MAPPING SURVEY DATA

Ann Cowling, Ray Chambers, Ray Lindsay, Bhamathy Parameswaran, ABARE
Ann Cowling, ABARE, GPO Box 1563, Canberra, ACT 2601, Australia

KEY WORDS: GIS, kernel regression, maps, survey data presentation

Abstract

ABARE conducts large scale annual surveys of the major Australian agricultural industries, collecting detailed production and financial information. Estimates of the main farm performance indicators derived from this survey have traditionally been published in the form of tables. However, they are now presented in the form of maps as well as tables, as maps have been found to be the preferred method of presentation for the majority of ABARE's clients. This paper describes the nonparametric regression method used to calculate the spatial estimates underlying these maps, and also, more briefly, their method of presentation using ARC/INFO, a GIS package.

1. Background

ABARE is the applied economic research organisation attached to the Department of Primary Industries and Energy. ABARE carries out a comprehensive program of economic research and analysis of Australia's agricultural, resource and energy industries. The resulting information is disseminated to government, industry and the community at large, thereby assisting the government and industry in making well informed production and policy decisions.

As an important part of its information gathering activities, ABARE conducts annual surveys of selected Australian agricultural industries. These surveys provide a broad range of information on the economic and physical characteristics of farm business units. Data from the surveys are used to obtain estimates of population means and distributional characteristics, and in econometric modelling of the farming sector.

ABARE's largest survey is the Australian Agricultural and Grazing Industries Survey (AAGIS), which covers farm establishments with an estimated valued of agricultural operations (EVAO) of $A22 500 or more in the last census and engaged in the broadacre industries — that is, farms involved mainly in one or more of cereal crop production, beef cattle production, and sheep and wool production. For the last two years, around 1650 farms have been included in the AAGIS sample, which is stratified by geographic area, industry and estimated value of agricultural operations. The sample farms are irregularly located, and their distribution varies in density throughout Australia. For the past fourteen years, the latitude and longitude of the sample farms has been recorded as a regular part of the collection. It is this knowledge of the location of the surveyed farms that has enabled the survey data to be mapped using the method described below.

Until about eighteen months ago, AAGIS results were presented as tables of numbers showing averages for all Australia, each state, and industries within states. However, the concern of rural industry and government about the combined impact of drought in some areas of Australia and the decline in certain commodity prices highlighted the need for timely and detailed information of regional trends in farm performance. An effective way of presenting this information is to map the regional variation in average farm performance of the surveyed farms. Recent improvement in computing power and the availability of high quality and affordable mapping packages have made this form of presentation a practical alternative to tables.

There is an increasing demand for presentation of data in the form of maps, both from within ABARE and from external clients, as the maps are a successful form of exposition for a number of reasons. First, the data presented in a map are easily interpreted. When presented with too many tables, in contrast, it is very easy for a client to overlook local variations or be 'swamped' by numbers. Next, maps make it easy for a client to relate the geographic variation in one variable with that of another. Finally, a colour map has great visual impact.

In this paper it is shown how kernel smoothing techniques can be used on survey data to produce maps which give a good indication of the local geographic variation of a surveyed variable. Two methods of mapping the smoothed data are discussed, both of which use ARC/INFO, a GIS software package.

2. Estimation and mapping of local averages

At ABARE, kernel smoothing is used to produce maps which show estimates of local averages of farm survey variables, and corresponding maps which show the expectile analogue of the interquartile range at the same set of points. The underlying model is based on the assumption that the distribution of a farm specific variable — say, farm business profit — can be decomposed into a

local distribution (variability in profit due to the geographic region in which the farm is located — for example, due to climate, soil, infrastructure, and to some extent enterprise mix and farm size) and a farm specific effect (the amount by which the farm's profit differs from that of its neighbours due to, for example, differences in management style).

Over the past twenty years, the theory and methods of kernel smoothing have made rapid advances. Recent general references include Eubank (1988), Müller (1988) and Härdle (1990). Kernel smoothing is a type of nonparametric regression in which no assumptions are made about the shape of the true regression function, or about the distribution of the errors except that they are independent.

A brief explanation of kernel smoothing for the case of a one-dimensional location vector L follows. For any observed value l of L, the expected value of the response variable Y is given by the regression function, $f(l)$. If n data points $(l_i, y_i), i = 1, \cdots, n$ have been observed, the regression relationship is modelled by

$$y_i = f(l_i) + \varepsilon_i$$

where the ε_i are independent observation errors with mean zero and variance $\sigma^2(l)$.

A natural choice for the local average at any point l is the mean of the values $\{y_i\}$ of the response variable Y for those observations with locations close to l, since observations from points far away will tend to have very different mean values. The local average is defined as a weighted mean

$$\hat{f}(l) = \sum_{i=1}^{n} W_i(l) y_i$$

where the weights $\{W_i(l)\}$ depend on all the locations $\{l_i\}$ of the sample observations. The weights are constructed using a function K known as the kernel, which is continuous, bounded and symmetric and integrates to one. Various weight sequences have been proposed (Priestly and Chao 1972; Gasser and Müller 1979). At ABARE, Nadaraya–Watson weights (Nadaraya 1964; Watson 1964) are used in the estimator. These weights have the form

$$W_i(l) = \frac{K\left(\frac{l - l_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{l - l_j}{h}\right)}$$

where h is a scale factor known as the bandwidth.

In general, such functions give observations close to l relatively more influence on the local average at this location. Points more than one bandwidth from l are given zero weight by the kernel function, and hence have no influence on the value of the local average.

Where observations are sparse, a fixed bandwidth window may contain few points and the corresponding estimator may therefore have a very high variance. In the 'k nearest neighbours' method of nonparametric smoothing (Benedetti 1977; Stone 1977; Tukey 1977) this is avoided by using a different bandwidth for each point l. The bandwidth at the point l is the distance to the kth nearest neighbour of l, so that there are always exactly k points in the bandwidth window.

This method also partially overcomes a boundary problem. Within one bandwidth of the boundaries of the data (if the bandwidth is fixed), there are fewer observations and the kernel weights become asymmetric. Using the k nearest neighbours method, the bandwidth is automatically adjusted to maintain the number of observations. Other methods of dealing with boundary effects have also been proposed (Gasser and Müller 1979; Rice 1984).

The bandwidth, or the number of nearest neighbours, regulates the degree of smoothness in the estimated function $\hat{f}(l)$ by controlling the size of the neighbourhood around l. If a large bandwidth or a large number of nearest neighbours is chosen, the fitted curve is very 'smooth'. It does not follow the data well, and the estimator $\hat{f}(l)$ may be biased. If a small bandwidth or a small number of nearest neighbours is chosen, the fitted curve becomes very 'rough' and, as already noted, the variance of $\hat{f}(l)$ can be inflated. Choosing the bandwidth (or, equivalently, the number of nearest neighbours) involves a tradeoff between over-smoothing and under-smoothing.

At ABARE a combination of the two methods is used: fixed bandwidth kernel weights are used over the majority of Australia, but where data are sparse, nearest neighbour kernel weights are used.

