REGIONAL VARIATION AND CROP SEPARABILITY IN A THEMATIC MAPPER BASED CROP INVENTORY OF NEW YORK STATE *

S.W. Buechel, W.R. Philipson, and W.D. Philpot

CLEARS, Cornell University,Hollister Hall
Ithaca, New York, 14853, U.S.A.

1987

ABSTRACT

Image stratification has long been considered a means of improving automated classification results. Many stratification criteria have been applied in practice, but the absolute effect of specific factors has received less attention. The impact of regionally definable parameters on Landsat Thematic Mapper spectral characteristics of crops is examined through statistical analysis of thirty sites across New York State. A normalized vegetation index (NVI), transformed divergence, and the accuracies associated with an unsupervised classification are tested as dependent variables against growing degree days, rainfall, topography, crop diversity, field size, soil drainage, frost-free season length, recent rainfall, and general yield characteristics. Linear relationships are found to be significant, and multiple regression is used to define the optimum set of criteria. The factors selected are shown to be dependent on crop type. Average frost-free season length was the most significant predictor of variation in classification accuracies. Growing degree days- base 40 and topography were the best predictors of variation for NVI.

1. INTRODUCTION

Regional variation has long been a nemesis of large-scale automated inventory involving satellite data. When significant target variability can be associated with land cover characteristics, stratification, either pre or post classification, has been one solution. Studies on the effectiveness of stratification in compensating for target variance effects in image classification have focused on Landsat multispectral scanner (MSS) data. Rohde (1978) reported that for the classification of inundated agricultural land, accuracies increased 5-10% with the use of three strata based on land cover types. Stratification was also used post-classification to correct for spectral overlap in natural cover types when it was associated with specific landforms (Rohde, 1978; Pettinger, 1980).

* Presented at the Twenty-First International Symposium on Remote Sensing of Environment, Ann Arbor, Michigan, October 26-30, 1987

Work with stratification in crop inventory suggests more refined delineations. Draeger et al. (1973) described improvement in classification accuracy, as well as inventory cost using stratification associated with field size, the relative proportion of crop types, field conditions and the proportion of non-agricultural land as divisional criteria. The USDA National Agricultural Statistics Service has incorporated stratification in their crop inventory methodology (e.g., May et al., 1983), using strata based on percent land in agriculture and, for some states, crop composition. Stoner (1983) found that single date classification on data stratified by soil drainage, topography and land use resulted in accuracies equivalent to non-stratified multidate classification for corn, soybeans, tobaccco and forest in North Carolina. Hay et al. (1982) described an automated stratification technique based on variation in crop temporal characteristics. These studies illustrate the potential usefulness of stratification, and some of the criteria which have proved effective for delineating strata. None, however, consider the relative importance of specific regional factors as the basis for stratification.

The objective of this study is to associate specific regional agronomic characteristics with variation in crop separability across New York State. A similar study was undertaken by Batista et al. (1985) which considered the effects of specifically defined scene characteristics on supervised MSS classification accuracies of corn and soybeans in the Midwest. The criteria found significant included field size, proportion of crop in the scene, crop diversity, soil order, soil drainage class, percent slope, long term average yield, maximum yield, position within the Midwest Corn Belt, weather, and crop development stage. The focus of this study differs in addressing early season crop inventory using Landsat-Thematic Mapper (TM) data in a more complex environment.

2. STUDY DESIGN

Thirty study sites were selected across New York State based on the range of selected environmental parameters and TM image availability (Figure 1). Image coverage required four Thematic Mapper scenes. The best coverage was obtained for the latter half of June (17 June 1984, 22 June 1984, 14 June 1986, 25 June 1986) and only early season identification is considered here. Each study site defined an area of approximately 195 sq km. Prior to analysis, a "dark object" subtraction atmospheric correction was applied to each subscene. Digital counts from water bodies in each subscene provided the baseline for the correction.

Crop cover data were obtained from the USDA Agricultural Stabilization and Conservation Service (ASCS) historical records. Field locations were transferred to aerial photographs and manually digitized to the screen-displayed TM subscenes. Five crops were considered [illegible] cut-hay, oats, wheat, and pasture.

Ten environmental factors, historically implicated in target variance, were selected to se[illegible] independent variables:

Topography (TP) - variation is represented as the frequency of change and degree of slope on a scale of 1 to 4, from flat(1) to steep(4).

Yield group (YD) - the seven levels reflect the expected yield characteristics of the site as defined by the New York State Cooperative Extension soil management guidelines; from low(1) to high(7).

Seasonal Rainfall (RN) - seasonal inches of rainfall as recorded by the closest representative weather station up to the date of image acquisition.

