Sat Ag Srvy 1973

INVESTIGATION OF THE FEASIBILITY OF MAKING GLOBAL WHEAT PRODUCTION ESTIMATES USING REMOTE SENSING DATA

by

Center for Remote Sensing Research
School of Forestry and Conservation
University of California
Berkeley, California 94720

1973

September, 1973

INTRODUCTION

Based on a request by the National Aeronautics and Space Administration, personnel of the Center for Remote Sensing Research, University of California, have compiled a series of preliminary discussion papers dealing with the subject of techniques for making worldwide estimates of annual wheat production, with emphasis on the possible role of remotely sensed data in the estimation procedure.

The material contained in this series of papers is not intended to constitute an exhaustive or complete discussion of the subject, but rather consists of information and thoughts on the subject resulting from a few days of effort on the part of those involved. Hopefully, however, these papers can provide a useful base for the much more thorough and detailed studies which might follow. In some cases, the information contained in the papers is based on extensive research conducted in the past by the Center for Remote Sensing, while in others it has been acquired through a rapid library search and may be quite speculative in nature.

The techniques discussed have been developed with assumption that only information in the public domain, i.e., excluding that acquired by clandestine means, would be available. Remote sensing data sources considered included those non-military satellites expected to be operational in the fall of 1974.

In general, our goal has been to ascertain what parameters might be useful in predicting wheat production at any point in time during the annual cycle, to ascertain the relative importance of each parameter, and to speculate on ways in which these parameters might be estimated or measured. In addition, it is felt that the variation in different geographic areas and times during the crop growth cycle of the actual list of parameters and their importance may be quite significant.

Specifically, the topics discussed in the attached papers consist of the following:

I. A discussion of the general procedure that might be used in an operational inventory.

II. Statistical analyses of the relationships between a number of pertinent parameters including acreage, yield, and production.

III. A discussion of the estimation of crop acreage using remote sensing data, based primarily on past work conducted at the CRSR.

IV. Possibilities for developing relatively direct models to estimate crop yield using primarily visual (i.e., spectral and textural) inputs relating to the appearance of the crop at various times.

V. More sophisticated continuous and discrete function production estimation models using a variety of environmental parameters in addition to those in (II) above.

VI. Multistage sampling techniques to be used in support of large scale regional or worldwide crop inventories.

GENERAL RECOMMENDATIONS

The work completed to date has indicated two general directions in which future efforts should be directed to answer many of the questions raised during this brief study. First of all, a very thorough search should be made of sources of ancillary data available regarding both past and current wheat production in the study areas which might be used to supplement remote sensing techniques. It seems certain that particularly within the U.S. Department of Agriculture's Statistical Reporting Service, Economic Research Service, and Foreign Agricultural Service, information and knowledgeable persons exist who would be invaluable in a survey such as discussed here. Certainly it would be grossly inefficient to bypass these sources. Secondly, of course, a great deal of work would have to be done to test and verify the predictive models discussed and to accurately ascertain expected costs of acquiring and processing the necessary data before any informed decisions could be made regarding the actual adoption of any particular system on an operational basis.

A rapid, cursory study such as this raises more questions than are answered. It is our hope therefore that at least some of the more important and meaningful questions have been asked.

The first question that should be answered is "How good have estimates been in the past?"

The Department of Agriculture predicts wheat yield for both spring and winter wheat. The predictions for winter wheat begin with a pre-December prediction and then in April, monthly updates are given through October. For spring wheat, the first prediction is made in June and continues monthly through October.

The prediction for winter wheat yields on the average are more reliable than the prediction for spring wheat. This is probably due to the apparent variability in acreage planted to spring wheat from year to year. The production estimate for months prior to harvest were compared to the final after harvest for the period from 1956 to 1966. The average error for winter wheat predictions for these years beginning in December were 8.4 percent, for April, 6.3 percent, for May 5.4 percent, for June 5.4 percent, for July 3.0 percent, for August 1.4 percent, and for September 1.3 percent. It is apparent that as the growing season progresses, the estimates of yield get better and better. However, there are some rather wild excursions from these average predictions. In 1958, the December prediction was off by 23 percent; in 1960 it was off by 17 percent; 1962 was off by 15 percent and in 1963 it was off by 13 percent. The first predictions after the winter, made in April, were considerably better. However, in 1958 they were off by —18 percent; in 1960 they were off by —12 percent, and in 1962 they were off by +12 percent. By May, the worst case error was down to 11.2 percent; by June it was 9.4 percent, by July it was down to 4.5 percent; by August it was down to 2.3 percent, and by September it reached 2.6 percent.

The first estimate for spring wheat is made in June. The averages are considerably worse. In June, the average error was 13.3 percent, in July 9.9 percent, in August 6.9 percent, in September 2.4 percent, and October 1.6 percent. In June the excursions from the final value were as high as 40.5 percent in 1961, 28 percent in 1958, 23 percent in 1962, and down to a low of 12 percent in 1964. In July, the 1958 error was 24 percent; the 1962 error was 22 percent, and all other values were below 10 percent. By August the worst case error was down to 11.4 percent in 1958, and all other values were below 10 percent. By September the errors were all below 6.7 percent.

I. GENERAL PROCEDURE

The following outlines the procedures that would be required to use ERTS-1, ERTS-B, meteorology satellites, supporting aircraft data and ground data to complete a global inventory of a single crop or crop type such as wheat or cereal grains.

Step 1: Image of the Globe

First using ERTS-1 imagery now on file, a worldwide mosaic of all of the land surfaces of the world would be created. This mosaic should be generated by the latest in computer techniques where control points from the imagery would be related to existing map control points and satellite navigation data would be utilized. From the information obtained through the mathematical analysis of this control information and global information, all redundant "picture elements" would be removed from digital tapes and a precision mosaic would be generated and stored on tape. If ERTS-1 imagery does not exist cloud-free for the entire world, it could only be obtained from ERTS-B or other multi-spectral data sources such as Skylab, which may allow the filling in of holes in the digital mosaic. The major difficulty in accomplishing this task would be the extraction of the control points from the imagery. It would need to be done only once, but would have to be done for all areas where wheat is grown. If navigational information from the satellite is adequate, it may require only a few control points per orbit rather than control points from individual frames as is needed now. The major cost in such an operation is the computing of the transform after the transformation algorithm has been determined.

The mosaic could then be converted to current tape format. From these tapes, images could be produced for the next step which would be human photo interpretation.

Step 2: Photo Interpretation and Agricultural Area Delineation

From the computer compatible tapes generated in the mosaicing procedure in the previous step, color composite images would be produced for use in the photo interpretation stage. The resolution of ERTS data may be more than adequate. Possibly, every other or every third point and line of the image would be adequate to produce an image for delineation of agricultural versus non-agricultural lands in the wheat growing areas of the world. If every other point or every third point and line were adequate, it would mean that a 300 x 300 mile area could be delineated on one image in this gross agricultural-non-agricultural separation step. This separation could be done directly on the hard copy image. The image would be mounted on a digitizing system that would not interfere with the photo interpreter's speed and accuracy of delineation. A system similar to the graph pen currently in use at Center for Remote Sensing, or possibly, an extremely high resolution color CRT utilizing a light pen system could be used. The coordinate information and annotation of the various areas delineated by the human would be entered through a computer terminal by the interpreter. No intermediate step of hard copy preparation or line following by an operator after interpretation would be necessary. A portion of this separation may be done automatically. For example, it has been demonstrated that a single band of MSS data may be adequate to discriminate water, unproductive bare soil, snow packs, and desert areas from everything else. The photo interpreter would not be presented with the problem of delineating these features, reducing his total interpretation time. This pre-processing could be put at the front end of the system and eliminate further processing of this data during the inventory, reducing the cost of processing significantly. The human input would still be cost-effective in the separation of some wildland and urban areas from agriculture. This combined human and computer stage would significantly reduce the cost of further processing, and increase the accuracy of the discriminant analysis that would follow. It would be a one-time operation that would last for several years and would be updated as needed rather than being completely redone as new agricultural areas developed or old areas went to other users.

Step 3: Worldwide Stratification

The delineation produced by the photo interpreter and computer would be used in conjunction with geographic, climatic, and cultural information to stratify the world's agricultural land. These strata would be non-contiguous in that they would be worldwide in nature. For example, agricultural lands similar to those in Kansas would be identified in Chile, Australia, Europe, China, India, USSR, the United States, Brazil and all the other significant areas of the world. This strata would be based on cropping practices, soil types, local climate, geography, geology, distance from coast, distance from river systems, sources of water, and many other factors. This may be the most difficult and important step in the processing, in that it requires time to gather many sources of information to make decisions on stratification and it appears that much of the variation in wheat yield is explained statistically through stratification. This step is another one-time operation that would last for several years, requiring periodic updating as local conditions changed. Much of the information for this stratification already exists in terms of the climate and geography associated with various land masses and has been compiled by several investigators.

Step 4: Training the Classifier

For any given date or series of dates of imagery that would be used in the discriminant analysis, training sets would be developed for each of the worldwide strata as defined. This training could be done by conventional means. That is, extracting training fields from the raw imagery, or perhaps when possible from low altitude aircraft identification and spectral data acquisition at the same time. Experience has shown that in many cases it is necessary to identify and train on crops other than the one of direct interest in the inventory to give the discriminant analysis program alternatives to the crop of interest when the spectral signatures are similar. This may be true in some strata and not in others, but must be considered. By monitoring environmental conditions throughout a single strata, transformations may be applied to the training set to allow the extension of the spectral signature throughout the strata. Without this transformation in the spectral signature to correct for varying conditions throughout the stratum, very poor discrimination will result, unless weather conditions, soil conditions, planting dates, and other factors influencing crop maturity and spectral reflectance happen to be the same worldwide for a given stratum.

With the large amount of data that must be handled in this worldwide stratification, some automatic or semi-automatic means of processing the data must be utilized. Because of the nature of the data, a pattern recognition algorithm, similar to the one used for crop identification through spectral patterns, could be used. The variables mentioned above that are important to the stratification would be quantified in a unique, continuous fashion. These quantified parameters would be utilized as features in the pattern recognition algorithm. For example, minimum temperature, maximum temperature, mean annual temperature, mean annual rainfall, average pre-planting rainfall, average post-planting rainfall, distance from coast, historical yield per acre, variance of yield per acre, average field size, variance of field size, wheat to non-wheat ratio, historically, soil productivity, index of mechanization, and degrees latitude may all be variables that could be used as features for the pattern recognition process. To utilize this technique, areas where ground truth could be acquired would be clustered into homogeneous or as nearly homogeneous types as possible. These homogeneous areas or clusters would be used as the training set for the classifier and all other areas of the world would be matched by the pattern recognition algorithm to the training set. After the data is made available to the computer the computational task would be relatively simple and inexpensive.

Step 5: Crop Identification and Acreage Estimation from ERTS Multispectral Scanner Data

Within each of the strata as defined by the photo interpreter and computer earlier, the training generated in Step 4 would be used to classify the agricultural lands to separate wheat from all other crops. With this discriminant analysis, an estimate of the acreage of wheat would be obtained from ERTS data. This estimate would have errors associated with classification accuracy, boundary effect, field size, and crop condition. The analysis may be done on a point-by-point basis, or it may be done on a sampling scheme where every other point, every third point, every fifth, sixth, or tenth, point and line may be considered. A clustering or blocked sampling scheme may be more efficient than a point sampling scheme because field by field information that may be useful in ratio estimation procedures described is section VI. The sampling scheme will depend on the processing that would be required in the subsequent steps in the estimation of yield and corrections for acreage. A variable probability sampling scheme where the human quick look estimate of wheat acreage for very large contiguous areas would be used to determine the general location of the areas where samples for computer processing would be selected. The theoretical justification of this process is covered in section VI. Because of the errors associated with this estimate, it could not be the final stage in the area estimation procedure.

Step 6: Photo Ground Sampling

A sampling design as described in section VI of this report would be used to subsample selected areas within the computer processed areas accessible via aircraft and then, if possible, subsequent ground samples of these areas to determine the relationship between the ERTS acreage estimate, the photo ground acreage estimates, and the yield. The size of the photo gound sample would be determined by the accuracies obtained in the remote sensing stage or the ERTS stage of the estimation. The better job that can be done from the remote sensing platform, the fewer samples that will be needed on the ground. In some cases, no ground samples will be needed, only aircraft samples. When there is a positive consistant linear relationship between the stages in the estimation, very few samples at each stage will be needed. The results of the sampling procedure would give the correction to the acreage and yield information within each of the strata. After experience is gained, sampling may be done on an adaptive basis where samples obtained from previous inventories may be updated rather than being completely replaced when an updated production estimate is made. This sampling with partial replacement may be useful from year to year as well as from month to month for a single year.