The kernel weights at the two dimensional locations used for mapping ABARE's survey estimates are product kernels. In general, if the ith farm's location is specified by the d-dimensional location vector $\mathbf{l}_i = (l_{i1}, l_{i2}, \cdots, l_{id})^T$ then the product kernel weights defining the smooth at any location $\mathbf{l}$ are given by

$$W_i(\mathbf{l}) = \frac{\prod_{k=1}^{d} K\left(\frac{l_k - l_{ik}}{h_k}\right)}{\sum_{j=1}^{n}\prod_{k=1}^{d} K\left(\frac{l_k - l_{jk}}{h_k}\right)}.$$

For spatial smoothing based on latitude and longitude, $d = 2$, with the one dimensional kernel weights in the N–S and E–W directions multiplied together to give the final product kernel weight. Non-zero product kernel weights occur in a rectangular window centred at **l**. The number of non-zero weights in the window is counted, and if it is less than a prespecified number, the number of nearest neighbours is automatically increased in both the N–S and E–W directions and the one dimensional kernels and product kernel are recalculated until the number criterion is met. Since the maps are intended to show population, rather than sample, characteristics, the product kernel weights are then multiplied by the survey weights to get the final smoothing weights used for calculating the local average.

As has been noted, bandwidth selection is critical for controlling the smoothness of the surface. At present, the bandwidth is chosen by minimising a penalty function which is the sum of (a) the sum of the squared differences between the regional totals of the smoothed and unsmoothed Y values for a fixed set of 27 agricultural regions which cover Australia, and (b) the summed variance of the smoothed values in each region. The first term (a) in this penalty function gets smaller as the smoothed values approach the unsmoothed values, that is, as the bandwidth decreases. On the other hand, the second term (b) gets smaller as the bandwidth is increased. Minimising this penalty function therefore corresponds to 'trading off' the ability of the smoothed data to replicate standard Agricultural Region estimates against the degree of smoothness of the local average surface fitted to the survey data. There are a number of other methods that could be used for this purpose, including cross-validation and numerous plug-in methods; see for example Härdle (1990).

The smoothing program used at ABARE calls FORTRAN subroutines from S-PLUS (Statistical Sciences 1991). The default output is the local average at each survey farm location, but local averages at any other points can also be produced. In particular, the points on a 50 by 50 rectangular grid can be used. Two versions of a completed map, using survey farm and grid points, are shown in figures 2 and 4 respectively.

The smoothing program contains a module for nonparametric M-quantile regression (Breckling and Chambers 1988) which is used to fit a smooth surface to the expectiles (Newey and Powell 1987) of the Y-distribution at any location. The difference between the smoothed 75th and 25th expectile surfaces (the smooth expectile analogue of the interquartile range) is then mapped to show areas of high and low variability in the data. Not surprisingly, this smooth interexpectile range tends to be highest in areas where the farms are sparsely located and the farm-to-farm variability in Y is therefore highest. The interexpectile range map corresponding to figure 2 is shown in figure 3.

The validity of the independent error assumption can be checked using semivariograms (Cressie 1991). In general, there are indications that after fitting the local average surface there is still some unexplained spatial variation left in the data, which means that the estimated surface is too smooth. It is possible that the model would be improved by allowing the errors to be correlated. However, there is also evidence in the statistical literature that this would not significantly improve the estimates: Laslett, McBratney, Pahl and Hutchinson (1987) found that with their soil pH data, Laplacian smoothing spline estimates, which are similar to kernel smoothing estimates and in which errors are assumed to be independent, outperformed kriging estimates which model the spatial correlation in the data.

Future research at ABARE will include a simulation study of alternatives to the simple product kernel currently used, investigation of the robustness of the procedure against misspecification of farm location and bandwidth, and comparison of other methods of bandwidth choice with the current method.

Because of confidentiality of the survey data, care must be taken in mapping the smoothed data for publication to ensure that the locations of surveyed farms are not thereby revealed. Another requirement is output quality compatible with desktop publication packages. Two procedures for generating the final maps that satisfy these requirements have been developed using ARC/INFO, a GIS package.

In the first method, a Thiessen polygon is calculated around each farm. The polygon defines the area closer to that farm than to any other farm. The farm location is not in the centre of its polygon, and the polygon shape does not resemble the shape of the farm, and so the polygons conceal the locations of the survey farms, as shown in figure 1. The whole of each polygon is coloured according to the smoothed value of Y at the farm location in that polygon. Usually ten colours are used in each map and the

estimated population deciles of the smoothed data are used as boundaries for the colour areas. The maps shown in this paper are black and white analogues of these colour maps.

In the second method smoothed values on a dense rectangular grid are used in place of smoothed values at the farm locations, and a further minor interpolation of the data is carried out in ARC/INFO. A continuous 3-dimensional surface which passes through the smoothed values at the grid points is built in two steps. As a first approximation, a faceted surface formed of triangles obtained by Delauney triangulation is constructed, and then a bivariate fifth degree polynomial is fitted within each triangle using Akima's algorithm (Akima 1978). The resulting continuous surface is then contoured using the estimated population deciles of the smoothed values. Figure 4 is an example.

In this second method of presentation, the locations of the survey farms are not used in any way, thereby completely concealing the location of the survey farms. It also gives smooth contours, and the result is not as patchy as the polygon based maps. Moreover, it is preferred by ABARE's graphics staff because it reduces the number of areas to be separately coloured and has lower storage requirements, enabling the maps to be more readily manipulated in desktop publishing packages. Its disadvantage is that it uses more computing time in the ARC/INFO stage.

Since the above procedures interpolate across all of Australia, including areas where there is no agricultural activity, the final stage of the map production in ARC/INFO is the 'blanking out' of those areas of Australia where there are few or no farms involved in the particular broadacre industry represented by the map. As figures 5 and 6 show, different areas are blanked out for different industries.

Future research at ABARE will include changing the procedure so that there is no interpolation across areas where there is no agricultural activity, and an evaluation of the effectiveness of such changes, using simulated data with and without discontinuities in these areas.

3. Conclusion

At ABARE, nonparametric spatial smoothing has been found to have many applications to data collected in farm surveys, and new applications are continually being developed. The first application of spatial smoothing was in mapping survey estimates, which graphically demonstrated the usefulness of the procedure. Regional differences at a level finer than the state level are currently of interest for allocation of government and other resources. The maps are a very effective tool for demonstrating variation at such a level. Moreover, information presented in the form of a colour map is more readily understood than information in a table, and makes comparison of different regions simpler and faster, as colours seem to be easier to compare than numbers. Maps produced using this technique are now included routinely in ABARE publications.

Following the acceptance of spatial smoothing in maps, it has subsequently been included in other projects, involving for example regional estimates of cost and price elasticities, estimates of farm income risk, and differences in financial performance between farms affected and unaffected by salinity. An area of current research is the application of spatial smoothing ideas to sample weighting. Spatially smooth sample weights should improve small area estimates by reducing their relative standard errors while maintaining control on their biases.

References

Akima, H. (1978). A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points, *ACM Transactions for Mathematical Software* **4**, 148–159.

