POLICY ISSUES ASSOCIATED WITH THE UTILIZATION OF GEOGRAPHIC INFORMATION SYSTEMS (GIS) IN THE U.S. NATIONAL AGRICULTURAL STATISTICS SERVICE (NASS)

By

Rich Allen, Associate Administrator, NASS
Ron Bosecker, Director, Research Division, NASS
George Hanuschak, Associate Director, Research Division, NASS

I. ROLE OF GIS IN NASS

1. The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture is responsible for conducting the Census of Agriculture (at 5-year intervals, 1997 is the current Census year) and a very extensive survey program to support the publishing of all official current agricultural statistics for the United States (U.S.). In order to fulfill its complex mission, the NASS staff has utilized numerous statistical and technological tools over the last 135 years of its existence. Some historic tools were: use of rural postal route and postcard surveys, design and use of mechanical crop meters which calculated percent of different crops along rural road routes, construction and utilization of area sampling frames, design and use of objective crop yield survey techniques, construction and use of a list sampling frame, combination of list and area sampling and estimation techniques known widely as multiple frame sampling, use of statistical software systems, development and use of computer assisted data collection techniques such as computer assisted telephone interviewing, and use of space borne remotely sensed data to supplement and aid in area frame construction, crop area estimation and crop monitoring. Even with all these tools added to the tool kit over the decades, the primary source of data remains as farmer and agribusiness reported data.

2. Geographic information systems technology is one of the more recent tools in the tool kit and NASS staff are just beginning to tap some of the potential benefits of this new tool. Another relatively new tool being utilized by NASS staff is data warehousing and extraction software. An important concept is to not get too caught up with a new tool as a potential panacea. As with any new tool, it is important to know what problems it can help solve and also to be trained on the proper use of that tool, as well as all the other necessary tools.

3. The first question to answer is, “What is the appropriate role for GIS technology in the U.S. agricultural statistics system?” NASS research staff began the search for this answer in 1990 by procuring GIS software, hardware and peripherals and also acquiring training in GIS principles and software utilization proficiency. There were two major categories of potential GIS use in NASS. The two categories were: spatial or geographic data displays, and survey sampling and/or estimation.

4. The research staff started with the easier of the two categories which was spatial or geographic data displays. They examined displays for crop monitoring that utilized NASS

sample survey data and estimates, remotely sensed data, and weather data. A number of these displays were shown in last year's paper for this ECE GIS workshop. Also, the vegetative index data maps that are one aid in monitoring crop conditions are located on the NASS Web site at **http://www.nass.usda.gov/research**. Data displays are provided every 2 weeks throughout the crop season and metadata are included as well. The metadata describe the data sources and features as well as strengths and limitations of the data. The data displays are usually quite useful in observing vegetative greenness over large areas and comparisons over time. However, because of spatial resolution (1 sq. km.) and occasional atmospheric interference from clouds or haze, one must be aware of limitations in the data content. Also available from the NASS Web site is the summarized data from the Weekly Weather and Crop Report and Crop Condition Report. These reports are also available on a weekly basis and provide information about crop stage and condition and weather data. The crop portion comes mainly from expert opinion of USDA agricultural extension agents across the country. They submit a weekly report to NASS on crop stage and condition, which NASS staff summarize at the regional, state and sometimes at the sub-State agricultural statistics district level. A recent addition to the Web site is the definitions used for these weekly expert opinion surveys. The vegetative index maps and this weekly expert opinion data complement each other and both have different strengths and limitations. This is just another example of the value of good metadata to data users. These weekly sources of information are supplements to the Monthly Crop Reports, which are based on large statistically representative samples of farmers who report monthly yields to NASS and on objective yield survey plant counts and fruiting characteristics throughout the growing season.

5. In addition to the spatial data displays provided to the public at large, such as the vegetative index maps, some internal displays are created for crop commodity analysts. These displays depict sample survey data at geographic levels such as county, which due to data confidentiality protections can only be used for internal crop analyst review. Some examples of this type of application are: viewing monthly farmer reported yield data from a small sample at the county level and doing month to month and year to year graphic comparisons as well and viewing the weekly crop stage and condition data at the county level.

6. NASS is also using GIS in the construction and maintenance of its area sampling frame. Percent of land cultivated strata are constructed, using digital map files and digital remote sensing data, and are geo-coded. Within a land cultivated strata, blocks of land, called count units, that encompass 10 or so ultimate sampling units, called sample segments, are identified and geo-coded. For count units selected for the sample, segments are broken out within those count units and sample segments randomly selected. There are approximately 17,000 sample segments selected for a June and a Fall Agricultural Survey. For each sample segment, the approximate centroid is used as the geo-coding. A customized special purpose software system developed jointly by NASS and the National Aeronautics and Space Administration (NASA) with GIS like features was used from 1990-1997. Now, the area frame construction and maintenance process is being converted to commercial off-the-shelf GIS and image analysis software. PC ARCVIEW and ERDAS are being used for this task.