Step 7: Estimation of Yield

Estimates of yield would be obtained directly in many cases by applying a weighting factor to the acreages estimated for each of the classes generated through the training and discriminant analysis in Steps 4, 5 and 6. In some cases, spectral differences will not be observed due to the actual yield and therefore, cannot be predicted by training by subclasses associated with yield, therefore auxiliary information would be introduced through the models similar to those described in section lV and V of this report to provide the yield figures. Auxiliary variables such as rainfall, indexes of environmental stress, or harvesting techniques would be used. The following section discusses the relative importance of some of these variables.

II. RELATIVE IMPORTANCE OF PARAMETERS AFFECTING WHEAT PRODUCTION

From the data analyzed and information from existing production models, it was determined that stratification of wheat producing lands by soil type, cultural practices and general meteorological conditions is the most significant step in reducing the variance of the estimate total wheat production of the U.S.

It is also the most importnat parameter in consistently estimating wheat yield per acre on a regional basis. After the stratification is completed it was found that wheat yield and production was a function of different parameters or combinations of parameters dependent on the stratum being studied.

To limit the computer effort but study this intra-stratum variability Arizona was selected to represent the irrigated wheat area, Kansas was selected to represent the winter wheat area, and North Dakota was selected to represent spring wheat area. Within each of the selected study areas acreage planted, acreage harvested, acreage abandoned, precipitation, and year were analyzed to determine the amount of variation in production attributable to each and the correlation between the parameters affecting production.

Arizona wheat production, comprised mainly of irrigated wheat, seems to be the simplest production estimation problem. In the past 5 years, because of development of new strains of wheat resistant to lodging, the wheat production of the area has increased drastically. In the period from 1965 to 1972, the production has varied some 82 percent. Using linear relationship between planted acreage and production, the error associated with the estimate of production is reduced to 3.5 percent with a worst case prediction of 5.2 percent. Using harvested acreage, an improvement of only .1 percent was achieved. No rainfall information or other climatic variables were considered in this analysis because of the small residual error after the variability due to acreage was removed. It was found through examination of data that reduced production has been due to low spring temperature and/or low average temperatures, which may indicate incident radiation conditions or temperatures below optimum for wheat production. The high correlation between planted wheat and production is probably due to the consistency of the environment and the fact that water is supplied through irrigation.

With U.S. winter wheat production as the dependent variable in the regression analysis and planted area as the independent variable, the error was reduced to 17.3 percent from the original 18.8 percent. Adding rainfall (from the previous harvest through April) reduced the variation to 10.8 percent. Using harvested area only, the error was reduced to 15.7 percent. The difference between planted and harvested or abandoned acreage was used as an independent variable reducing the error to 8.8 percent. With abandoned and harvested acreage and date used as independent variables, the variation was 7.3 percent with a worst case error in the above prediction equation of 8.8 percent.

With U.S. spring wheat the relationship is considerably different. The coefficient of variation for total production is 12.1 percent. When planted acreage is used as the independent variable to predict production, the error is 7.9 percent. Adding rainfall reduced the C.V. to 2.0 percent. When using harvested acreage only, the error is reduced to 5.4 percent. When planted and harvested and percent abandoned wheat is used, the error is reduced to 5.0 percent with a worst case error of 6.5 percent.

In summary acreage planted is the most important variable in determining production in irrigated wheat areas. In winter wheat areas rainfall is the most important factor followed by other catastrophic factors with area planted contributing very little to the overall variability. In the spring wheat areas, acreage planted explains the largest portion of the variability followed by rainfall with other catastrophic factors contributing the least of overall variability.

The lack of correlation between acreage and production may be due to the cropping practices used under the planting allocation imposed by the government. However, acreage as an indicator may become even less reliable as the less productive lands are forced into use by the removal of planting allocations and higher prices paid for wheat in the expanding food market. This lack of reliability would be due to the higher probability of crop failure after planting on the marginal lands and the increased response of the wheat crop to other variables in the system under the higher stress conditions.

III. WHEAT IDENTIFICATION AND ACREAGE ESTIMATION

From data obtained in Maricopa County, Arizona, it appears that multi-date and multi-spectral images will be required to separate wheat from all other crops present in an area. Wheat can be identified and separated from all other crops in the last 30 days prior to harvest. To do this, however, it requires multi-date and multi-spectral imagery. The results of late season multi-spectral discriminant analysis of four square mile test areas in Maricopa County show a 100% identification of wheat on a field-by-field basis with no crops being confused with wheat and no wheat fields being confused with other crops. Two months prior to harvest, 84% of the wheat fields were correctly identified. The 16% misidentified were all called alfalfa. However, compensating errors occurred where alfalfa and barley were called wheat so that the overall acreage estimate after the compensating errors was 96%. These separations were made from all crops present, which were alfalfa, sugar beets, cotton, safflower, barley, bare soil, urban and native vegetation. The study was conducted using 1970 black and white multiband imagery from the RB 57 photo flights. The images were scanned in common register using a microdensitometer to simulate ERTS spectral and spatial resolution. The study was aimed at total crop inventory rather than a single crop inventory of wheat. It is felt that training on all crops present is required to obtain the high level of accuracy presented in this study. Discriminant analysis using thresholding techniques on the single crop of interest will probably provide lower percentage corrects on field identification. No studies have been conducted in identification of wheat under ERTS-1 at the Center for Remote Sensing Research because the 1972 image acquisition from ERTS started after wheat had been completely harvested and the 1973 data that contains the wheat inventory information has not yet been made available. It is felt, however, that the results from ERTS-1 will be at least as good as the results obtained from the scanning of multispectral black and white imagery from aircraft.

false & must investigate data prior to classification

It was indicated by Williams, Mrain, Baker, and Coiner, that when the investigator has adequate knowledge of local environment and local crop cycle, planted wheat may be identified very early in the season. In this study in Kansas, wheat was separated from sorghum for grain, corn for grain, corn for silage, alfalfa hay, and sugar beets. They discovered that, in September, wheat could be separated from all other crops 89% of the time. In compensation errors where non-wheat was called wheat, gave them an overall accuracy of 99% in the estimation of wheat, field-by-field and in acreage estimate. The decision rules used by the photo interpreter were equivalent to th those used in automatic discriminant analysis when human stratification of the image was used as an initial classification. Therefore, it is felt that in areas similar to Kansas the automatic classifier should do as well. It should be noted, however, that in this study and in the early identification of wheat in Maricopa County, Arizona, the large percentage of the total estimate for wheat (20%) is from non-wheat which would lead to a high variance in the estimate of wheat

[illegible] do [illegible] utilization for Spectral data over the long run. Thus reducing the confidence one would have in the estimate of wheat acreage early in the season. Near harvest time nearly all wheat is properly identified creating a high degree of confidence in the results. Because of the high correlation in many areas between wheat yield and wheat acreage planted, the confidence in the estimate for wheat production would be correspondingly low for estimates made early in the year.

Because of the spatial resolution of the ERTS system, the proper identification of an individual field or "PIXEL" does not guarantee acreage estimates within acceptable limits. To investigate this source of error and the confusion class source of error, two studies were conducted at the Center for Remote Sensing Research to determine the relationship between ERTS acreage estimates and acreage from high flight imagery. First, to investigate the field size problem, randomly selected contiguous fields identified on the output of the discriminant analysis were compared with the actual planted acreage as interpreted and measured on high flight imagery for alfalfa, corn, and asparagus. Second, to investigate the field size and confusion crop problems simultaneously, sampling units were selected using the procedure described in Section VI of this report for asparagus.

Fifty-two fields were selected from the intensive test cells within San Joaquin County representing corn, asparagus, and alfalfa. Field size ranged from 15 to 550 acres. The linear model $Y_i = \beta_o + \beta_i X_i + e_i$ used to develop a prediction equation where Y_i = the photo ground estimate of the acreage of the field and X_i is the ERTS estimate of the acreage for the field and e_i is the error associated with each sample. The coefficients a_i relating the ERTS estimate to the ground estimate were estimated using the Least Square Procedure (LSE). From this analysis the amount of variation explained by the regression relatonship between the ERTS and photo/ground estimates was computed for each of the three crops, asparagus, alfalfa, and corn. The results are summarized in Table 1 for the classified ERTS discriminant analysis estimates of acreage.

Table 1

	ASPARAGUS	CORN	ALFALFA
Source	Mean Square Error	Mean Square Error	Mean Square Error
Regression	224,007.85	150,780.01	82,341.72
Residual	3363.72	24.83	524.23
Total	227,371.57	150,804.84	82,865.95

The results show that much of the variance of field size on the ground can be explained by the estimate obtained from ERTS data. There is, however, a significant difference between the actual acreage measured on high flight imagery and the acreage estimated from ERTS that is not explained by the regression relationship and, furthermore, the regression relationship appears to be dependent on crop type rather than being a general relationship that can be applied to all crops.

The variation in the estimate for individual fields, while significant, is only part of the problem of estimating crop acreage. The other part of the problem is the inclusion or exclusion of areas because of misidentification during the discriminant analysis procedure. To evaluate this problem an independent sample of asparagus, one of the most difficult crops to obtain acreage estimates for using discriminant analysis because of the confusion with native vegetation and bare soil, was selected using the multistage sampling procedure for agricultural acreage estimation.* It was selected because the human can easily identify this crop on the aerial photography that would be used as the second stage in the acreage estimation.

Segment

Eight primary sampling units (PSUs) were selected proportional to the acreage estimated by the discriminant analysis. The PSUs were transferred to 1:120,000 Color Infrared imagery where the area of the growing crop of fields identified by photo interpretation as asparagus were measured. Plotting the Y_i's (photo estimate) versus the X_i's (ERTS estimate) (Figure III.1) indicates that the relationship does not pass through the origin. Because of this, the linear regression model $Y_i = \beta_o + \beta_1 X_i + e$ was used and a least squares estimate for β_o and β_1 computed. The resulting prediction equation was: $\hat{Y} = -292. + 2.033 (X_i)$.

The coefficient of variation of the mean of the sample ($\bar{Y}$) was 41%, which was reduced to 13.8% by the regression relationship between the X_i's and the Y_i's.

Although these studies did not specifically cover wheat, they do document the problems of acreage estimation from relatively low resolution satellite systems and indicate an approach that can be used to overcome the problem at the minimum cost.

*See Section VI of this report for a detailed description of the method.

The confusion of wheat fields with surrounding grass lands and other cereal grains may make the estimation of acreage without sampling and regression estimation extremely unreliable. It may, however, be possible to develop correction procedures that can be applied more generally through detailed studies of the wheat problem.

IV. DIRECT PREDICTION MODELS

The simplest and possibly the most practical model in the short run may be one which utilizes direct observation and measurement of conditions at discreet times to estimate wheat yield and production. Spectral characteristics, textural characteristics, and area measurements in many cases are highly correlated with biological production and site conditions (Miller). With this directly observed information and auxiliary information from weather satellites concerning mean temperatures, mean rainfall by month, and abnormally high or low conditions of rainfall and temperature, we can provide most of the information necessary to complete the prediction model. In exceptional cases where lodging, freezing, excessive rainfall, draught, economic conditions, unusual snow conditions, insect, and disease conditions, subtractions would be made from yield to compensate for the area damaged, weighted by the appropriate reduction factor. For most of the factors concerned, adjustments would be made for conditions other than normal. Initially, the production would be estimated based on historical information. After data is acquired from the satellite, adjustments would be made, both positive and negative, to bring the estimate in line with the most current data. For example, with winter wheat, the early estimate of planted area along with rainfall to date and current economic conditions, available seed and labor, would be used to project the production of wheat. As time passed and information was obtained on accumulated rainfall, freezing conditions, snow cover (providing insulation and moisture), and updated economic information, these would be applied to the model at that point in time to predict total production. This procedure would continue through the winter months to the time of thaw, and the relationship between thawing, rainfall, the first appearance of green, and the winter's snowfall, would be used to update the prediction. One of the most important, directly observable parameters would be the appearance of green after the spring thaw. The date of green and the textural patterns associated with the green appear to be strong indicators of the soil moisture and expected crop vigor associated with annual plants. As the crop began to grow and mature in the early part of the growing season, spectral characteristics such as the ratio between the chlorophyll absorption band and the infrared band could be used as an indication of actively producing tissue (Miller). The length of time between first appearance of green and the drying of the wheat is an indication of stored soil moisture and temperature conditions associated with total plant production and the subsequent yield of wheat. As the crop began to dry, the patterns between the drying and still green material would be an indication of the residual soil moisture and soil depth, and therefore, the expected yield from the area. These characteristics of greening and drying would be observed and quantified for both the wheat and annual grass lands to be used as an index in the model for wheat production. This textural information would also be very important in the classification of site potential for growing wheat. In range studies (Carneggie) a high correlation between soil depth and the relative length of the active growing season has been found. Coefficients are available for wheat produced for a given amount of rain on a given site. Coefficients are also available that relate minimum and maximum temperatures to biological productivity. Based on the information obtained on meteorological or climatic conditions and site conditions relating to soil fertility, it is felt that one of the most important stages in the prediction of wheat yield will be the stratification of the world's wheat growing areas into similar areas relating to potential. This stratification or determination of analogous areas will be extremely important to any yield prediction model.