Benedetti, J. K. (1977). On the nonparametric estimation of regression functions, *Journal of the Royal Statistical Society* **B 39**, 248–253.

Breckling, J. and Chambers, R.L. (1988). M-quantiles, *Biometrika* **75**, 761–771.

Cressie, N.A.C. (1991). *Statistics for spatial data*, John Wiley: New York.

Eubank, R.L. (1988). *Spline smoothing and nonparametric regression*, Marcel Dekker: New York.

Gasser, T. and Müller, H.G. (1979). Kernel estimation of regression functions. In: *Smoothing Techniques for Curve Estimation*, eds. Gasser and Rosenblatt. Springer-Verlag: Heidelberg.

Härdle, W. (1990). *Applied Nonparametric Regression*, (Econometric Society Monograph Series), Cambridge University Press: New York.

Laslett, G.M., McBratney, A.B., Pahl, P.J. and Hutchinson, M.F. (1987). Comparison of several spatial prediction methods for soil pH, *Journal of Soil Science* **38**, 325–341.

Müller, H. G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*, (Lecture Notes in Statistics 46). Springer-Verlag: Berlin.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications* **9**, 141–142.

Newey, W.K. and Powell, J.L. (1987). Asymmetic least squares estimation and testing, *Econometrica* **55**, 819–847.

Priestly, M.B. and Chao, M.T. (1972). Nonparametric function fitting, *Journal of the Royal Statistical Society* **B 34**, 385–392.

Rice, J. A. (1984). Boundary modification for kernel regression, *Communications in Statistics* **A 13**, 893–900.

Statistical Sciences (1991). *S-PLUS Users' Manual,* Seattle.

Stone, C.J. (1977). Consistent nonparametric regression, *The Annals of Statistics* **5**, 595–645.

Tukey, J.W. (1977). *Exploratory data analysis*, Addison-Wesley: Reading, MA.

Watson, G.S. (1964). Smooth regression analysis, *Sankhya* **A 26**, 359–372.

Figure 1 Thiessen polygons constructed around selected ABARE survey farms. Farm location is shown as a small square within each polygon.

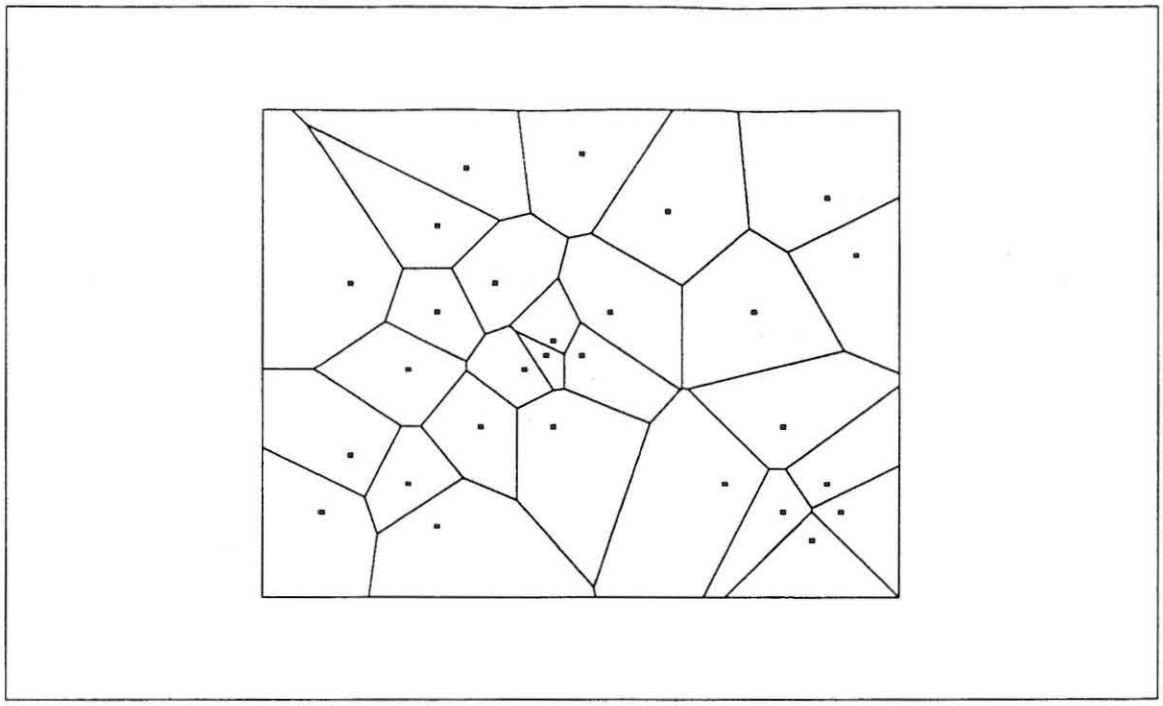


Figure 2 Polygon map of farm business profit in 1991-92, all broadacre farms ($)

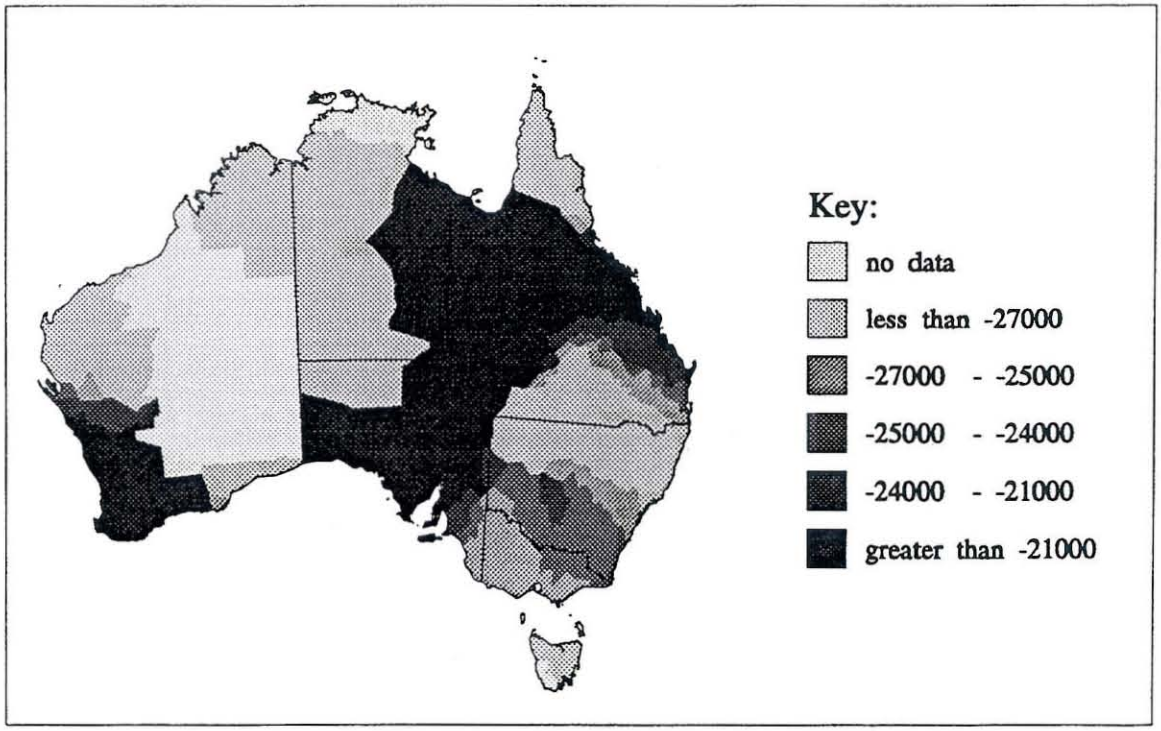


Figure 3 Polygon map of interexpectile range of farm business profit in 1991-92, all broadacre farms ($)