Wetness (WT) - the inches of rainfall recorded for the two days prior to and including the image acquisition date.

Soil Drainage (SD) - generalized soil drainage characteristics derived from the New York State soil associations map (Cline and Marshall, 1978). The data were digitized and overlaid on the site locations; from poorly drained (1) to well drained (4)

Crop Diversity (DV) - the number of different crops in the data set for a site.

Field size (FZ) - the number of pixels defining a field or the average number of pixels per field for a crop type, depending on context.

Growing degree days, base 40 and base 50 (G4,G5)-represent the amount of time for potential photosynthetic activity elapsed to a given date(as a function of temperature). It is defined here as:

$$\frac{\text{daily maximum temperature} + \text{daily minimum temperature}}{2} - \text{base X} = \text{GDD}_{\text{base X}}$$

Base 50 data are routinely recorded by NYS weather reports. Base 40 data was calculated from historical weather data tapes (the base number refers to degrees Fahrenheit).

Frost-free season (FFS) - the historical average length in days

Excepting field size, each criterion is an average across the study site. Other field specific information was not available.

Three measures associated with crop separability were used to evaluate the relative importance of the ten environmental variables:

1. NVI -the variation in NVI (band 4-band3/band4+3) as a measure of the quantity and consistency of vegetative cover, an important criterion influencing the spectral variability of a crop.

2. DIVERGENCE-the variation in the optimized divergences (Swain and King, 1973) between crop digital count distributions as a direct measure of inherent spectral separability

3. CLASSIFICATION-a test of crop separability as expressed in the results of an unsupervised classification

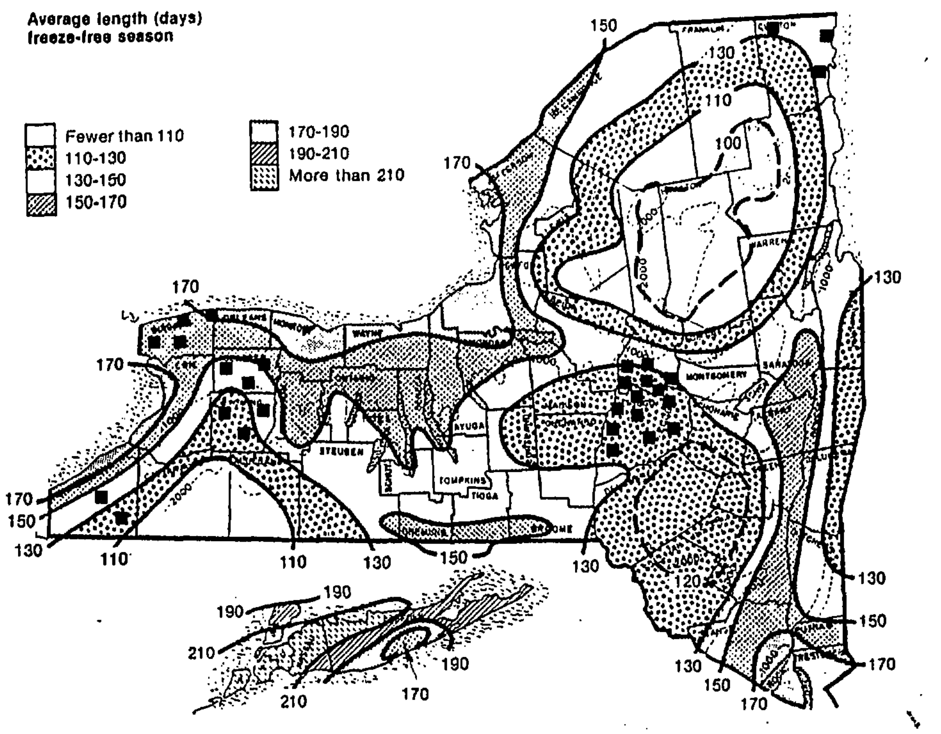

FIGURE 1. Frost-free season length variation in New York State(Division of Atmospheric Sciences, Cornell University) and study site locations

4. DATA ANALYSIS and RESULTS

4.1 NVI

The Normalized Vegetation Index was calculated for each field in the data set. The data set of 1863 fields was divided in two. One half was used to determine the best linear regression among the variables, the other as a test of the results. The SAS RSQUARE (SAS,1985) procedure was used to test every possible combination of variables. The R^2 and Mallow's Cp (Wood,1973) values were used to select an optimum set of parameters The variables found to account for maximum variation in the NVI of the specific crops, while minimizing Cp (a measure of the significance of the contribution of additional parameters) appear in Table 1.