Another use of pattern recognition similar to the crop identification algorithms may be valuable in the overall process of predicting crop yield. In the final stages of predicing crop yield, based on all of the input parameters measured during the growing cycle within a given stratum, pattern recognition may be the most efficient means of separating the areas into similar yield classes.

For each field identified as wheat by the discriminant analysis procedure of the spectral information from either single or multi-date ERTS imagery, estimates for the various parameters affecting wheat yield would be obtained. These parameters would then be utilized as features in the pattern recognition program.

Examples of such features for the final yield prediction are total rainfall prior to planting, planting date, minimum temperature prior to emergence, snow cover, rainfall after emergence, preharvest spectral characteristics, degree day, time from planting to harvest, green wave duration (the time from greening to drying of the wheat during the season), minimum temperatures at various points during cycle, maximum temperature at various points during cycle, estimates of soil moisture, estimates of rainfall post planting, spectral characteristics of fields, and many other such parameters. Many of the parameters would be estimated on a regional basis, while others such as rainfall, may be estimated on a more localized basis from overlayed meteorological information from satellite sources. Other parameters would be estimated by interpolating between point samples from non-scanning or nonimaging satellites.

PROPOSED FUTURE MODELS FOR WHEAT YIELD ESTIMATION

Section 0.5 Some Comments on Modelling

The act of constructing and utilizing mathematical abstractions of real world systems or processes to gain information concerning equilibrium conditions among system variables is called modelling. Models can be time invariant, that is their response pattern for a given set of inputs is constant. On the other hand, models that give a different response for the same input set at different times are termed dynamic. The model variables can be deterministic (zero variance) or probabilistic in character.

The natural system we wish to model, wheat production, is governed primarily by probabilistic rather than by deterministic relationships. The first two of the three mathematical models proposed below reflect this fact and are largely stochastic (probabilistic) in nature. It is supposed here that a probabilistic model tends to provide the best "fit" of a probabilistic real system. However, the third model described has a deterministic emphasis. Inclusion here results from its ability to be expressed as a circuit theory analogy. Hence this model could potentially be simulated quickly through conventional circuit analysis to give reasonable yield estimates at very competitive costs.

None of the three mathematical models need be considered the "best" to the exclusion of all others. Rather, it is proposed that if fully developed each should complement and enhance the prediction accuracy of the other. Relationships among the models are suggested for this eventual multi-model design.

Section 1.0 Introduction to the Form of Model Type I

The first mathematical model proposed to predict wheat yield is composed of a series of polynomial regression equations. Each equation allows estimation of yield at harvest or effective yield at market based on the physical, chemical, biological, cultural, and economic environment existing in any given preharvest time period i and based on the most recent previous prediction of yield (if any). In this discussion yield will be defined on a weight per acre basis. Then the model for wheat yield based on any previous time i in the wheat growth and pre-growth cycle can be expressed by the following equation:

$$\begin{aligned}\hat{Y}_{i_{\text{harvest}}} = b_o &+ (b_1\,\text{Physical}_1 + b_2\,\text{Physical}_2 + \ldots + b_c\text{Physical}_c)\\ &+ (b_{c+1}\text{Chemical}_1 + b_{c+2}\text{Chemcial}_2 + \ldots + b_d\text{Chemical}_d)\\ &+ (b_{d+1}\text{Biological}_1 + b_{d+2}\text{Biological}_2 + \ldots + b_e\text{Biological}_e)\\ &+ (b_{e+1}\text{Cultural}_1 + b_{e+2}\text{Cultural}_2 + \ldots + b_f\text{Cultural}_f)\\ &+ (b_{f+1}\text{Economic}_1 + b_{f+2}\text{Economic}_2 + \ldots + b_g\text{Economic}_g) \qquad \text{eq. 1.1}\\ &+ b_{g+1}\hat{Y}_{i-1_{\text{harvest}(i-1\neq 0)}} + e_i\end{aligned}$$

for i = 1 to i = harvest or market = n

where

$\hat{Y}_{i_{\text{harvest}}}$ = predicted yield at harvest or market based on pre-harvest time period i

$Physical_\ell$, $Chemical_\ell$, $Biological_\ell$, $Cultural_\ell$, $Economic_\ell$ = physical, chemical, biological, cultural, and economic variables that drive the wheat system. These predictors are considered the independent variables allowing prediction of the dependent variable $\hat{Y}_{i_{harvest}}$.

$\hat{Y}_{i-1_{harvest}}$ = predicted yield at harvest or market based on pre-harvest time period i—1. $Y_{i-1_{harvest}}$ only included in equation 1.1 if $i - 1 \neq 0$.

b_o = yield axis intercept

$$= \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 - \ldots - b_k\overline{X}_k$$

where the X_j represent the independent variables.

$b_1 \ldots b_{g+1}$ = partial regression coefficients of $\hat{Y}_{i_{harvest}}$ on the independent variables defined above.

= the increase in $\hat{Y}_{i_{harvest}}$ if X_i is increased one unit while all other X variables are held constant

e_i = an error term arising from either measurement error of eventual Y (yield) or from stochastic error resulting from the influence on Y of omitted independent variables

Alternatively we may state e_i as e_{iq} or e_q (a given i assumed) where q denotes a given set of values for the variables in eq. 1.1. Then the basic assumptions in the above model are that:

(1) e_q is a random variable with mean equal to zero and variance σ^2 (unknown), alternatively stated as $E(e_q) = 0$, $V(e_q) = \sigma^2$.

(2) e_g and e_h are uncorrelated, $g \neq h$ (g and h representing q values), giving $\mathrm{cov}(e_g, e_h) = 0$.
It follows that

$$E(\hat{Y}_{iq_{harvest}}) = \beta_o + \beta_1 X_{1q} + \beta_2 X_{2q} + \ldots + \beta_{,} X_{kq}$$

where β_j are the population means of the corresponding b_j, and that $V(\hat{Y}_{iq_{harvest}}) = \sigma^2$, and that $\hat{Y}_{ig_{harvest}}$ and $\hat{Y}_{ih_{harvest}}$, $g \neq h$, are uncorrelated.

(3) A final additional assumption necessary to carry out later F and t tests is that e_{iq} is normally distributed, i.e., $e_q \sim N(0, \sigma^2)$. This implies that e_g and e_h may be considered independent as well as uncorrelated.

Section 1.1 Derivation of Independent Variables

The above model is stochastic in that there is an error term attached to the estimated yield, $\hat{Y}_{i_{harvest}}$. However, deterministic relationships are often useful to arrive at values for some of the independent variables X_j for a specific time i. Two examples of these deterministic relationships are given below, one for temperature and one for evaporative stress. Others could be added depending on the degree of sophistication in the independent variables desired.

Section 1.1.1 Temperature

To improve the prediction of plant response to the daily (diurnal, 24 hours) temperature cycle, four temperature variables are formulated. All four variables are hypothesized (based on literature review) to be strongly related either to the rate of net biomass assimulation in wheat or its biological capability to function. These four temperature "types" are (1) the high temperature range (TR'_{hot}) at which wheat must spend more of its energy reserves to remain alive than it can assimilate (net negative biomass accumulation). Higher portions of this temperature range give rise to denaturing of proteins and thus plant green biomass destruction. An intermediate range of temperature ($TR'_{effective}$) allows net positive assimilation. At lower temperatures (TR'_{cool}) the plant must use more energy than it stores to maintain itself and at still lower temperatures (TR'_{cold}), plant protein destruction occurs. These temperature ranges, their threshold values, and integration (shaded areas) over a 24 hour period are represented in terms of a diurnal cycle in Figure M.1 below.

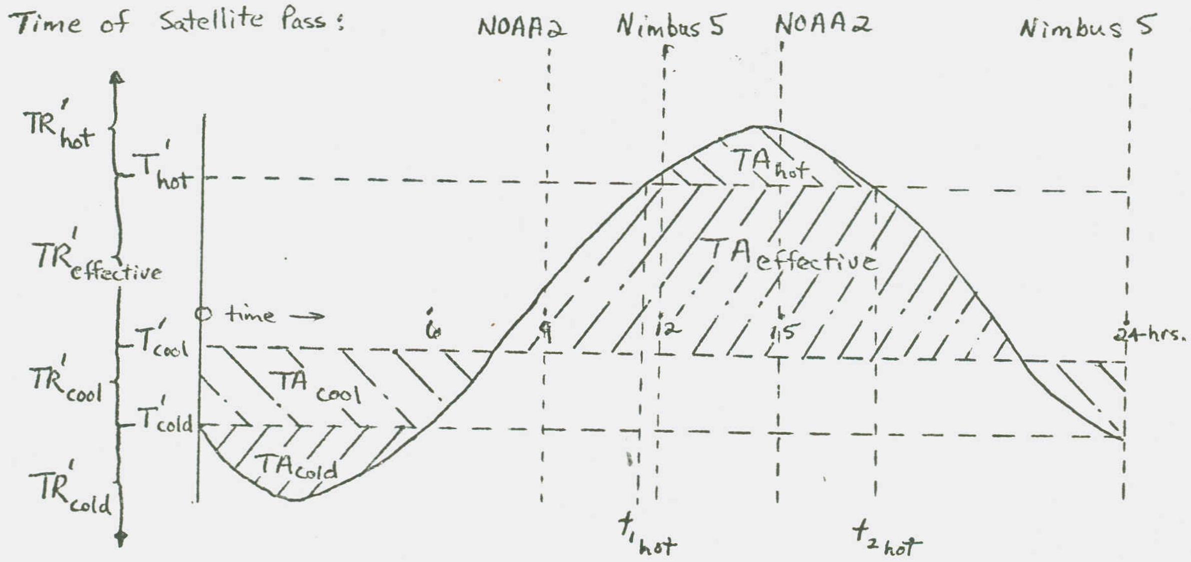


Figure M.1.

The above diurnal temperature cycle can be expressed mathematically as a cosine function. Actual diurnal cycles tend to have both their positive and negative peaks skewed somewhat to the right (e.g., Gates, 1965). However, if the investigator is interested in integrating temperature ranges over time to give yield predictions (X., in this case TA'_{hot}, $TA'_{effective}$, TA'_{cool}, TA'_{cold}) the cosine function allows a good approximation for these variables. More sophisticated diurnal cycle analysis could use Fourier series to define a function more closely fitting (i.e., lower residual sum of squares) actual cycle values.

The temperature at any point in time t may be represented by the following deterministic equation:

$$T'_t = A \cos\left(\left(\frac{2\pi}{1\ \text{day}}\right)t - \phi\right) \qquad \text{eq. 1.2}$$

where

T'_t = temperature at time t standardized to the scale of A (see below)

$t = 0 = \frac{0}{24}$ at 0000 hours

$t = 1 = \frac{24}{24}$ at 2400 hours

$\frac{2\pi}{1\ \text{day}} = w =$ the constant of proportionality between time displacement and angular displacement. Here 1 day (24 hours) corresponds to the period of the cycle.

ϕ = phase angle that describes the relationship between the time t at T_{max} (maximum temperature point) and t = 0.

$$A = \frac{T_{max_{approx}} - T_{extreme\ ave.}}{\overline{T}_{max_{approx}} - \overline{T}_{extreme\ ave.}}$$

where

$$T_{max_{approx}} = \frac{\text{T measured at 1200 hours by Nimbus 5 and T measured at 1500 hours by NOAA2}}{2}$$

$T_{extreme\ ave.}$ = approximate average of diurnal temperature extremes measured at 900 hours by NOAA2 (could also be obtained by ground measurements).

$\overline{T}_{max_{approx}}$ and $\overline{T}_{extreme\ ave.}$ are the respective yearly averages of the above two variables.

Utilizing the above definitions we have

$$T'_t = \frac{T_t - T_{extreme\ ave.}}{\overline{T}_{max_{approx}} - \overline{T}_{extreme\ ave.}} \qquad \text{eq. 1.3}$$

where T_t = temperature in °C.

T_{max} is used here to set A and ϕ since good satellite estimates of this value (Nimbus 5 at 1200 hours and NOAA2 at 1500 hours, see Figure M.1) exist while such measurements do not exist for T_{min}. ϕ will be a function of the time of year, shading, wind, and weather systems. Only time of year will be considered in this treatment. Since seasonal average temperatures lag potential solar beam irradiation averages (Oort, Scientific American, 1970), the lowest average temperature day of the year in the northern hemisphere will be selected as approximately

January 21 for this example. Adding 182.5 days gives July 23 as the day having the average hottest temperature. Review of the literature (e.g., Gates, 1965) indicates that cooler days tend to have earlier T'_{max} peaks. Hence for illustration, 1200 hours is taken as the time of T_{max} for January 21 and 1430 hours for July 23. Then

$$\phi_{Jan.21} = 12 \text{ hrs.} \times \frac{2\pi}{24} = \pi$$

$$\phi_{July\ 23} = 14.5 \text{ hrs.} \times \frac{2\pi}{24} = 1.21\pi$$

$$\phi_{Jan.22 \to July\ 22} = \pi + (\text{Julian Date} - 21)\left(\frac{14.5 - 12.0}{182.5}\right)\left(\frac{2\pi}{24}\right)$$

$$\phi_{July\ 24 \to Jan.20} = 1.21\pi - (\text{Julian Date} - 203.5)\left(\frac{2.5}{182.5}\right)\left(\frac{2\pi}{24}\right)$$

Now to integrate to obtain TA_{hot}, $TA_{effective}$, TA_{cool}, and TA_{cold} the intersection of the lines $T' = T'_{hot}$, $T' = T'_{cool}$, and $T' = T'_{cold}$ with $T' = A \cos\left((2\pi)\ t - \phi\right)$ must first be found.