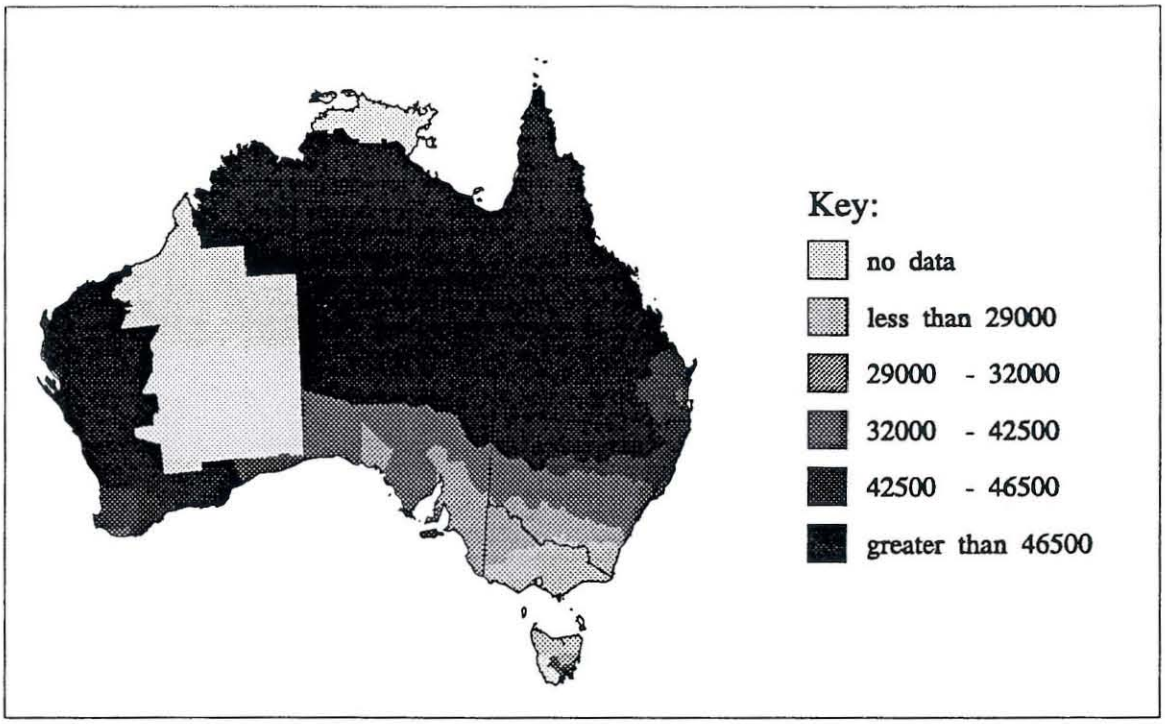


Figure 4 Contour map of farm business profit in 1991-92, all broadacre farms ($)

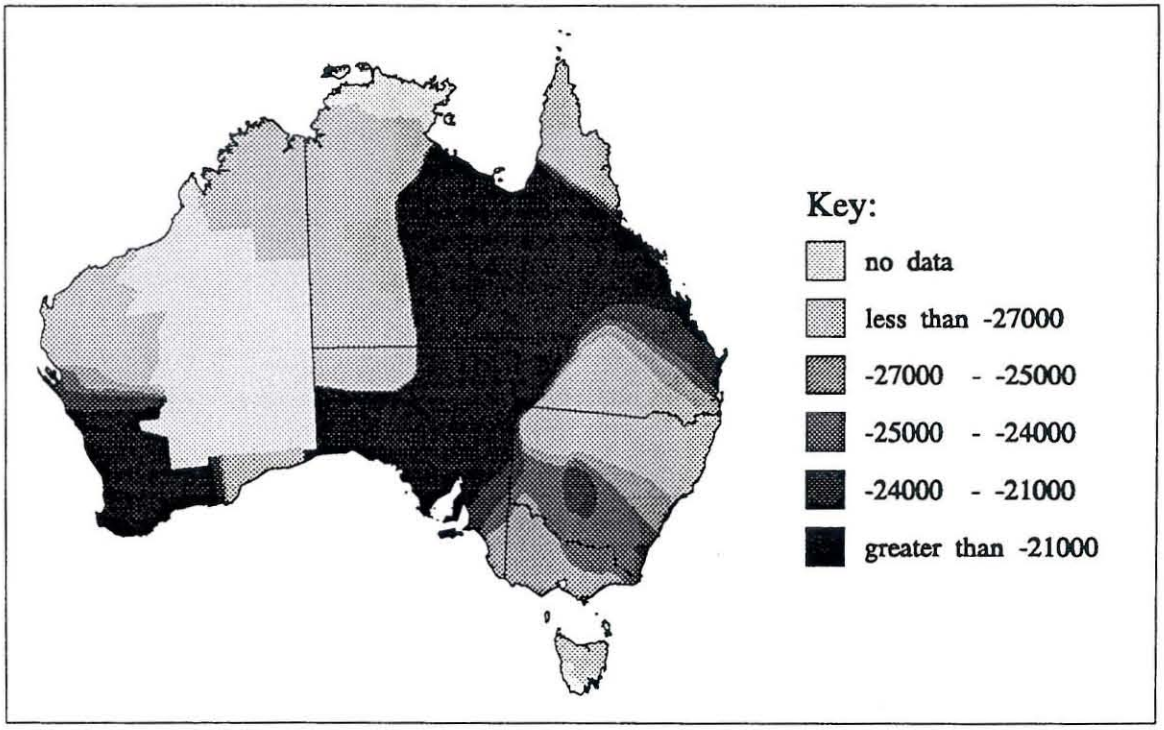


Figure 5 Polygon map showing expected increase in wheat yield from including grain legumes in the crop rotation, 1989-90, wheat and other crops industry (t/ha)