TABLE 1. REGRESSION RESULTS - NORMALIZED VEGETATION INDEX

Crop	Variables **	R2	Cp	N*
CORN	FES TP G5 SD	.736	2.2	552
CUTHAY	FZ RN SD G5 WT	.930	9.6	63
HAY	G4 YD RN WT	.913	2.7	112
OATS .	RN YD FZ G4	.495	.46	82
PASTURE	RN G5 WT TP	.923	.36	32
WHEAT	G4 FZ RN TP	.890	.46	92

*number of observations

** DIVERSITY (DV) was not included in this data set

__ the variable contributed a large proportion of the R2

To find the best fit for all crops, a regression was run under the SAS GLM procedure. G4, RN, FZ, G5, WT and TP were selected for inclusion as the initial predictors yielding an R2 = .734 with a high probability of significance (p=.0001). Variables were deleted based on the lack of significance of the partial F tests. The remaining variables were tested for the significance of their relationship as a function of crop type. The final best fit was obtained with G4 (as a function of crop type) and TP explaining 73.8 percent of the variation in the data (p=.0001). The model was tested on the other half of the data to determine the extent to which the model building procedure influenced the results. An R2 = .726 was obtained (p=.0001) suggesting that the effectiveness of G4(crop) and TP in explaining over 70 percent of the variability in NVI for all crops is not unique to the inital data set.

The inability of the independent variables to explain a significant portion of the oat field variability in NVI (Table 1) may be due to the inclusion in the data set of both oats for hay and oats for grain. In late June, oats in New York State should be mature and heading if planted for grain, but potentially a wide variety of states if some fields have also been cut for hay. Factors relating to this source of variation were not model options.

4.2 Divergence

Transformed divergences were calculated to test the variability in the spectral distributions of the crops as an indicator of their potential optimized separability. TM bands 1-5 and 7 were used in the analysis. The pixel population for each crop type was clustered into a maximum of five clusters using the ERDAS modified version of the NASA/JSC ASTEP clustering algorithm (ERDAS,1987). The statistics associated with the clusters were used to compute the divergence between each crop and the cluster groupings of the other crops within a site. Each divergence value was weighted according to the relative importance of the cluster. The weighted average transformed divergence served as the dependent variable in the analysis. The best fit for the variables, based on the highest R2 with the lowest Cp are shown in Table 2.

TABLE 2. REGRESSION RESULTS - TRANSFORMED DIVERGENCE

Crop	Variables	R2	Cp	N
CORN	RN YD WT G5 FFS	.431	1.89	31
HAY	YD DV WT FFS TP	.251	1.45	24
OATS	TP YD SD G4 G5	.480	3.90	21
PASTURE	G4 DV RN TP G5 WT	.830	3.28	15
WHEAT	DV YD G5 G4 SD RN FFS TP FZ	.901	18.9	13

__ the variable contributed a large proportion of the R2

The best result for regression over all crops included FFS, SD, and RN ($R2=.215$) The F test for fit was not significant.

4.3 Unsupervised Classification

An unsupervised classification (ERDAS "Clustr" program, ERDAS,1987) was run on each of the 30 sites using the six non-thermal TM bands. Pixels from the entire data set for each site were used in defining the clusters. A limit of five clusters per crop type was set. An example of the variation in accuracies found for a single crop appears in Figure 2. The crop accuracies (omission error only) for each individual cover class served as the dependent variable in the regression. The "best fit" results are listed in Table 3.