Setting

$$T'_{threshold} = A \cos\ ((2\pi)t - \phi) \qquad \text{eq. 1.4}$$

and solving for t will provide the needed intersections.

Then for T'_{hot} we have

$$T'_{hot} = A \cos((2\pi)t - \phi)$$

$$\frac{T'_{hot}}{A} = \cos((2\pi)t - \phi)$$

$$\arccos \frac{T'_{hot}}{A} = (2\pi)t - \phi$$

$$\frac{\arccos \frac{T'_{hot}}{A} + \phi}{2\pi} = t \Rightarrow t_{1_{hot}},\ t_{2_{hot}}$$

Similar solutions hold for T_{cool} and T_{cold} and in general we have

$$\frac{\arccos \dfrac{T'_{threshold}}{A} + \phi}{2\pi} = t \Rightarrow t_{1_{threshold}},\ t_{2_{threshold}} \qquad \text{eq. 1.5}$$

Now proceeding directly to the integrations we have

$$TA_{hot} = \int_{t_{1_{hot}}}^{t_{2_{hot}}} A \cos((2\pi)t - \phi)\, dt - T'_{hot}(t_{2_{hot}} - t_{1_{hot}}) \qquad \text{eq. 1.6}$$

$$TA_{effective} = \int_{t_{1_{cool}}}^{t_{2_{cool}}} A \cos((2\pi)t - \phi) dt - TA_{hot} \qquad \text{eq. 1.7}$$

If $t_{2_{photoperiod}} - t_{1_{photoperiod}} < t_{2_{cool}} - t_{1_{cool}}$

then use

$$TA_{effective} = \int_{t_{1_{photoperiod}}}^{t_{2_{photoperiod}}} A \cos((2\pi)t - \phi) dt - TA_{hot} \qquad \text{eq. 1.8}$$

where $t_{1_{photoperiod}}$ = sunrise − skylight period

$t_{2_{photoperiod}}$ = sunset + skylight period

Tentatively set skylight period as .75 hour.

$$TA_{cold} = \int_{t_{1_{cold}}}^{t_{2_{cold}}} A \cos((2\pi)t - \phi) dt - T'_{cold}(t_{2_{cold}} - t_{1_{cold}})$$

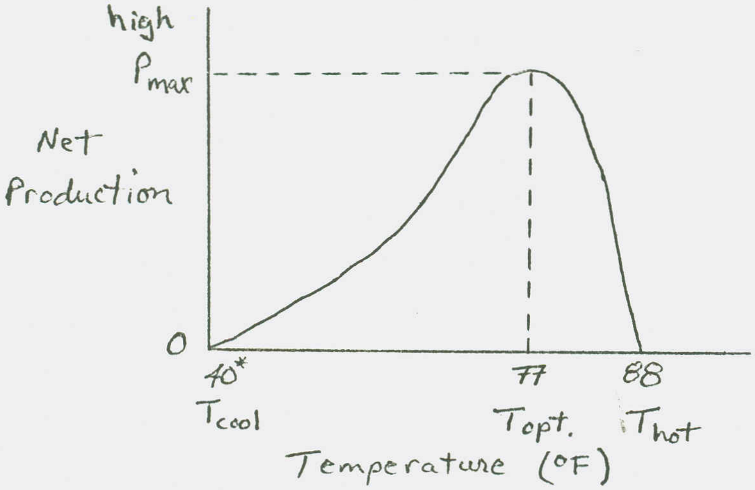
eq. 1.9

$$TA_{cool} = \int_{t_{2_{cool}}}^{t=1} A\cos((2\pi)t - \phi)dt + \int_{t=0}^{t_{1_{cool}}} A\cos((2\pi)t - \phi)dt - TA_{cold}$$

eq. 1.10

It should be noted that the above integration limits are specific for the given ϕ, T'_{hot}, T'_{cool}, and T'_{cold} of Figure M.1. Different values of these parameters may give rise to different integration limits.

A further refinement to the $TA_{effective}$ predictor will now be given. Review of the literature (e.g., Leonard and Martin, 1963) indicates that rate of net positive assimilation of biomass is not constant in the temperature range T_{cool} to T_{hot}. Rather a function relationship in Figure M.2 is indicated.



* Temperature Values from Leonard & Martin (1963).

Figure M.2.

Hence a given temperature within the T_{cool} to T_{hot} range must be weighted according to its ability to give rise to net production. To perform this weighting procedure it is proposed that 1° to 2°C class intervals be defined in the T'_{cool} to T'_{hot} range in Figure M.1 and the area ($TA_{effective}$) under the cosine function between $t_{1_{cool}}$ and $t_{2_{cool}}$ for each such class interval be integrated separately. Then each such area should be weighted by the average production for that temperature class divided by the maximum production possible (P_{max}). These weighted areas would be summed to give $TA'_{effective}$:

$$TA'_{effective} = \sum_{v=1}^{n} \left[\left(\frac{P_v}{P_{max}} \right) \left(TA_{v_{effective}} \right) \right] \qquad \text{eq. 1.11}$$

The three cardinal points for growth: T_{cool}, $T_{optimum}$, and T_{hot} will vary according to the variety and strain of wheat and according to its growth stage (Levitt, 1969). Data from Leonard and Martin (1963) indicate that for most types of wheat, T_{cool} varies approximately from 38.5°F to 41.0°F, $T_{optimum}$ ∿77°F, and T_{hot} from 87.5°F to 88°F. A simplified approach would be to take the averages given by Leonard and Martin and assume a constant production versus temperature curve through the wheat life cycle. Or a curve could be fitted to temperature -- yield data by life cycle or growth stage to give production values from T_{cool} to T_{hot}. Such curve fitting processes might utilize functions of the form

$$Y = b_o + b_1T + b_2T^2 + b_3T^3 \qquad \text{eq. 1.12}$$

or

$$\text{Log } Y = c_o + c_1 \log T + c_2T \qquad \text{eq. 1.13}$$

where T could be T_{max} or T_{min} measured on the ground or estimated with satellite data, c_1 should be positive and c_2 negative (Freese, 1964).

TA_{hot}, $TA'_{effective}$, TA_{cool}, and TA_{cold} may be added day-by-day to give total figures for some time stage i of the wheat cycle. They may then be used in the regression model of eq. 1.1 as independent variables in the prediction of wheat yield $\hat{Y}_{i_{harvest}}$.

Section 1.1.2 Evaporative Stress

The second example of the use of a deterministic relationship to provide input to a probabilistic model is given by the determination of evaporative stress. Many sophisticated formulas can be given to derive instantaneous evaporative stress (Federer, 1970). Most, however, require one or more ground measurements of parameters not available from satellites or most weather stations.

One apprach to determination of effective evaporative stress using satellite data is a modification of the Eddy Correlation method (Federer, 1970). Here vaporation, E, is defined as

$$E = \rho \overline{w'q'} \qquad \text{eq. 1.14}$$

where

ρ = air density as g/cm^3

w' = instantaneous vertical component of wind

q' = instantaneous deviation of specific humidity $(gH_2O/g\,air)$, q, from its mean

$\overline{w'q'}$ = average deviation of the instantaneous water-vapor flow from its mean value.

Now letting

$$q = \frac{0.622}{p} e \quad \text{(Federer, 1970)} \qquad \text{eq. 1.15}$$

where e = vapor pressure and p = atmospheric pressure, we may define an evaporative stress index, EI, as

$$EI = -\rho \frac{1}{q} \xi \qquad \text{eq. 1.16}$$

ξ = slope of vertical temperature structure near the surface

ξ proportional to vertical wind

ρ and q are as defined above

and negative sign to account for negative ξ under adiabatic lapse rate conditions exhibiting decrease in temperature with an increase in altitude (i.e., lack of temperature inversions).

Values for variables in eq. 1.15 and eq. 1.16 should be derivable from Nimbus 5 and NOAA2 radiometer data. Future satellites with improved spacial resolution would tend to give more "area-specific" data for those variables. For most accurate yield estimates, meteorological data spacial resolution should be at the scale of individual fields.

A probabilistic alternative to the determination of an evaporative stress independent variable may be given by utilization of the surface energy balance equation (Federer, 1970):

$$L_vE = Rn - H - S - M - Sn \qquad \text{eq. 1.17}$$

where

L_vE = energy used in transpiration

L_v = latent heat of vaporization

E = the mass loss of water

Rn = net radiation

H = sensible heat (heat gained or lost by air above the surface)

S = heat gained or lost by soil and vegetation

M = heat used in metabolism (primarily photosynthesis minus respiration)

Sn = heat lost on freezing of snow and heat gained on melting of snow.

All terms in the above equation have units of energy per unit area per unit time and, when integrated over time, are expressed as energy per unit area.

If it is assumed that most of the energy required for evaporation is supplied by solar radiation (Federer, 1970), Rs, and

$$Rn = (1 - a)Rs + \varepsilon R_{td} - \varepsilon \sigma T_0^4 \qquad \text{eq. 1.18}$$

where

Rn = net radiation

a = albedo (reflected solar radiation)

Rs = downward solar radiation

ε = absorptivity or emissivity of the surface for thermal radiation

σ = is the Stefan-Boltzmann constant

R_{td} = thermal radiation emitted downward by the atmosphere

T_0 = surface temperature

$(1 - a)Rs$ = solar radiation absorbed

εR_{td} = thermal radiation absorbed

$\varepsilon \sigma T_0^{\ 4}$ = thermal radiation emitted

Then evaporative stress may be made a function of Rs via a model as in eq. 1.19

$$\hat{E} = b_0 + b_1 Rs + e \qquad \text{eq. 1.19}$$

where b_i and e are defined in a similar fashion to eq. 1.1 and $\hat{E}$ equals predicted relative evaporative stress.

Since satellite data is or will be available on many of the other parameters in eq. 1.18, their inclusion might increase the accuracy of prediction for $\hat{E}$. Thus

$$\hat{E} = b_0 + b_1 \left[(1 - a)Rs\right] + b_2 \varepsilon + b_3 T_0^{\ 4} \qquad \text{eq. 1.20}$$

It should be noted that Rs on a clear day may be obtained by the following equation discussed by Frank and Lee (1966):

$$Rs_{\substack{clear\\day}} = \frac{I_0}{r^2}\left(2t \cdot \sin\theta \cdot \sin\delta + \frac{1}{\omega} \cdot \cos\theta \cdot \cos\delta \cdot 2 \cdot \sin\omega t\right) \qquad \text{eq. 1.21}$$

Here

$Rs_{\substack{clear\\day}}$ = maximum potential downward solar beam irradiation

I_0 = solar constant

r = radius vector, ratio of the earth-sun distance at a particular time to its mean

θ = terrestrial latitude

δ = solar declination

w = angular velocity of the earth's rotation, 15° per hour

t = number of solar days

Solar energy transmitted on cloudy days may be significantly less (Bates, 1965 and Connor, 1973) differentially by wavelength. Its effect on evaporative stress and net biomass production should be a point of further investigation.