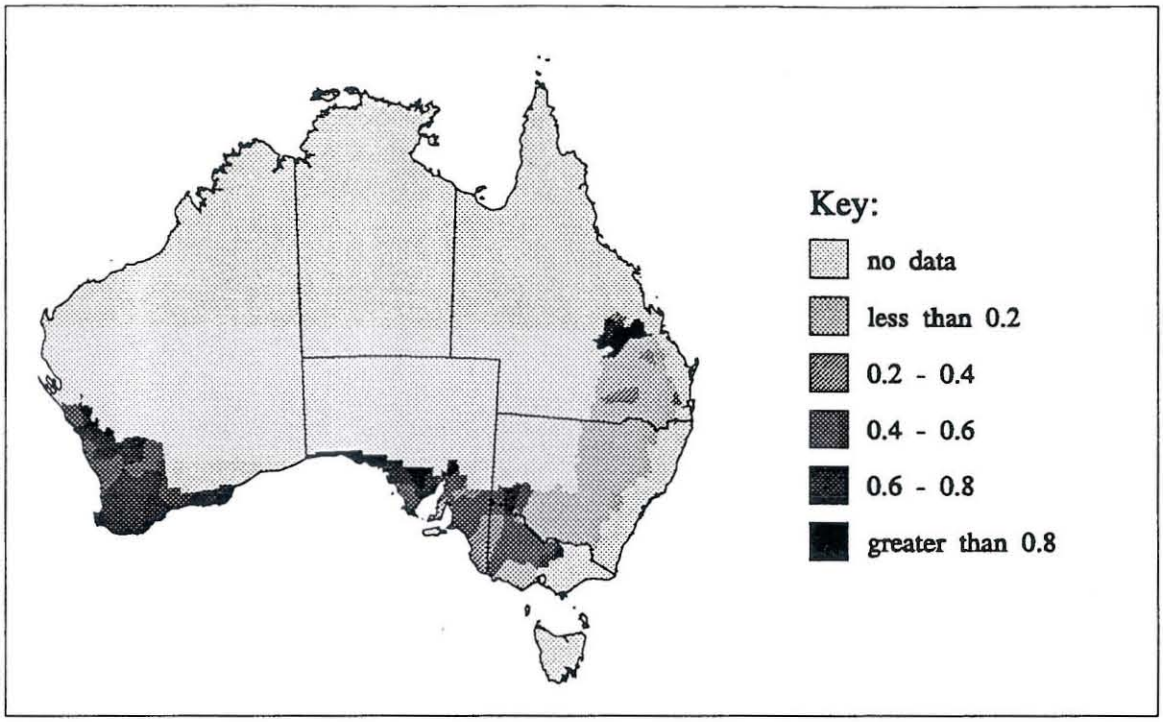
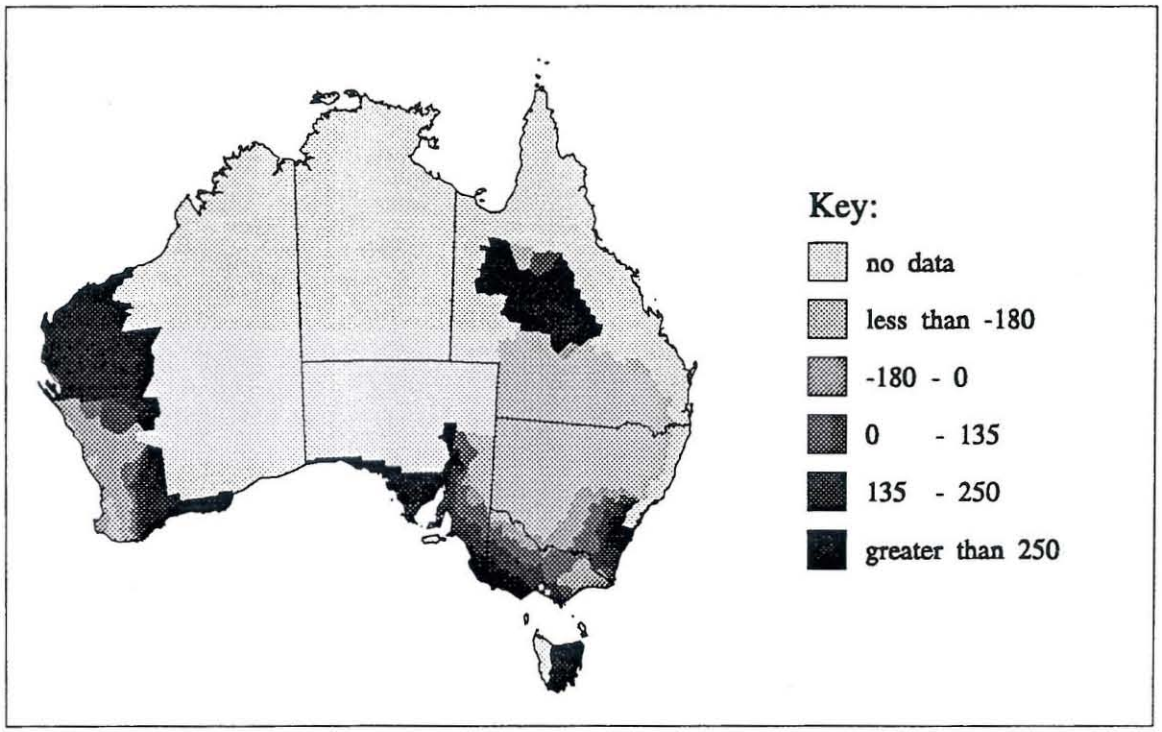


Figure 6 Polygon map showing expected change in wool production, 1991-92 to 1992-93, farms with 100 or more sheep in 1991-92 (kg)



REMOTE SENSING PROGRAM OF THE NATIONAL AGRICULTURAL STATISTICS SERVICE: FROM A MANAGEMENT PERSPECTIVE

George A. Hanuschak and Michael E. Craig, U.S. Department of Agriculture
National Agricultural Statistics Service

I. SUMMARY

The National Agricultural Statistics Service of the United States Department of Agriculture has been utilizing digital earth resource observation satellite data since the launch of Landsat 1 in 1972. There are currently three applied research efforts in the U.S. agricultural statistics program. These are area crop area estimation, crop condition assessment and geographic information system (GIS) utilization for farm chemical and other agricultural survey data. These three research applications are in various stages of development and implementation.

The major research application is the use of Landsat thematic mapper data in combination with area sample frame based ground-gathered data to improve the precision of rice and cotton acreage estimates in the Mississippi Delta region of the U.S. Landsat thematic mapper is a sensor on polar orbiting earth resource observation satellites. The crop area estimates are calculated in an operational timeframe and provided to the Agency' Agricultural Statistics Board as input to the official estimates released by the Agency during the crop season. The well documented aregression estimator approach is used. A contributed paper at this conference authored by Mitch Graham discusses the statistical procedures in detail. The Delta region was selected because of the excellent separation characteristics of rice and cotton from competing spectral land covers and because of the North-South orientation and relatively small growing region compared to the Midwest or Great Plains regions of the U.S. Landsat data is used and regional, state and county level estimates are calculated. In addition county level classification color coded theme map products are provided to the state offices. This project began in 1991 and will be done on an annual basis. The Agency has a long history of similar projects with Landsat Multi-Spectral Scanner Data from 1972-1990. Accurate cost estimates have been kept for the time series 1972-1992 for these projects for cost benefit analysis comparing the new method to the conventional area frame ground-gathered data approach. The statistical measure of performance used is the relative efficiency which is the ratio of the variance of the ground

data only direct expansion estimator (numerator) and the variance of the regression estimator (denominator).