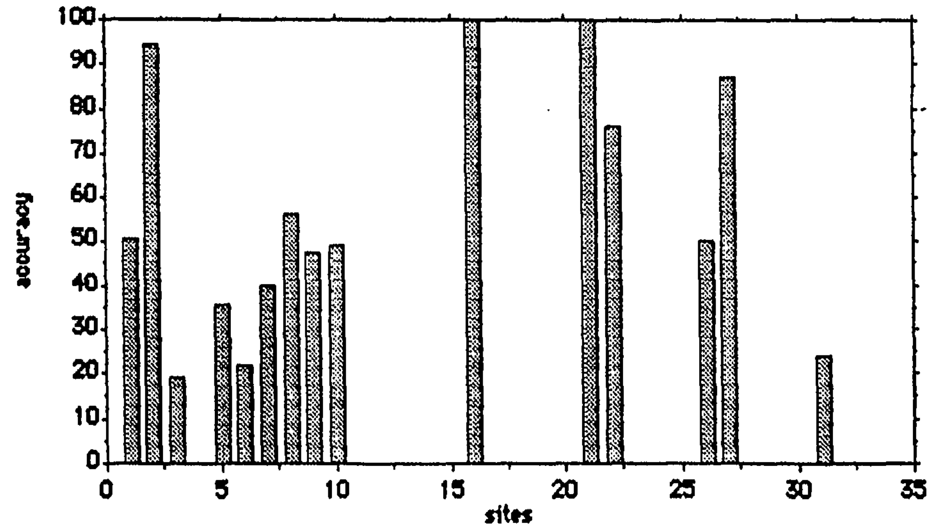

FIGURE 2 Variation in classification accuracies for cuthay

TABLE 3. REGRESSION RESULTS - UNSUPERVISED CLASSIFICATION ACCURACIES

Crop	Variables	R2	Cp	N
CORN	SD FFS DV TP WT	.88	5.89	30
CUTHAY	G4 WT DV FFS FZ	.67	4.09	17
HAY	G5 G4 FZ TP WT	87	.51	23
OATS	FFS DV G5 TP FZ SD YD	.85	7.09	26
PASTURE	FFS WT SD	.85	.57	15
WHEAT	FFS YD RN SD G4 G5 TP DV	.99	8.20	15

__ the variable contributed a large proportion of the R2

Due to very high individual correlations, the first variable in every case contributed the largest amount to the set. When the crops were run as one data set the best fit included FFS, DV, WT and FZ yielding an R2 of .753 (p=.0001). Clearly, differences in average frost free season length account for considerable variation in accuracies among the sites.

5. DISCUSSION

Crop type was consistently found to be an important factor in describing the relationship between crop separability and the environmental variables tested. This result confirms the finding of Batista et al. (1985) using MSS data. The crops which were individually less well described can be associated with sources of variation not accounted for by the chosen criteria. Pasture and wheat, neither of which were subject to such sources were the best fit of the individual crops across all measures. In addition to varying between crop types, the variables selected were not consistent by crop across the three measures tested. This suggests that optimal stratification will vary by crop as well as the classification strategy employed.

Regression on entire data sets did yield results which may be useful for multi-crop inventory. The variation in NVI was described with an R2 of .738 including growing degree days (base 40) and topography. The variation in unsupervised classification accuracy was best predicted (R2=.753) as a function of frost-free season length, crop diversity, amount of recent rainfall, and field size.

A significant linear model for divergence over all crops was not found. The optimized clustering appears to enhance the problem attributed to an inadequate range of independent variables for describing the variation in corn, oats, and hay. When clusters are defined based on the entire data set, as in the classification procedure, accuracy is adequately described in a linear relationship with a subset of the criteria. The average frost-free season length was the most significant single descriptor.

This study described the role of regionally generalized parameters in describing variation in crop separability using a generic classification strategy and the full spectral dimensions of the non-thermal TM data in an environment where classification accuracies for a crop may vary from 0 to 100 percent (Figure 2). It was determined that significant linear relationships could be found. Tentative best descriptor relationships are presented. Further testing is required to develop the models and determine the appropriate gradation of factors for stratification. The effects of more specifically defined criteria, as well as the impact of an operationally optimized classification procedure must be considered.

REFERENCES

Batista, G.T., M.M. Hixson, M.E. Bauer (1985) " Landsat MSS crop classification performance as a function of scene characteristics," **International Journal of Remote Sensing** 6(9): 1521-1533.

Draeger, W., J. Nichols, A. Benson, D. Larrabee, W. Senkus, C. Hay (1973). "Regional agriculture surveys using ERTS-1 data," **3rd Earth Resources Technology Satellite-1 Symposium**, v.I: 117-125.

ERDAS,Inc. (1987) "Image Processing System User's Guide," **ERDAS User's Guide**, ERDAS, Inc., Atlanta , Ga.

Hay,C., L. Beck, E. Sheffner (1982) "Use of vegetation indicators for crop group stratification and efficient fullframe analysis," **Proceedings of the International symposium on Remote Sensing of Environment** v.II:737-745.

Pettinger, L.R. (1980) " Environmental stratification - a method to improve Landsat digital analysis accuracy and land cover map utility," **Proceedings of the 14th International symposium on Remote Sensing of the Environment** v.III: 1587-1599.

Rohde, W.G. (1978) " Improving land cover classification by image stratification of landsat data," **12th International Symposium on Remote Sensing of the Environment** v.I: 729-741.

SAS Institute, Inc (1985). **SAS User's Guide - Statistics Version 5 Edition**, SAS Institute, Inc., Cary N.C.

Stoner, E.R. (1983) "Stratification of sampled land cover by soils for Landsat-based estimation and mapping," **AgRISTARS Technical Report Nasa/NSTL/ERL-224**.

Swain, P.H. and R.C. King (1973) "Two effective feature selection criteria for multispectral remote sensing," **Proceedings of the 1st Joint Conference of Pattern Recognition**, IEEE, : 536-540.

Wood, F.S., (1973) "The use of individual effects and residuals in fitting equations to data," **Technometrics** 15(4):677-695.