Section 1.2 Model Type I as a System of Regression Models

Once values for physical, chemical, biological, cultural, and economic aspects of the environment have been obtained directly from ground and/or satellite measurement or via deterministic and/or probabilistic methods such as those just discussed, then those values may be plugged into eq. 1.1. Equation 1.1 may be restated in terms of a sum of environment vectors:

$$\begin{aligned} Y_{i_{\text{harvest}}} = {} & b_0 + \vec{T}_i \vec{b}_{T_i} + \vec{P}_i \vec{b}_{P_i} + \vec{E}_i \vec{b}_{E_i} + \vec{S}_i \vec{b}_{S_i} \\ & + \vec{B}_i \vec{b}_{B_i} + \vec{C}_i \vec{b}_{C_i} + \ldots + \vec{EM}_i \vec{b}_{EM_i} + e_i \end{aligned} \qquad \text{eq. 1.22}$$

where

$\vec{T}_i$ = a vector of temperature variables, e.g., TA_{hot}, $TA'_{effective}$, TA_{cool}, TA_{cold}

$\vec{P}_i$ = a vector of precipitation variables, e.g., mean monthly ppt., 20-day moving average of precipitation, snow depth

$\vec{E}_i$ = a vector of evaporative stress variables, e.g., EI, $\hat{E}$

S_i = a vector of soil variables, e.g., soil moisture, water table level, soil depth, conductance, pH

B_i = a vector of biological variables, e.g., percent of field area infested by a pathogen or insect

C_i = a vector of cultural variables, e.g., amount of irrigation, fertilization, application of insecticides and/or herbicides; planting and harvesting techniques

EM = a vector of economic and market constraints

b_{x_i} = the vector of partial regression coefficients corresponding to variables within the vector group X

Model Type I can now be written as a system of regression models yielding a wheat production prediction based on environment for individual time periods and on previous predictions of yield. Use of eq. 1.1 gives the system as

i=1 $$Y_{1_{harvest}} = b_{1,0} + \sum_{r=1}^{h} \overrightarrow{X_{1r}}\,\overrightarrow{b_{1r}} + e_1$$

i=2 $$Y_{2_{harvest}} = b_{2,0} + \sum_{r=1}^{h} \overrightarrow{X_{2r}}\,\overrightarrow{b_{2r}} + \hat{Y}_{1_{harvest}}\, b_{2,h+1} + e_2$$

o o
o o
o o

i=n $$Y_{h_{harvest}} = b_{n,0} + \sum_{r=1}^{h} \overrightarrow{X_{nr}}\,\overrightarrow{b_{nr}} + \hat{Y}_{i-1_{harvest}}\, b_{n,h+1} + e_n$$ eq. 1.23

where $\overrightarrow{X}_{i,r}$ and $\overrightarrow{b}_{i,r}$ refer, respectively, to a given environment variable vector r (e.g., $\vec{T}$) and its corresponding partial regression coefficient vectors (e.g., $\vec{b}_T$) for time period i. The $\hat{Y}_{i-1_{harvest}}$ represent the feedback components of Model Type I. Each $\hat{Y}_{i-1_{harvest}}$ allows the $\hat{Y}_{i_{harvest}}$ (i = some K) to be corrected for the environment or sequence of environments (i = 1,...,K-1) whose sum effect is represented by $\hat{Y}_{i-1_{harvest}}$.

The length of any time period i could be one day (24 hours). However, longer periods of time corresponding to phenological stages are most useful and efficient for modelling plant yield (Sauer, unpublished). These phenological stages or phenophases include (1) winter dormancy or first visible growth; (2) one or more vegetative growth periods; (3) floral buds, open flowers, and ripening fruit; and (4) buds, flowers, green and ripe fruit. The plant in a given phenophase will tend to respond in a characteristic fashion to environmental stimuli. A given order and magnitude of a set of environmental events will tend to have a different effect on yield according to phenophase.

Each combination of morphology and physiology defining a phenophase may last several days to several weeks. Thus the applicable time period for each regression equation in eq. 1.23 may differ. Then to obtain values for the independent variables (predictors) in each equation, values for individual days must be summed over the number of days in the given pre-harvest time period i.

In addition to phenophases, other pre-harvest time periods might include a set of pre-planting intervals. The environment of these pre-crop periods could be used to provide early predictions of wheat yield where water, temperature, or soil, etc., conditions might be limiting.

It should be noted that when an environmental variable indicates a pre-harvest termination of the crop (e.g., (1) death resulting from a catastrophic event or (2) planting cancellation due to an unsatisfactory pre-planting environment) for a given area, then the yield prediction for that wheat life cycle is set to zero and i is reset equal to one. The threshold values for such termination must be determined for a given wheat strain before use of the model.

Section 1.3 Sensitivity Analysis for Model Type I

Part I: ANOVA

The first portion of the sensitivity analysis involves the construction of an analysis of variance (ANOVA) table (see Figure M.3). This table allows the investigator to determine the significant independent variable or variables controlling crop yield response.

Each row pair in the table gives, first, the variation (given under MS) about the average yield $\overline{y}_{harvest}$ attributable to regression on the particular independent variables utilized. The second row of each row pair gives the error or variation about $\overline{y}_{harvest}$ not attributable to the given regression. Assuming an appropriate model*, dividing the error variation into the regression variation allows a test of the null hypothesis that the regression coefficient equals zero ($H_o : \vec{\beta j} = 0$). This test is carried out with use of the statistical F distribution. Acceptance of H_o means that there is no statistically significant change in wheat yield with a unit change in the particular

*See Appendix I for a discussion of methods to determine the correctness of regression models.

Figure m.3

Analysis of Variance (ANOVA) Table

Definitions: KK = maximum possible no. of independent variables in the model.

u = no. of wheat cycles for which there is X_j and $Y_{i\,\text{harvest}}$ data for time period i.

j range*	Source	Degrees of Freedom (d.f.)	Sum of Squares (SS)	Mean Square Error (ms) ms = SS/d.f.	F-test of H_0
$j = 1, KK$	Y on X_j error	1 $u-1-1$	$b_{x_j}\sum_{g}(X_{jg}-\overline{X}_j)(Y_{i_{harv.}}-\overline{Y}_{i_{harv.}}) = b_{x_j}\sum_{g}x_j y = R^2_{y \cdot x_j}$ $1 - R_{y \cdot x_j}$		$H_0: \hat{\beta}_j = 0$
$j = 1, KK-1$	Y on X_j, X_{j+1} error	2 $u-2-1$	$\sum_{j=z_1}^{z_2}(b_{x_j}\sum_{g}x_j y) = R^2_{y \cdot x_{j=z_1}, \dots, x_{j=z_2}}$ $1 - R^2_{y \cdot x_{j=z_1}, \dots, x_{j=z_2}}$		$H_0: \hat{\beta}_j = \hat{\beta}_{j+1} = 0$
$j = 1, KK-(KK-1)$	Y on $X_j, \dots, X_{j+(KK-1)}$ error	KK $u-KK-1$	$\sum_{j=1}^{KK}(b_{x_j}\sum_{g}x_j y) = R^2_{y \cdot x_1, \dots, x_{KK}}$ $1 - R^2_{y \cdot x_1, \dots, x_{KK}}$		$H_0: \hat{\vec{\beta}}_j = 0$
$j = 1, KK-1$	Y on $X_{j+1} \mid X_j$ error	1 $u-1-1$	$R^2_{y \cdot x_j, x_{j+1}} - R^2_{y \cdot x_j}$ $1 - R^2_{y \cdot x_j, x_{j+1}}$		$H_0: \hat{\beta}_{j+1} = 0 \mid \hat{\beta}_j$
$j = 1, KK-2$	Y on $X_{j+2} \mid X_{j+1}, X_j$ error	1 $u-1-1$	$R^2_{y \cdot x_j, x_{j+1}, x_{j+2}} - R^2_{y \cdot x_j, x_{j+1}}$ $1 - R^2_{y \cdot x_j, x_{j+1}, x_{j+2}}$		$H_0: \hat{\beta}_{j+2} = 0 \mid \hat{\beta}_{j+1}$ $\hat{\beta}_j$
	⋮				
$j = 1, KK-(KK-1)$	Y on $X_{j=KK} \mid X_{j=KK-1}, \dots X_j$ error	1 $u-1-1$	$R^2_{y \cdot x_j, \dots, x_{j=KK}} - R^2_{y \cdot x_j, \dots, x_{j=KK-1}}$ $1 - R^2_{y \cdot x_j, \dots, x_{j=KK}}$		$H_0: \hat{\beta}_{j=KK} = 0 \mid \hat{\vec{\beta}}_{j \neq KK}$
$j = 1, KK-2$	Y on $X_{j+3}, X_{j+2} \mid X_j$ error	2 $u-2-1$	$R^2_{y \cdot x_j, x_{j+1}, x_{j+2}} - R^2_{y \cdot x_j}$ $1 - R^2_{y \cdot x_j, x_{j+1}, x_{j+2}}$		$H_0: \hat{\beta}_{j+3} = \hat{\beta}_{j+2} = 0$ $\mid \hat{\beta}_j$
	Y on all possible remaining permutations of variable set$_1$ versus variable set$_2$, i.e. Y on $\vec{X}_{1j} \mid \vec{X}_{2j}$ where the j values of set$_1$ and set$_2$ are mutually exclusive, but with this restriction may take on values 1 to KK and each vector space may be from 2 to KK-2 elements long.	$\vec{X}_{1j}$ element length			
	error	$u-(\vec{X}_{1j}\text{ length})-1$			

set of independent variables. Rejection of H_0 at a given α (alpha) or probability level (e.g., with $\alpha =$ .05, i.e., with 95 percent probability) indicates there *is* a significant change in wheat yield with a unit change in the particular set of independent variables. In this case, it can be concluded that one or more variables in the given set of independent variables are statistically important in predicting wheat yield.

In order to determine which specific variables contribute most to the prediction of yield (i.e., explain the most variation about the average yield, $\bar{y}_{harvest}$) the following procedure must be utilized. First a pair row analysis of variance is performed on a series of regression equations, each succeeding regression equation including one more independent variable. Then the contribution of each independent variable, $X_{j=\ell}$, to prediction of yield is analyzed. The procedure is to determine the additional reduction in variation about $\bar{y}_{harvest}$ by regression on the particular independent variable, given that the error or residual variation has already been reduced by regression on a specified set of independent variables. Division of the variation accountable to the given variable ($MS_{x_{j=\ell}}$) by the error variance ($MS_{x_{j=a\rightarrow z, j\neq \ell}} - MS_{x_{j=\ell}}$) allows a test of H_0. If H_0 is rejected for a given α , then it can be concluded that $X_{j=\ell}$ is significantly related to yield. By expanding this procedure to a particular set of independent variables given that another set has already been used to reduce the residual variation, a systematic analysis of the contribution of each variable and each variable set to the explanation in variation about $\bar{y}_{harvest}$ could be conducted. Based on this analysis variables could be eliminated from the regression model to allow a more cost-efficient prediction of yield. Variables deleted would be those tending to explain variation in a similar fashion to other variables (i.e., significant correlation present) measured more inexpensively and/or with greater accuracy. Such would be the case where either variable or variable set gave a non-significant reduction in the residual variation unaccounted for by regression firstly with the other.

Part II: Confidence About the Predicted Yield Given X_j Without Error

The second measure of performance of the model is the confidence about the prediction $\hat{Y}_{iq_{harvest}}$ resulting from a given wheat cycle set, q, of values for the independent variables during time period i. In this case these X_{jq} values are assumed to be determined without error.

Since Model Type I is probabilistic in nature, a statement of confidence about predicted wheat yields can be constructed. This statement consists of a claim that the actually observed yield, $Y_{iq_{harvest}}$, will lie in a given value range, known as the confidence interval, with a specified probability (commonly .90, .95 or .99). The confidence interval is centered on the predicted value. This statement may be alternatively described as a measure of how precisely the yield estimate may be stated for a given level of confidence, that is probability. The confidence interval, for a given level of confidence, around a predicted value of wheat yield may be defined mathematically as:

$$CI = \pm\ t_{(d.f.,\ 1-\alpha/2)} \left[\left(\widehat{V(Y_{iq_{harvest}}))}\right)_1^{\frac{1}{2}} \div (U-1) \right]$$

eq. 1.24

where

t = a bell-shaped statistical distribution known as "Students t"; in terms of the present example it may be conceptualized as the difference between the predicted value and the eventual actually measured value divided by the standard deviation of the predicted value, for given d.f. and α.

d.f. = degrees of freedom = $(u-1) - KK - 1$.

$(u-1)$ = number of wheat cycles on which the regression is based.

KK = number of independent variables in the regression equation.

α = the probability that the eventually observed wheat yield, $Y_{iq_{harvest}}$, will actually fall within the calculated confidence interval; this probability is alternatively known as the confidence level.

$\widehat{V(\hat{Y}_{iq_{harvest}})}_1$ = estimate of the variance of the predicted wheat yield assuming no error in the measurement of X_j.

In order to compute the confidence interval an estimate of the variance for the yield prediction $\hat{Y}_{iq_{harvest}}$ for a given i and q must be determined. It can be shown (Draper and Smith, 1966) that

$$\widehat{V(\hat{Y}_{iq_{harvest}})}_1 = X_{iq}\ (X\ X)^{-1} X_{iq} \cdot s^2_{y\cdot x_{11}\ldots x_{hk},\ \hat{Y}_{i-1_{harvest}}} + s^2_{y\cdot x_{11}\ldots x_{hk}, Y_{i-1_{harves}}}$$

eq. 1.25*

*Footnote: Note that $\hat{Y}_{i-1_{harvest}}$ is treated as an independent variable and is thus included here implicitly under all X notation except in the subscript to s^2.

where q = u, k = no. of independent variables in a given environmental vector, and where

$$s^2_{y\cdot x_{11}\cdots x_{hk},\hat{Y}_{i-1_{harvest}}} = \frac{\sum\limits_{q=1}^{u-1}\left(Y_{iq_{harvest}} - \hat{Y}_{iq_{harvest}}\right)^2}{(u-1) - KK - 1} \qquad \text{eq. 1.26}$$

$$= \text{sum of squares due to regression}$$

$$= \vec{Y}'\vec{Y} - \vec{b}'\vec{X}'\vec{Y} \qquad \text{eq. 1.27}$$

and where also

$$s^2_{y\cdot x_{11}\cdots x_{hk},\hat{Y}_{i-1_{harvest}}} = \sum_{r=1}^{h}\sum_{v=1}^{k}\left(\sum_{q=1}^{u-1}(X_{irvq} - \overline{X}_{irv})(Y_{iq_{harvest}} - \overline{Y}_{i_{harvest}})\right) \qquad \text{eq. 1.28}$$

$$= R^2_{y\cdot x_{11}\cdots x_{hk},\hat{Y}_{i-1_{harvest}}}. \qquad \text{eq. 1.29}$$

Since the size of the confidence interval about a given wheat yield estimate is one convenient statistical measure of the relative predictive value of that estimate, a convenient representation of the CI should be given. It is known as the confidence interval half-width. For standardization purposes it is necessary to express the confidence interval half-width as a percent of the estimate. Thus we have

$$\underset{\text{half-width}}{\text{CI}} = 100 \cdot \left[\left(t_{(d.f.,\ 1-\alpha/2)} \widehat{(V(\hat{Y}_{iq_{harvest}}))}^{\frac{1}{2}}_{1} \div (U-1) \right) \div \hat{Y}_{iq_{harvest}}\right] \qquad \text{eq. 1.30}$$

as a measure of performance of the model.