Larger values of the relative efficiency reflect a larger gain due to adding Landsat data into the estimator process. For the 1991 and 1992 Delta project, the average statistical relative efficiency for rice was 3.5 and for cotton it was 3.9. That is, the sample size on the ground would have to be increased by a factor of 3.5 to 3.9 to match the precision of the Landsat-based acreage estimate. These were cost effective improvements in the precision of the State level crop acreage estimates with no additional respondent burden on farm operators.

In addition, county level estimates and crop specific classification (color theme) maps are provided. Statistical methodology for the county (small area) estimator is provided in detail in a contributed paper by Michael Bellow at this conference. The color coded theme maps provide the complete spatial distribution of crops that conventional sample ground gathered data cannot.

The second research utilization of complete spatial and remotely sensed data involves the use of vegetative indices calculated from National Oceanic and Atmospheric Administration's Advanced Very High Resolution Radiometer (AVHRR) sensor. The AVHRR is a sensor on polar orbiting weather satellites. NASS has been slow to get into this area because of its very extensive ground-gathered objective yield forecasting and estimation program already provided excellent information on crop conditions and yields. However, due to the daily satellite passes and the spatial nature of the AVHRR data there is now interest in calculating and mapping vegetative indices similar to the operational program that Statistic's Canada has had since 1988. NASS is currently populating a historic data base of AVHRR data and testing the hardware system to support this type of activity. NASS is using the Land Analysis System (LAS) software from the Earth Resource Observation Satellite (EROS) data center in Sioux Falls, South Dakota and NASA's Goddard Space Flight Center. This program is still in the development stage but the goal is an operational program that would sell on a subscription basis special crop condition assessment data and color map products similar to the Statistics Canada program.

The third and newest area is the use of geographical information systems for providing management additional information about agricultural survey data by taking advantage of the spatial aspects of the data and by overlaying several layers of data such as farm chemical applications, soil types, slope and water flow, crop and land use covers, etc. NASS is in the very early stages of the utilization of GIS based data and related analysis. NASS has procured the ARCINFO GIS software system and also has a Sybase relational data base software system that integrates with ARCINFO. NASS is in the process of populating a farm chemical data base and GIS sample survey data layer at the moment. After completion of these tasks, other layers will be considered as analysis goals and potential become better clarified.

Overall NASS is a fairly extensive user of space based remotely sensed data and related spinoff technologies such as the process of electronic digitization of frame and sample boundaries in its' U.S. Agricultural Statistics Program. However, in relation to the overall NASS mission of providing agricultural statistics on hundreds of items throughout a year, the portion of NASS's

program that utilizes remotely sensed data is not large. The Agency has, however, been able to successfully supplement its' existing probability based (area, list and multiple frame sampling) estimation program by utilizing digital and image space based remotely sensed data for selected geographic areas.

II. CROP AREA ESTIMATION IN THE MISSISSIPPI DELTA REGION OF THE UNITED STATES

NASS staff used Landsat Thematic Mapper Data to operationally calculate improved crop acreage estimates for rice and cotton in the Mississippi River Delta Region in 1991 and 1992. The Landsat Thematic Mapper used in conjunction with area frame based ground-gathered data in the form of a regression estimator. The ratio of the variances, also called the relative efficiency, of the regression estimator and the ground data only direct expansion estimator is the measure of statistical gain from using Landsat Thematic Mapper Data.

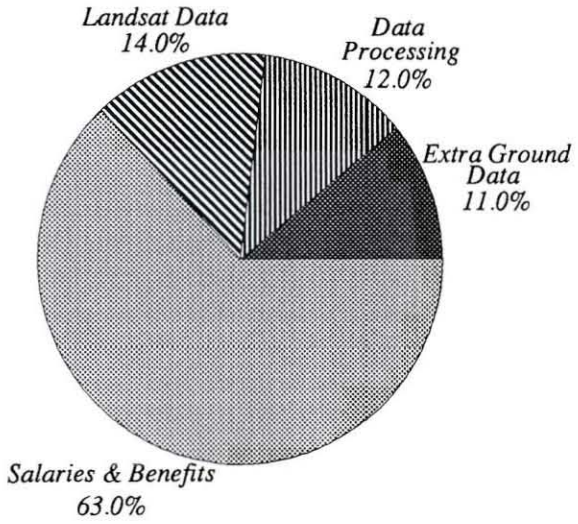
In 1991 and 1992, for rice the average relative efficiency averaged 3.5, for cotton it was 3.9 and for soybeans it was 1.9. The relative efficiency can also be interpreted as the factor by which the ground data area frame sample size would have to be increased by to match the results of the regression estimator. Due to cloud cover and scene availability factors, the Landsat coverage area was divided into both multitemporal and unitemporal analysis regions. In addition, county level estimates were also calculated. Coefficients of variation for the county level estimates for the major rice counties ranged from 3.9 to 10.0 percent. Also, color coded crop classification maps were provided to the State Statistical Offices. The full details of this project are in a recent paper by Bellow and Graham (Aug. 1992).

All the estimates were calculated using the extensive PEDITOR in-house software system developed by NASS and the National Aeronautics and Space Administration Ames Research Center Staff over the years. The PEDITOR system is a quite extensive analysis system for using remotely sensed data in combination with an area frame sample of ground gathered data to calculate regression estimator based crop area estimates and their associated variance. The system has over 100,000 lines of PASCAL code and is used in several other countries around the world. A paper by Jacques Stakenborg (1989) reveals why the European Community's Joint Research Centre chose it over commercial systems for an extensive remote sensing for agricultural statistics project over a ten year period in Western Europe. The main reason PEDITOR was chosen by the European project staff was its efficiency in calculating regression estimates over large land areas. The mosaicing and statistical features are optimized for use with a regression estimator approach. The PEDITOR system's current status has recently been summarized in a paper by Ozga, Mason and Craig (Aug. 1992).

Accurate cost data has been collected and preserved in a data base by project managers since the mid-late 1970's. Thus, NASS has been able to look at Landsat projects (both Multi-Spectral Scanner and Thematic Mapper) from a rudimentary cost/benefit perspective over the years (1975- 1992). The cost side of the equation has been relatively easy to measure. However, as in any cost/benefit analysis the assumptions made about the benefits are a key ingredient to the validity of the analysis. Current total Landsat project costs per State are approximately $175,000. Of the total, 63% is for salaries and benefits, 14% is Landsat data purchases, 12% is all data processing costs

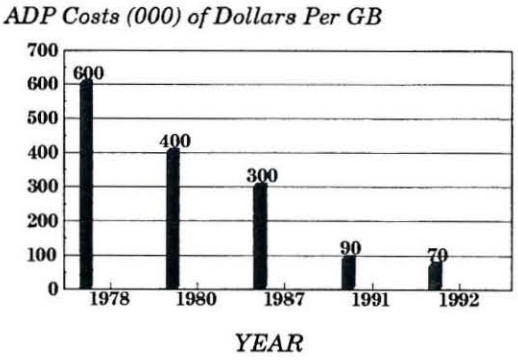
including amortized equipment costs on an annual basis, and 11% is a second visit to ground data sites where fields were not already planted on the first visit (See Figure 1.) In addition, costs per State for the already operational ground survey are approximately $60,000 for the States involved in the project. Landsat project costs have been dropping due mainly to advances in computer technology and in concert with dropping prices for any given level of technology. Ground data collection costs on the other hand are increasing due to inflation in salary, hotel and mileage costs for survey interviewers.

