels classified and the mean number of acres reported, if the slope regression was very different. In any event if the sample estimate of the mean number of pixels classified to the crop of interest was a "good" estimate of the population mean number of pixels the little effect would be seen from the dependent regression line as opposed to the independent regression line. If the sample estimate of the number pixels classified to the crop of interest was not "close" to the population mean than the difference between the acres regression estimate from the dependent sample and the true relationship could become larger as the sample estimate departs from the population number of pixels. Using fig-1 we can see the difference more clearly. The D line was the dependent relationship with the point

An alternative to occassionaly used with area regression estimates is to estimate combined regression parameters rather than separate regression parameter for each strata. Details for this procedures were outline by Cochran in Sampling Techniques. Basically the separate estimate has a smaller variance than the combined unless the regression slope are the same a cross the combined strata.

The combined regression slope estimates were summarized in table 13 and 14. The result did not change when be analyzed the combined regression. The Missouri soybean data indicated a significant difference between the dependent regression estimates and the independent regression estimates. The dependent slope estimate was always less than the independent, which was similar to the result found for separate regression slope estimates for each strata. We were then led to the same conclusion concerning the effect of using dependent regression slope estimates. For the combined regression slope estimates the combined strata means

T