Computation of CI half-width for given X_j (independent variable) combinations will show the sensitivity of a time period i model to given predictors for a given region of the world.

Part III. Confidence About the Predicted Yield Given X_j with Error

The third measure of model performance is defined similarly to that of Part II. The difference for this case is that the X_j (independent variables) are measured with some error, expressed as

$d = X - \xi$, ξ equal to the true value of the variable.

If $\nu = Y - \eta$, η equal to the true value of Y, and d and ν are independent and normal with respective variances of σ_d^2 and σ_ν^2, then we may determine $\widehat{V(\hat{Y}_{iq_{harvest\ 2}})}$. Keeping (1962) shows that by using the method of grouping $V(\beta_j)$ may be estimated by

$$\widehat{V(\beta_j)} = \widehat{V(\sigma_\nu^2 + \beta_j^2 \sigma_d^2)} = s_y^2 - 2\beta_j s_{x_j y} + \beta_j^2 s_{x_j}^2 \qquad \text{eq. 1.31}$$

here $\hat{\beta}_j = \dfrac{\bar{Y}_{i3} - \bar{Y}_{i1}}{\bar{X}_{irv3} - \bar{X}_{irv1}}$ where the $Y_{iq_{harvest}}$ and the X_{irvq} (for given irv) have been grouped into three classes according to the relative value of the X_{irvq}.

Then expanding and substituting into formulae given by Steel and Torrie (1960) we have for q = u

$$\begin{aligned}\widehat{V(\hat{Y}_{iq_{harvest\ 2}})} &= s^2_{y \cdot x_{11} \cdots x_{hK}, \hat{Y}_{i-1\,harvest}}(1) + s^2_{y \cdot x_{11} \cdots x_{hK}, \hat{Y}_{i-1\,harvest}}\left(\frac{1}{u-1}\right) \\ &+ \left[\sum_{j=1}^{KK}\left(s_y^2 - 2\hat{\beta}_j s_{x_j y} + \hat{\beta}_j^2 s_{x_j}^2\right)\right](X_{ijq} - \bar{X}_{ij})^2 \\ &+ 2\left[\sum_{j=1}^{KK}\sum_{\ell=2}^{KK}(s_y - \hat{\beta}_j s_{x_j})(s_y - \hat{\beta}_\ell s_{x_\ell})\right] \\ &\cdot (X_{ijq} - \bar{X}_{ij})(X_{i\ell q} - \bar{X}_{i\ell}) .\end{aligned} \qquad \text{eq. 1.32*}$$

The confidence interval is thus

$$\hat{Y}_{1q_{harvest}} \pm t_{(d.f.,\,1-\alpha/2)} \sqrt{\widehat{V(\hat{Y}_{iq_{harvest}})}_2} \div (U-1), \quad (u-1) - KK - 1 \text{ d.f.} \qquad \text{eq. 1.33}$$

* Footnote: Manipulation of $\hat{Y}_{i-1\,harvest}$ is handled similarly to equation 1.25 with respect to notation.

From this interval a confidence interval half-width for $\hat{Y}_{iq_{harvest}}$ may be stated as a ± percent of the estimate for a given α level of probability (see Part II). Computation of this quantity for given X_j combinations would show the sensitivity of a time period model to given predictors and their respective precision of measurement for a given region of the world.

Part IV

The fourth aspect of model sensitivity consists of the comparison

$$D = (\hat{Y}_{iq_{harvest}} \text{ CI half-width})_{X_j \text{ with error}} - (\hat{Y}_{iq_{harvest}} \text{ CI half-width})_{X_j \text{ without error}} \quad \text{eq. 1.34}$$

This equation should be applied to all combinations of statistically significant X_j composing models satisfying e_{iq} assumptions.

Part V: A Measure of Accuracy of the Model

A measure of accuracy for Model Type I can be defined as the root mean square error of the yield estimate about the actually observed value. The accuracy of the model for a given time period i may then be expressed as a percent of the mean observed value as follows:

$$\text{Accuracy of } \hat{Y}_{i_{harvest}} = 100 \cdot \left(E\left[\hat{Y}_{i_{harvest}} - \overline{Y}_{i_{harvest}}\right]^2 \right)^{\frac{1}{2}} \div \overline{Y}_{i_{harvest}} \quad \text{eq. 1.35}$$

The model may be considered inaccurate or biased if an F-test involving the ratio of the mean square for predictions about their corresponding observed values (representing lack of model fit) to the variance of actually observed values indicates rejection of the hypothesis of no significant difference between those two ratio components (after Draper and Smith 1966). That is, the model is biased if

$$\frac{\sum_{q=1}^{u} (\hat{Y}_{iq_{harvest}} - Y_{iq_{harvest}})^2 \div u - KK - 1}{\sum_{q=1}^{u} (Y_{iq_{harvest}} - \overline{Y}_{i_{harvest}})^2 \div (u - 1)} > F_{\alpha}; u - KK - 1, u - 1 \qquad \text{eq. 1.36}$$

where α = a chosen confidence level for $u - KK - 1$ and $u - 1$ degrees of freedom.

From eq. 1.35 we have the mean square error for the yield estimate, $Y_{iq_{harvest}}$, as

$$MSE(\hat{Y}_{iq_{harvest}}) = E(\hat{Y}_{iq_{harvest}} - Y_{iq_{harvest}})^2 \qquad \text{eq. 1.37}$$

It can be shown (Raj 1968) that

$$MSE(\hat{Y}_{iq_{harvest}}) = V(\hat{Y}_{iq_{harvest}}) + (B(\hat{Y}_{iq_{harvest}}))^2 \qquad \text{eq. 1.38}$$

$$= V(\hat{Y}_{iq_{harvest}})(1 + \frac{B(\hat{Y}_{iq_{harvest}})}{\sigma(\hat{Y}_{iq_{harvest}})})^2 \qquad \text{eq. 1.39}$$

where $V(\hat{Y}_{iq_{harvest}}) = \sigma^2(\hat{Y}_{iq_{harvest}})$ and $B(\hat{Y}_{iq_{harvest}})$ represents the bias associated with $\hat{Y}_{iq_{harvest}}$. Substituting into eq. 1.39 the appropriate estimate of $V(\hat{Y}_{iq_{harvest}})$ and its square root, we can solve for $B(\hat{Y}_{iq_{harvest}})$. If we let $\overline{Y}_{i_{harvest}}$ represent the average of observed yield, then bias may be defined as

$$B(\hat{Y}_{iq_{harvest}}) = E(\hat{Y}_{iq_{harvest}}) - Y_{i_{harvest}} \qquad \text{eq. 1.40}$$

The important result to be obtained here (Raj 1968) is that as long as

$$\frac{B(Y_{iq_{harvest}})}{(V(\hat{Y}_{iq_{harvest}}))} \leq 0.1, \qquad \text{eq. 1.41}$$

the foregoing confidence interval sensitivity analysis will be relatively unaffected.

Other methods of detecting model bias, as well as ways to correct it are given in Appendix I.

2.1: Model Type II

A second probabilistic approach to mathematical modelling is that offered by Markovian Process. In essence, this method involves defining an initial state for wheat and then multiplying it by the probability that this wheat state will give rise to a given wheat yield class at harvest time. The mathmatical expression for this model is shown below (Chorafas, 1965).

$$\underset{\substack{\text{(matrix of transition}\\ \text{probabilities)}}}{\begin{bmatrix} P_{11} & P_{12} & P_{13} \cdots P_{1n} \\ P_{21} & & \\ \cdot & & \\ \cdot & & \\ \cdot\cdot & & \\ P_{n1} & P_{n} & P_{n3} \cdots P_{nn} \end{bmatrix}} \times \underset{\substack{\text{state}\\ \text{vector}\\ \text{at}\\ t=1\\ (t_{i=1})}}{\begin{bmatrix} Y_{r=1} \\ \\ Y_{r=2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{r=n} \end{bmatrix}} \qquad \underset{\substack{\text{state}\\ \text{vector}\\ \text{at}\\ t=2\\ (t_{i=harvest})}}{\begin{bmatrix} Y_{s=1} \\ \\ Y_{s=2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{s=n} \end{bmatrix}} \qquad \text{eq. 2.1}$$

P_{rs} are the conditional probabilities for the events Y_s given that for each such event a corresponding event Y_r has occurred during the immediately preceding trial. P_{rs} may be defined by calculations of relative frequencies from historical data. These frequencies are, of course, dependent on past environments for a given time period, that is, $P_{rs} = f(\vec{T},\vec{E},\vec{S},\vec{P},\vec{B},\vec{C},\vec{EM})$ where these vectors where defined in the discussion of Model Type I.

The matrix of transition probabilities is defined as a stochastic matrix if it is square, $P_{rs} \geq 0$, and the sum of each transition state (row) equals one. A stochastic matrix in combination with the initial probability P'_{ro} completely define a Markoff Chain. A requirement for a sequence of states to be considered a Markoff chain is that the transition probabilities must not depend on states earlier than the operand.

To define the initial probability P'_{ro} or state of the wheat system, let Y_r equal a predicted value of yield (i.e., a potential yield class), $\hat{Y}_{t_i}$, at time equals i. Intervals of i might correspond to phenological states of wheat. Let $P_{t_i\ t_{i\,\text{harvest}}}$ represent the stochastic (transition) matrix where $t_{i\,\text{harvest}}$ represents time at harvest with respect to initial time i. Equation 2.1 then becomes

$$\begin{array}{cccc}
\begin{bmatrix} \vec{P}_{t_i \to t_{i\,\text{harvest}}} \end{bmatrix} & \begin{bmatrix} 1 = Y_r = Y_{t_i} \\ 0 = Y_r \neq Y_{t_i} \end{bmatrix} & = & \begin{bmatrix} P_{1s} Y_r = 1 \to n \\ \vdots \\ P_{ns} Y_r = 1 \to n \end{bmatrix} \\
\text{stochastic matrix} & t_i \text{ state vector} & & t_{i\,\text{harvest}} \text{ state vector} \\
\multicolumn{4}{c}{\longleftarrow \text{markoff chain} \longrightarrow}
\end{array} \qquad \text{eq. 2.2}$$

Given $\hat{Y}_{t_i}$ from a regression model as in Model Type I or from some simpler predictor*, multiplication of $\vec{P}_{t_i \to t_{i_{harvest}}}$ $\hat{Y}_{t_i}$ will give the probability distribution of $\hat{Y}_{t_{i_{harvest}}}$, where $\vec{P}_{t_i \to t_{i_{harvest}}}$ is based on past frequency data for the time interval t_i to $t_{i_{harvest}}$.

$\vec{P}_{t_i \to t_{i_{harvest}}}$ gives the conditional probabilities for attaining specific events (in this case specific harvest yield classes) given that for each such event a corresponding event (potential yield class) existed or was predicted to exist at time t_i.

$\vec{P}_{t_i \to t_{i_{harvest}}}$ should be constructed for each initial time i selected in the growth cycle of wheat. The individual $P_{i,r,s}$ will represent the relative frequency at which predicted or potential yield class r at time i will end up in class s at harvest time. The frequency values will be obtained from actual observation of $\hat{Y}$ at time i versus the eventual yield, Y. It follows, similar to Model Type I, that over time the predictive ability of the transition matrices $\vec{P}_{t_i \to t_{i_{harvest}}}$ will be significantly improved as more $\hat{Y}_{t_i}$ versus Y_t observations are made. This improvement process should be greatly enhanced by more accurate assessment of factors controlling growth (hence more accurate $\hat{Y}_{t_i}$) and by measurement of eventual yield provided by satellite data. In the above sense, both model Types I and II

*Example of a simpler model:

Y_{t_i} = actively metabolizing biomass = $b_0 + b_1 X_{ratio}$ eq. 2.3

where $X_{ratio} = \dfrac{\text{reflectance in .005 } \mu\text{m band centered on .8 } \mu\text{m}}{\text{reflectance in .005 } \mu\text{m band centered on .68 } \mu\text{m}}$

.8 μm band indicates actively metabolizing biomass
.68 μm band indicates solar radiation absorption by plant chlorophyll
Reference (Tucker et al, 1973)

are capable of "learning" via feedback from previous data.