FIGURE 1: Delta Project Costs



When total project costs are compared over time and divided by billions of bytes of input Landsat data processed, the project cost drop is dramatic (see Figure 2.) This was due to two main factors. The first has already been cited as the dropping prices of an ever improving computer technology. The second is staff productivity as more States and land areas were done with a constant number of research staff. With the Landsat Thematic Mapper sensor, it is estimated that a relative efficiency in the 3.0-4.0 range is required to be cost effective. Thus, the results for rice and cotton are judged to be cost effective improvements. This is especially the case since the fairly dramatic improvements in the precision of the State level crop acreage estimates do not add to total respondent burden which is a major concern in U.S. agricultural surveys. The county level estimates and maps with measurable precision are additional benefits. However, the success across years and seasons is still dependent on the degree of cloud cover during the critical crop discrimination windows which are usually only 30 - 40 day windows at best. The probability of success for these projects would be increased substantially by having eight day coverage (two Landsat TM systems) instead of the planned one at a time Landsat 6 and 7 systems.

FIGURE 2: ADP Costs Per Gigabyte (GB) of Import Data (1987-1992)



III. VEGETATIVE INDICES

NASS has recently (last 18 months) begun to explore the possibilities of crop condition assessment utilizing vegetative indices from NOAA's AVHRR data. The Agency has been slow to get into this area because of its very extensive and sound ground-gathered survey data program to forecast and estimate crop yields. The conventional

program utilizes both objective crop counts and measurements such as corn ears, ear length and circumference, field and laboratory weights etc. for each crop plus farmer reported yields. Both types of data have long well established time series and provide a relatively high performing system for forecasting and estimating crop yields. The most comprehensive document of the U.S. system for forecasting and estimating yields was by Huddleston (August 1978). For a current update, the Agency survey manuals and a paper at this conference by Birkett would be the best source.

However, due to the daily satellite passes and the spatial nature of AVHRR, and complete national coverage NASS research staff saw some new potential. In addition, close cooperation with Statistics Canada's Agriculture Division enabled NASS to observe their AVHRR vegetative index program which became operational in 1988. These facts combined inspired NASS research staff to initiate a program. NASS has begun a cooperative agreement with the Remote Sensing Laboratory of USDA's Agricultural Research Service to investigate vegetative indices as related to crop conditions. Condition assessment encompasses such topics as comparison of current year crop(s) growth to previous year(s), comparison of crop growth within a given year between States or counties, and drought and crop disease monitoring. The AVHRR-based Normalized Difference Vegetative Index (NDVI) produced biweekly by the EROS data center will be specifically evaluated. Early research in crop condition assessment will center on the evaluation of NDVI color line printer plots, building a historic data base of NDVI and on the potential use of the NDVI for yield models. The AVHRR vegetative index data provides virtually complete national spatial coverage every two weeks. The spatial resolution of the data is one square kilometer. Thus, when combined with other geographic boundaries such as State and county in a GIS, many different geographic levels of data aggregation and comparisons are made possible. Tabular and color theme map data, when put in a GIS can be aggregated or displayed by any polygon of interest. Thus, the vegetative index data has potential to be one input variable in crop yield models. The Agency plans to value add to this data by using other existing Agency data sets including area sampling frame strata. The subject of the Agency's area sampling frame is covered in detail in an invited paper by Jeff Bush and Carol House at the conference. The Agency's area sampling frame and Landsat crop specific classifications can be used as masks to narrow down the polygons of interest for the AVHRR vegetative index. The polygons of interest can then exclude non-agricultural land and in come cases, can provide crop specific polygons for input to crop specific yield models. A DEC VAXStation workstation has been purchased for this project; it will utilize a current version of the Land Analysis System LAS software developed by the U.S. Geological Survey and the National Aeronautics and Space Administration's Goddard Space Flight staff.

IV ENVIRONMENTAL DATA AND GEOGRAPHIC INFORMATION SYSTEMS

The newest major addition to the Agency's survey program are farm chemical application data in various forms. Survey programs have been designed and implemented (1989 - current) to measure farm chemical applications at the farm level and at the individual field level on a sample survey basis. As part of the U.S. President's Water Quality and Food Safety Initiatives, NASS has become the

surveyor of farm applied chemicals. As part of these initiatives, the tasks of putting these data in a data base and into a geographic information system were also assigned to NASS. NASS has utilized SYBASE (a UNIX based relational data base system) and ARCINFO (a GIS system) as the software to provide the necessary platforms for storing, retrieving and analyzing the sample survey farm chemical data. Data at the published level and micro data will soon be entered into these systems. Confidentiality of farmer reported data will be strictly protected as only use for official government statistical purposes will be allowed and individual data will not be revealed in any form of publication. In addition, a small pilot project was initiated to look at Global Positioning Systems (GPS) recorders for getting accurate coordinates of field locations. A recorder was used to label points within several sample segments in Ohio. This technology, as reported by many other applications scientists, seems to meet most accuracy needs. However, the up front capital investment in equipment, software, training, etc. was judged to be too high for current Agency applications. However, as costs continue to drop, the GPS technology holds substantial promise for several Agency applications such as GIS, area frames, etc.

V. IN HOUSE COMPUTER SYSTEMS

To service these requirements, a wide range of microcomputer technologies are interfaced in- house. Large volume remote sensing analyses are performed on a VAXCLUSTER of a MicroVAX 3500 and a VAXStation 3100. Other technology research applications, such as GIS, are performed on a UNIX system which utilizes a SUN 4/380 server with SPARC and SUN IPC workstations (both stand alone and client server forms). Both servers have a 9-track tape and Exabyte tape cartridge capabilities in addition to several disk drives and other peripherals.

Smaller volume analyses utilize 386 and 486 personal computers as stand alone and/or client workstations to both the SUN and VAX servers. All servers, workstations and personal computers are connected together on an ETHERNET network using Network File Server, DECNET and TCP/IP protocols. Peripheral equipment includes high resolution color monitors, printers, scanners, video cameras, and digitization tablets. Other equipment includes laptop and notebook computers, such as GRID Pads and Zenith Supersports and Zeos

VI. LOOK TO THE FUTURE

The future of all three of the efforts described in this report crop acreage estimation using a regression estimator, vegetative indices, and geographic information systems is bright concerning the technology aspects. The pressure will be on economic factors and showing cost effective improvements or new products in the budget decision time schedules and framework.