2.2: Measure of Performance for Model II

The accuracy of the yield prediction for any given time period i in Model II may be assessed by determining the probability in the $\vec{Y}_s$ vector centered on the eventual observed value $Y_{t_{i_{harvest}}}$. The yield class interval on either side of $Y_{t_{i_{harvest}}}$ is obtained by normalizing the observed yield to its population parameters and then determining the yield class limit that will allow .5 - $\alpha/2$ probability between it and $Y_{t_{i_{harvest_{normalized}}}}$.

The total interval around $Y_{t_{i_{harvest_{normalized}}}}$ will then contain 1 $-\alpha$ probability, where α is a null hypothesis (H_0) rejection level.

To normalize $Y_{t_{i_{harvest}}}$ we proceed as follows.

$$Y_{t_{i_{harvest_{normalized}}}} = z = \frac{Y_{t_{iw_{harvest}}} - \eta}{\sigma_\eta} \qquad \text{eq. 2.4}$$

where

$w = w^{th}$ element of $Y_{t_{i_{harvest}}}$ population.

$$\eta = \frac{\sum_{w=1}^{m} Y_{t_{iw_{harvest}}}}{m} = \text{population mean for a given time period } i \qquad \text{eq. 2.5}$$

$$\sigma_\eta = \left[\frac{\sum_{w=1}^{m}\left(Y_{t_{iw_{harvest}}} - Y_{t_{i_{harvest}}}\right)^2}{m}\right]^{1/2} \qquad \text{eq. 2.6}$$

σ_η = population standard deviation

m = number of members in population

But since m is often much less than 50 we must consider η as a sample mean $\bar{Y}$ and thus must use the t distribution to "t-normalize" $Y_{t_{i_{harvest}}}$. Then

$$Y_{t_{i_{harvest}}}{}_{t\text{-normalized}} = t = \frac{Y_{t_{iw_{harvest}}} - \bar{Y}_{t_{i_{harvest}}}}{S_{\bar{Y}_{t_{i_{harvest}}}}} \qquad \text{eq. 2.7}$$

where

$$\bar{Y}_{t_{i_{harvest}}} = \sum_{w=1}^{m} \frac{Y_{t_{iw_{harvest}}}}{m} = \text{sample mean for a given time period } i \qquad \text{eq. 2.8}$$

$$S_{\bar{Y}_{t_{i_{harvest}}}} = \left[\frac{\sum_{w=1}^{m} \left(Y_{t_{iw_{harvest}}} - \bar{Y}_{t_{i_{harvest}}} \right)^2}{m - 1} \right]^{1/2} \qquad \text{eq. 2.9}$$

$S_{\bar{Y}_{t_{i_{harvest}}}}$ = sample standard deviation

w = w^{th} member of sample

m = number of members in sample

For both z and t note that there is only one $Y_{t_{iw_{harvest}}}$ obtained per wheat life cycle. This situation implies an improvement in estimation of means and standard deviations over time as more $Y_{t_{iw_{harvest}}}$ data become available.

To find the yield class limits containing $1-\alpha$ probability centered on $Y_{t_{i_{harvest}}}$ we must calculate the following confidence interval:

$$Y_{t_{i_{harvest}}} \pm t_{(d.f.,\ 1 - \ /2)}\ S_{\bar{Y}_{t_{i_{harvest}}}} \qquad \text{eq. 2.10}$$

where d.f. = degrees of freedom = m–1.

letting $t_{(d.f.,\ 1-\alpha/2)} s_{\bar{Y}_{t_{i_{harvest}}}} = c$

we can obtain the probability in the $\vec{Y}_s$ vector in the interval $2c$

centered on $Y_{t_{i_{harvest}}}$ from

$$\text{prob} = \sum_{s=Y_{t_{i_{harvest}}} - c = S_{lower}}^{s=Y_{t_{i_{harvest}}} + c = S_{upper}} Y_s \quad . \qquad \text{eq. 2.11}$$

If c is not equal to an integer value then interpolation between interval bounding yield classes in $\vec{Y}_s$ is performed to obtain partial yield class probabilities. If either S_{lower} or S_{upper} lies beyond the possible range of s in $\vec{Y}_s$ then the confidence interval is expanded in the possible direction an amount equal to the interval lost in the other.

The proposed measure of accuracy for Model II can then be stated as

$$\text{Model II Accuracy} = \frac{(1-\alpha) - \text{prob}}{1-\alpha} \cdot 100 , \qquad \text{eq. 2.12}$$

Caution is stressed in the use of the measure as the number and numerical width of yield classes may affect the outcome.

3.1: Model Type III

Difference equations offer a third approach to mathematical modelling. This deterministic procedure is defined by the following model (Smith and Williams, in publication):

$$\hat{Y}_{t_{yield}} = Y_{t_i} + \frac{dY_{t_i \to t_{yield}}}{dt} \cdot (t_{yield} - t_i) \qquad \text{eq. 3.1}$$

where

$\hat{Y}_{t_{yield}}$ = estimated yield at yield time t

Y_{t_i} = yield status (e.g., biomass present) at the end of some previous time period t_i (e.g., a phenophase)

$\frac{dY_{t_i \to t_{yield}}}{dt}$ = rate of change in yield status from t_i to t_{yield}; this rate of change may itself change according to, for instance, the phenophase of wheat.

The solution to equation 3.1 may be obtained by substituting

$$Y_{t_i \to t_{yield}} = Y_{t_i} e^{\sum_{i=t_{i+1}}^{i=t_{yield}} (\alpha_i t_i)} \qquad \text{eq. 3.2}$$

where

α_i = the intrinsic rate of increase during time period t_i

$$\alpha_i = f(\vec{T}_i, \vec{P}_i, \vec{E}_i, \vec{S}_i, \vec{B}_i, \vec{C}_i, \vec{EM}_i) \qquad \text{eq. 3.3}$$

where the function components were defined in the discussion of Model Type I.

Then equation 3.1 becomes

$$\hat{Y}_{t_{yield}} = Y_{t_i} + e^{\sum_{i=t_{i+1}}^{i=t_{yield}} (\alpha_i t_i)} \qquad \text{eq. 3.4}$$

This equation may be expressed as a circuit analogue and then solved in an analogue computer or may be solved via digital computers. Accuracy may be expressed as a percent difference between $\hat{Y}_{t_{yield}}$ and the $Y_{t_{yield}}$ observed evaluated with respect to $Y_{t_{yield}}$. For q=1,u observed wheat life cycles, accuracy may be expressed as a mean value together with a confidence interval for specified α (probability) level.

Appendix I: Correctness of the Model

Recall that in performing the regression analysis the following assumptions were made about the errors (residuals),

$$e_{iq} = Y_{iq_{harvest}\ (observed)} - \hat{Y}_{iq_{harvest}}, \quad q=1,u.$$

These were that the e_{iq} are independent, have zero mean, and a constant variance σ^2. A final assumption necessary to conduct F tests was that the e_{iq} were normally distributed. If the fitted model is correct then the residuals should not exhibit a violation of these assumptions.

If the above assumptions for e_{iq} are not met then bias will be introduced in the yield prediction $\hat{Y}_{iq_{harvest}}$. To determine if the regression model is correct an analysis of the residuals is made. This analysis may be performed firstly by plotting the e_{iq} as follows (Draper and Smith 1966): (1) overall to determine if the residuals resemble a normal distribution with zero mean; (2) versus time, (3) versus $Y_{iq_{harvest}}$, q=1,u, (see also Lee 1969), and (4) versus X_{ijq} (independent variables j for time period i) to determine trends in the resulting pattern of residuals. The trend patterns may be of four main types: (1) residuals lie in a horizontal band indicating a correct model satisfying the e_{iq} assumptions; (2) a widening or narrowing band indicating a non-constant variance and implying that a weighted least squares analysis should have been used; (3) a band of constant width with positive or negative slope indicating the departure from the fitted equation is systematic (e.g., negative residuals correspond to low t's, $Y_{iq_{harvest}}$'s, or X_{ijq}'s, positive residuals to high t's,

$Y_{iq_{harvest}}$'s, or X_{ijq}'s); or (4) an arched band of constant width indicating the need for extra terms in the model (e.g., quadratic terms) or the need for a transformation on the $Y_{iq_{harvest}}$'s before analysis. Caution must be exercised in analyzing the pattern of residuals as more than one trend may be present.

The second major method for analysis of the error terms, e_{iq}, involves statistics T_{pr} (Draper and Smith, 1966) for trend patterns 2 through 4 above. These statistics are:

For trend 2: $$T_{21} = \sum_{q=1}^{u} e_{iq}^{2} Y_{iq_{harvest}}$$

For trend 3: $$T_{11} = \sum_{q=1}^{u} e_{iq} Y_{iq_{harvest}}$$

T_{11} should equal zero if this trend is not present.

For trend 4: $$T_{12} = \sum_{q=1}^{u} e_{iq} Y_{iq_{harvest}}^{2} .$$

VI. PROCEDURE AND STATISTICAL MODEL WHERE GROUND DATA IS AVAILABLE

This section outlines a general ground data collection and sampling scheme that optimizes processing of remote sensing data and ground data by reducing the ground data needed through information gained from ERTS data for the inventory of a single crop.

The models proposed here rely heavily in the first 1 stage on the information extracted from the spacecraft data by both human and computer to provide the desired accuracy of the estimate. The second 2 stage is based on low altitude aircraft photography of sampling units selected using the satellite imagery. Because estimates of yield per acre of wheat are needed, two stages are required to obtain adequate information.

1 The first stage of the model starts with the human stratification of spacecraft imagery. At this point, political and administrative boundaries may also be superimposed on the imagery to define the geographic area of interest. 2 Next, to train the discriminant analysis program, fields identified as wheat by ground data or photo interpretation are located on small-scale photos for extraction from the ERTS digital tapes. [illegible]

The number of training fields required for each crop class depends on the variability of the spectral signature of the wheat present. This variability is caused by such factors as different cropping practices, local soil difference, genetic variations, and local climatic conditions. These fields are identified on and extracted from the spacecraft imagery and supplied as training to the discriminant analysis to obtain a point-by-point identification of wheat for the entire area by strata (as defined by the human interpreter). This provides an initial estimate of the acreage of wheat by strata.

3 The discriminant analysis results must then be sampled in some manner to determine the relationship between the discriminant analysis estimate and the true value or ground estimate of the resource. Sampling units (SU's) are defined by breaking the entire area into rectangular areas which in the case of the ERTS study were based on the coordinate grid generated by the MSS system. The size and shape of the rectangular areas is determined by the accuracy requirements, the change in variability of the estimates by the SU's as their size is changed, the cost of making further estimates on the SU's on conventional imagery.

Because the variance of the acreage and yield estimate appears to be proportional to the ERTS estimate the variable probability sampling will provide the most efficient sampling design. PPS

Variable Probability Sampling Model

Variable probability sampling and the associated estimators is a special case of the "mean of the ratios estimator" where samples are allocated proportional to the expected variance of the X_i estimate. For this model, the total value of the i^{th} SU, denoted by X_i, is evaluated using an indicator function.

$$X_i = \sum_{m=1}^{M} \sum_{j=1}^{J} I_m V_j$$

where $\begin{cases} I_m = 1 \text{ if } C_m = j \\ I_m = 0 \text{ otherwise} \end{cases}$

C_m = crop class for the m^{th} "pixel" (picture element) of the SU, as determined by the discriminant analysis,

M is the number of "pixels" per SU,

V_j is the expected yield per "pixel" of j^{th} wheat class

J = the number of crop classes.

The value (v_j) is assigned to rank the wheat strains and condition classes by their relative yields. Then (n) points are then selected from the list of SU's proportional to their estimated value.

The selected SU's are then carefully transferred to the corresponding high flight photography where precise field size measurements are taken. From high flight images, low altitude images, ground identification and historical data, an estimate of the area of wheat (A_i) in the SU is determined. Yield per acre (Y_i) for the field identified as wheat is made using conventional ground sampling methods.