As far as the technological aspects, the U.S. Government has recently increased its commitment to future Landsat's 6 and 7. The U.S. Government is firmly supportive of the NOAA/AVHRR program. It is currently funded to the year 2005. NASS is firmly supportive of area frame sampling as its statistical foundation to complete universe coverage without duplication in the frame. NASS of course, complements this with list and multiple frame sampling as well. NASS staff are also investigating panel surveys calibrated to the universe as a potential path to reducing total respondent burden. Geographic Information Systems are

proliferating throughout the public and private sectors on a worldwide basis.

The one down side on sensors is that for forecasting and estimating a dynamic event like crop production frequent satellite coverage is required. One Landsat TM or enhanced TM at a time, only gives 16 day coverage. For acreage estimation with the regression estimator, optimum classification windows are often only 30-45 days in length. Usually, that gives only 2 or 3 chances to get data during the optimum window. If those 2 or 3 chances are substantially cloud- covered, then the statistical gains of the regression estimator can drop dramatically. NOAA/ AVHRR gives daily coverage but with much different resolution than Landsat TM or SPOT. Thus, it enables large scale looks at the vegetative indices across time but doesn't provide a vehicle for estimating acreage accurately compared to ground-gathered data systems. Perhaps some private sector systems could be developed to better meet agriculture's needs.

The challenge will be to speed up the R&D process as much as possible to evaluate if cost beneficial application of these various technologies is appropriate under most likely declining budgets. Substantial progress has been made but work remains. The U.S. and other government commitment to space borne sensors seems to be at a quite healthy stage. The next 5 - 10 years will be crucial to complete R&D, and to apply the technology where it makes sense in a cost effective manner.

In addition, new sensors such as several nation's radar based systems and NASA's Earth Observing System Data and Information System (EOSDIS) will be new systems of data to evaluate. It is difficult to envision that preciously few research resources in NASS can address new sensors as well as current sensors. NASS staff will observe other research such as European and Canadian research on radar systems for agriculture and land cover and NASA research on EOSDIS. Radar sensors overcome the cloud problem but also have different characteristics and require different processing methods. If substantial demonstration of potential cost effective improvements are completed, then NASS research staff would re-evaluate its resource allocation. However, given current resource availability and NASS applications, we will continue to focus on Landsat TM, for crop acreage NOAA/ AVHRR vegetative index for crop condition, and geographic information systems especially related to environmental data such as farm chemical data. In fact, it will be a serious challenge to even address these three applications appropriately under cost and staff constraints.

VII.ACKNOWLEDGEMENTS

The authors sincerely acknowledge the contributions of many persons associated with this effort. Especially those units and staff listed here: Remote Sensing Section Staff Technology Research Section Staff Various Functional Unit Staff Members Throughout the Agency State Statistical Offices (Arkansas, Mississippi and Louisiana) Agricultural Research Service Staff USDA Remote Sensing Coordinator Environmental Statistics Staff

VIII.REFERENCES

Allen, J.D. and Hanuschak, G.A., 1988, "The Remote Sensing Applications Program of the National Agricultural Statistics Service: 1980 - 1987," U.S. Department of Agriculture, NASS Staff Report No. SRB-88-08

Battese, G.E., Harter, R.M. and Fuller, W.A., 1988, "An Error-Components Model for Prediction of County Crop Areas using Survey and Satellite Data," Journal of the American Statistical Association,83(401): 28 - 36.

Bellow, Michael and Graham, Mitchell, 1992. "Improved Crop Area Estimation in the Mississippi DeltaRegion Using Landsat TM Data," American Society of Photogrammetry and Remote Sensing Convention, Washington, D.C.

Bellow, M.E. and Ozga, M., 1991, "Evaluation of Clustering Techniques for Crop Area Estimation usingRemotely Sensed Data," American Statistical Association 1991 Proceedings of the Section on Survey Research Methods, Atlanta, GA, pp. 446 - 471.

Caudill, Charles E. and Hanuschak, G.A., 1983. "Management Issues of Integrating Earth ResourceSatellite Data in to the U.S. Department of Agriculture's Domestic Crop-Area Estimation Program,"Semi-Annual Meeting of the Institute of Management Science/Operations Research Society of America,Orlando, FL.

Cochran, William G., 1977. "Sampling Techniques," New York, NY: John Wiley and Sons, Inc.

Cook, P.W., 1982. "Landsat Registration Methodology Used by U.S. Department of Agriculture'sStatistical Reporting Service 1972 - 1982," Washington, D.C.

Craig, Michael E., 1992. "Applications of Advanced Technology for Agricultural Statistics." United Nations, Conference of European Statisticians, Bratislave, Czech and Slovak Federal Republic.

Hanuschak, George, 1977. "Landsat Estimation with Cloud Cover," Proceeding of the 1976 Symposium on Machine Processing Remotely Sensed Data, West Lafayette, Indiana.

Hanuschak, G.A., R.D. Allen and W.H. Wigton, 1982. "Integration of Landsat Data into the Crop Estimation Program of USDA's Statistical Reporting Service 1972 - 1982." Paper presented at the 1982Machine Processing of Remote Sensed Data Symposium, West Lafayette, IN.

Hanuschak, G.A., and K.M. Morrissey, 1977. "Pilot Study of the Potential Contributions of Landsat Data in the Construction of Area Sampling Frames," U.S. Department of Agriculture, Statistical Reporting Service.

Holko, Martin L., and Richard S. Sigman, 1984. "The Role of Landsat Data in Improving U.S. CropStatistics," Paper presented at the Eighteenth International Symposium on Remote Sensing ofEnvironment, Paris, France.

Huddleston, Harold F., 1978. "Sampling Techniques for Measuring and Forecasting Crop Yields,"Economics, Statistics and Cooperative Service, U.S. Department of Agriculture, Washington, D.C.

Johnson, R.A. and Wichern, D.W., 1988, "Applied Multivariate Statistical Analysis," Prentice Hall,Englewood Cliffs, N.J.

Ozga, M., Mason, W.W. and Craig, M.E., 1992. "PEDITOR - Current Status and Improvements," inProceedings of the ASPRS/ACSM Convention, Washington, D.C.

Ozga, M., W. Donovan and C. Gleason, 1977. "An Interactive System For Agricultural AcreageEstimates Using Landsat Data," Proceedings of the 1977 Symposium on Machine Processing of

RemotelySensed Data, West Lafayette, IN.

Sigman, Richard and Gail Walker, 1982. "Use of Landsat for County Estimates of Crop Areas: Evaluation of the Huddleston-Ray and Battese-Fuller Estimators," SRS Staff Report No. AGES 920909, U.S. Department of Agriculture, Statistical Reporting Service.

Stakenborg, Jacques, 1989. "Data Treatment for Crop Statistics," European Communities Joint ResearchCenter. Institute for Remote Sensing Applications Conference on the Application of Remote Sensing to Agricultural Statistics, Varese, Italy.

Winings, Sherman B., 1982. "Landsat Image Availability for Crop Area Estimation." Paper presentedat the Eighth International Machine Processing of Remotely Sensed Data symposium for Applications ofRemote Sensing, West Lafayette, IN.