The total production for the area (T) is estimated using the probability of selection (P_i) and the photo/ground estimate of SU value (Y_i) by

$$T = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{P_i}$$

$$P_i = \frac{X_i}{\sum_{i=1}^{N} X_i}; \qquad Y_i = A_i Y_i$$

The variance of the estimate for T is

$$s_{\hat{T}}^{2} = \mathrm{Var}\ (\hat{T}) = \frac{1}{n} \sum_{i=1}^{n} P_i \left(\frac{Y_i}{P_i} - \hat{T} \right)^2$$

If the photo/ground estimate (Y_i) were perfectly proportional to the remote sensing estimate (X_i), only one ground sample would be needed to determine the proportionality constant. More realistically, however, the number of ground samples (n) for the inventory would be estimated by:

$$n = \frac{Nt^2 \ s_{\hat{T}}^2}{N(AE)^2 + t^2 s_{\hat{T}}^2}$$

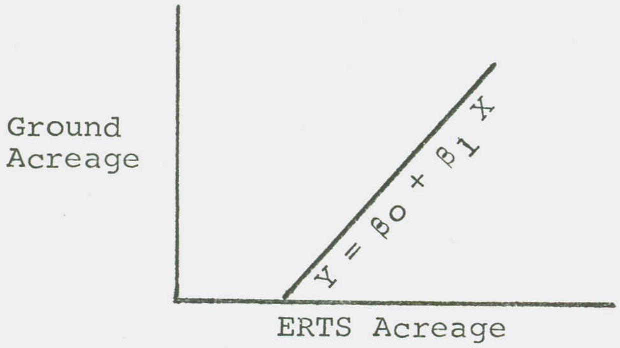
AE = allowable error, in units of wheat production

t = from "Students t" tables and

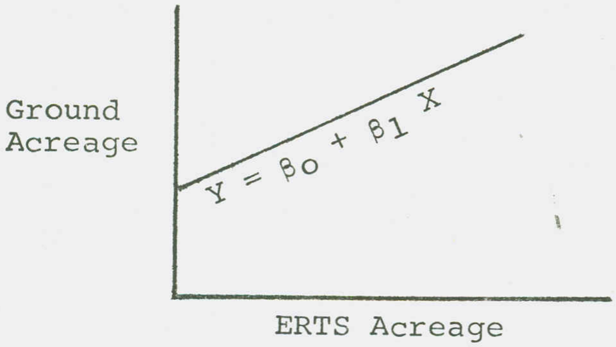
$s_{\hat{T}}^2$ = defined previously

This variable probability sampling model is appropriate when a single parameter such as acreage or yield of a single crop, value of all the crops present, or demand for irrigation water is desired and the source of X_i is proportional to Y_i. It must be replaced by a regression sampling model if the relationship between the ERTS estimate and the ground estimate does not pass through the origin.

It has been found that when this occurs consistently, either confusion classes occur in the same strata, causing acreage to be added (Figure VI.1) when non-wheat is called wheat, or acreage is subtracted (Figure VI.2) when wheat is placed in a non-wheat class.



(Figure VI.1)



(Figure VI.2)

When this occurs the Least Square estimators are appropriate using either model

$$Y_i = \beta_o + \beta_i X_i + e_i$$

where Y_i = photo/ground estimate for the i^{th} sampling unit

X_i = the ERTS estimate for either i^{th} SU

β_o & β_1 to be estimated using the Least Square error procedure based on the n sampling units.

The estimate of the total acreage for the area is then

$$\hat{Y} = \overline{Y} + \beta_1 \left(X^* - \overline{X}\right)$$

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$X^* = \frac{1}{N} \sum_{i=1}^{N} X_i$$

n = number of samples selected

N = total number of samples

Due to the complexity of the estimators for the variance of the parameters estimated and the associated sample size calculation, they are not presented here. The determination of sample size n is similar to the determination of n for the variable probability sampling model described earlier.

SUMMARY

The estimates of precision and accuracy are essential to the final acceptance of the estimates from the production model by the agencies currently producing and utilizing the estimates, and to the peace of mind of the investigators presenting the results. Therefore, the extension of these estimates to analogous areas throughout the world is an important area of study.

The application of these techniques to the wheat production problem will also be essential to obtain the estimates of accuracy and precision acceptable in the model at a cost that will allow effective application of the entire program.

References

Botkin, D.B. 1969. Prediction of net photosynthesis of trees from light intensity and temperature. Ecology 50 (5): 854 -- 858.

Brownlee, K.A. 1965. Statistical theory and methodology in science and engineering. John Wiley & Sons, Inc., New York. 590 p.

Chorafas, D.N. 1965. Systems and simulation. Academic Press, New York. 503 p.

Connor, D.J. 1973. GROMAX: A potential productivity routine for a total grassland ecosystem model. Technical Report No. 208, Natural Resource Ecology Laboratory, Colorado State University, Fort Collins. Part of Grassland Biome, U.S. International Biological Program. 27 p.

Cruff, R.W. and Thompson, T.H. 1967. A comparison of methods of estimating potential evapotranspiration from climatological data in arid and subhumid environments. U.S. Geological Survey Water-Supply Paper 1839-m. 28 p.

Draper, N.R. and Smith, H. 1966. Applied regression analysis. John Wiley & Sons, Inc., New York. 407 p.

Federer, C.A. 1970. Measuring forest evapotranspiration -- theory and problems. U.S. Forest Service Research Paper NE-165. Northeastern Forest Experiment Station, Upper Darby, Pa. 25 p.

Frank, E.C. and Lee, R. 1966. Potential solar beam irradiation on slopes. U.S. Forest Service Research Paper RM-18. Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado. 116 p.

Franklin, J.F., L.J. Dempster, and R.H. Waring (Editors). 1972. Research on coniferous forest ecosystems: first year progress in the coniferous forest biome, US/IBP. Proceedings of a symposium held at Northwest Scientific Association, Forty-fifth Annual Meeting, Bellingham, Washington. March 23 -- 24, 1972. Published by Pacific Northwest Forest and Range Experiment Station, Forest Service, U.S. Department of Agriculture, Portland, Oregon. 322 p.

Freese, F. 1964. Linear regression methods for forest research. U.S. Forest Service Research Paper FPL-17. Forest Products Laboratory, Madison, Wisconsin. 136 p.

Fritschen, L.J. 1966. Energy balance method. Reprinted from: Evapotranspiration and its role in water resources management (Conference Proceedings, Dec. 5 and 6, 1966, pp. 34, 35, 36 and 37), Published by American Society of Agricultural Engineers, St. Joseph, Michigan.

Gates, D.M. 1965. Energy, plants, and ecology. Ecology 46 (1 and 2), 13 p.

Gates, D.M. 1968a. Energy exchange between organism and environment. Reprinted from Biometeorology, Proceedings of the Twenty-Eighth Annual Biological Colloquium, 1967. Oregon State University Press, Corvallis, Ore. 22 p.

Gates, D.M. 1968b. Toward understanding ecosystems. Reprint: Advances in Ecological Research, Vol. 5, pp. 1 -- 35.

Jackson, M.T. and Newman, J.E. 1967. Indices for expressing differences in local climates due to forest cover and topographic differences. Forest Science 13 (1): 60 -- 71.

Johnston, R.S., R.K. Tew, and R.D. Doty. 1969. Soil moisture depletion and estimated evapotranspiration on Utah mountain watersheds. U.S. Forest Service Research Paper INT-67, Intermountain Forest & Range Experiment Station, Ogden, Utah. 13 p.

Jones, J.R. 1971. An experiment in modeling Rocky Mountain forest ecosystems. U.S. Forest Service Research Paper RM-75, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado. 19 p.

Keeping, E.S. 1962. Introduction to statistical inference. D. Van Nostrand Company, Inc., Princeton, New Jersey. 451 p.

Leary, R.A. 1970. System identification principles in studies of forest dynamics. U.S. Forest Service Research Paper NC-45, North Central Forest Experiment Station, St. Paul, Minnesota. 38 p.

Lee, Y. 1969. Reliability on predicted values in regression analysis for forestry research. Information Report BC-X-36, Forest Research Laboratory, Victoria, British Columbia. 17 p.

Leonard, W.H. and Martin, J.H. 1963. Cereal crops. MacMillan Co., New York.

Levitt, J. 1969. Introduction to plant physiology. The C.V. Mosby Company, St. Louis. 304 p.

Mankin, J.B. and Brooks, A.A. 1971. Numerical methods for ecosystem analysis. Publication No. 395, Ecological Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee. Part of Deciduous Forest Biome, International Biological Program. 99 p.

References - continued

McWhorter, J.C. and Brooks, B.P., Jr. 1965. Climatological and solar radiation relationships. Bulletin 715, Mississippi State University, Agricultural Experiment Station. 35 p.

Olson, J.S. 1963. Energy storage and the balance of producers and decomposers in ecological systems. Ecology 44 (2):322 -- 332.

Olson, J.S. 1964. Gross and net production of terrestrial vegetation. Journal of Ecology 52 (Suppl.): 99 -- 118.

Oort, A.H. 1970. The energy cycle of the earth. Scientific American, Fall.

Protter, M.H. and Morrey, C.B., Jr. 1963. Calculus with analytic geometry. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts. 572 p.

Raj, Des. 1968. Sampling theory. McGraw-Hill Book Company, San Francisco. 302 p.

Sauer, R. Unpublished. A simulation model for grassland primary producer biomass and phenology dynamics. Natural Resource Ecology Laboratory, Colorado State University, Fort Collins.

Smith, R.C.G. and Williams, W.A. In Press. Model development for a deferred grazing system. To appear in the Journal of Range Management.

Steel, R.G.D. and Torrie, J.H. 1960. Principles and procedures of statistics. McGraw-Hill Book Company, Inc., New York. 481 p.

Tucker, C.J., L.D. Miller, and R.L. Pearson. 1973. Measurement of the combined effect of green biomass, chlorophyll, and leaf water on canopy spectroreflectance of the short grass prairie. In Proceedings of the Second Annual Remote Sensing of Earth Resources Conference. Given at Space Institute, University of Tennessee, Tullahoma, Tennessee. 27 p.

Van Dyne, G.M. 1966. Ecosystems, systems ecology, and systems ecologists. Publication No. 161, Radiation Ecology Section, Health Physics Division, Oak Ridge National Laboratory. 31 p.

Watt, K.E.F. 1966. Systems analysis in ecology. Academic Press, New York. 276 p.

Wonnacott, T.H. and Wannacott, R.J. 1972. Introductory statistics (Second Edition). John Wiley & Sons, Inc., New York. 510 p.

REFERENCES

W. H. Leonard and J. H. Martin. "Cereal Crops". 1963. MacMillan Co., New York.

Sanderson. "Methods of Crop Forcasting". 1954. Harvard Univ. Press.

M. Y. Nuttonson. "Phrenological Temperature Requirements of Some Winter Wheat Varieties Grown in the Southeastern Atlantic Region of the U.S. and in Several of its Latitudinally Analogous Areas of the Eastern and Southern Hemispheres of Seasonally Similar Thermal Conditions". 1966. American Institute of Crop Ecology.

H. H. Cramer. "Plant Protection and World Crop Production". 1967. Farbenfabriken Bayer Ag. Leverkusen.

S. Martin and W. H. Leonard. "Principles of Field Crop Production". 1967. MacMillan.

M. Y. Nuttonson. "Wheat-Climate Relationships and the use of Phrenology in Ascertaining the Thermal and Photo-Thermal Requirements of Wheat". 1955. American Institute of Crop Ecology.

A. M. M. McFarquitar (ed.). "Europe's Future Food and Agriculture". 1971. North Holland Publishing, Amster.

U.S.D.A. "Major Statistical Series of the USDA -- How They are Constructed and Used". Vol. 8 of Crop and Livestock Estimates, Agricultural Handbook #118.

Foreign Agricultural Service, USDA. "Foreign Agriculture".

Foreign Agricultural Service, USDA. "Russian Agriculture"

"International Maize and Wheat Improvement Bibliography of Wheat". 1971. Scarecrow Press, N.J.

International Wheat Council. "World Wheat Statistics -- 1961-1963".

California Crop and Livestock Reporting Service. "California Field Crop Statistics 1962-1971". Statistical Reporting Service, USDA and Bureau of Agricultural Statistics, California Department of Agriculture.

CHAO. "Agricultural Production in Communist China 1945-1965". 1970. Univ. of Wisconsin Press.

Statistical Reporting Service, USDA. Misc. Publication #967. Dec. 1964. SRS, USDA.

Economic Research Service. "Grain Marketing in the Soviet Union with Emphasis on Wheat". 1961. USDA.

K. G. Brengle and H. G. Sitler. "Linear Programming Yields". Agronomy Journal. 58(6):637-638. Nov/Dec. 1966.

W. C. Johnson. "A Mathematical Procedure for Evaluating Relationships between Climate and Wheat Yields". Agronomy Journal. 51(11):635-639. Nov. 1959.

W. C. Johnson. "Seeding Time Soil Moisture vs. Yield". Agronomy Journal. 56(1):29-35. Jan. 1964.

F. R. Lowell. "The Wheat Market". 1968. Keltner's Statistical Service, Inc.

"1972 Arizona Agricultural Statistics" Arizona Crop and Livestock Reporting Service. April 1973.

"""Investigation of the Feasibility of Determining Yield of Rice, Wheat, and Sugar Cane by High Altitude Photography". Mark Systems, Inc. 1966.

Bureau of Agricultural Economics. "The Wheat Situation". Mar.'62--Nov.'69. Winn and Co. Canberra, Australia.

H. Arakawa. "World Survey of Climatology". 1969. Elsevier Publ.

Foreign Agricultural Service, USDA. "World Agricultural Production and Trade".