7. Recently, PC MAPINFO was procured for the NASS operational units, including the 45 State Statistical Offices. Some training has been conducted and is in progress. This brings us to the second major category of potential GIS applications which would involve sampling and estimation. One of the remaining challenges is associated with geo-coding NASS's list universe and sampling frame. Possible sources of information to geo-code with are: 5 digit postal zip codes, 9 digit postal zip codes, Census TIGER files, commercial digital map files, 911 emergency services routes and household locations, and any specific locational data collected with global positioning systems (GPS) devices. One serious problem is that at best one could get a household location from most of these sources. This, however, doesn't necessarily provide the location of the agricultural output associated with that household. There has been some initial discussion of using GPS devices to locate large economic scale agricultural facilities locations such as large cattle feedlots, buildings used for hog production and poultry production and perhaps large grain storage facilities. Some NASS State Statistical Offices have access to a limited number of such data sets created by State regulatory agencies. At this stage though, there is not a firm plan and/or infrastructure to accomplish this for all NASS offices. Thus, the geo-coding of even a portion of the list frame universe remains a serious challenge. The easiest geo-coding would be to use the centroids of the 9 digit zip codes associated with the mail address households, but the value of that is limited. The mailing address may or may not be in close geographic proximity to the agricultural production of the farm. In addition, some large and even moderate size farms involve numerous land parcels that can be spread out and involve land in several counties or, in some other rarer cases, across States. For example, if one used 9 digit zip information to sample within a watershed, considerable additional data collection would be required to identify the land operated only in the sampled watershed. Also to try to identify land that is in the watershed that wasn't represented by sampling operators by zip code, additional procedures would need to be developed. The 9 digit zip locations could be used for national or State views of certain types of farm operations (i.e., hog farms with greater than 1,000 head of inventory) or of identified samples, internal to Agency analysts. Even in these cases, metadata on the limitations and cautions associated with the accuracy of the locational information for a specified variable of interest should be added.

8. The brief description of GIS applications for NASS provided above leads us into the next two sections of this paper which are: data confidentiality issues and protections required, and statistical defensibility of products provided to the public.

II. DATA CONFIDENTIALITY PROTECTIONS

9. The search for data layers which can be used in GIS analyses have created more interest in NASS geographically located data and challenged the Agency to explore the release of additional data breakouts. NASS is examining all new requests in light of its confidentiality restrictions. In some cases, statistical defensibility is also an issue. NASS cannot release any individually identifiable data about an operation. The general guidelines used to prevent such identification is that no aggregate is published which represents fewer than three reports or for which one operation controls more than 60 percent of the total. The paragraphs below highlight the handling and assumptions that are being used for various data sets.

10. NASS county estimates data definitely meet confidentiality tests and form very logical data layers for land use or productivity studies. They also have the advantage of being available for multiple years so users can examine year-to-year variability or normalize data to an average or standard. County production data are available for most major crops so users can even add crop area figures together to account for more of the total land area. Crop county estimates data, however, do not carry any indication of variation from farm to farm within a county. The prohibition on variation information is done for two reasons. First, providing information on the range of amount of area in a crop within a county or even the range of average yields might reveal information about specific producers. Secondly, county estimates are created through the use of several data sources and a top down from State level, to district level to county approach so they are not simply composed of data from one survey which expand to county totals and yield averages.

11. The one new area of interest in NASS geographic data comes from environmental surveys. Individuals would like to study the relationship of inputs, such as various fertilizer levels and pest control methods on yield. Those studies have involved relatively small sample surveys and it might be easily possible to identify some operations due to size of holdings, combinations of crops grown, or other key data. In these studies, summary data have usually been released only as State level totals or averages since samples were drawn by States. State level data do not provide much sensitivity for GIS analyses. NASS has received numerous inquiries asking for individual information (which will not be provided due to confidentiality) or small geographic detail.

12. After review of the environmental survey results, NASS has created some district (collections of counties within a State) summaries or data sets which had State/district level identifiers only which provide more useful information. However, another common request is to summarize data by hydrogeomorphic regions such as river basins. These regions normally take in multiple counties, often cross State boundaries, and might split counties in some cases. NASS is normally not able to do anything about the split counties, although in some specialized situations all agriculture is known to be in a certain part of a county and it is obvious within which hydrogeomorphic region the data for a county should fall. NASS has done some resummarization of data into these special regions, if doing so would not reveal data for a particular county by comparing district level summaries and hydrogeomorphic regions. At the present time, only a few data sets have been created. Those have not had wide usage since the data had to be combined by using multiple expansion and nonresponse adjustment factors.

13. The interest in environmental data and the close scrutiny by users, ended up identifying one data set that NASS could release as a public use file. The sampling procedure for most field crop chemical use surveys was a probability proportionate to size selection from area frame expanded crop area data. The selection was in essence a selection of random points within the State. No confidential farm level data were used in sample selection and none were collected during the survey process. With no confidentiality restrictions, diskettes of the data were created and offered as an available NASS product. However, no geo-coding below the State level has been retained on the public use file. The first reason for limiting the coding was the chance that identifying a particular chemical being used in a small geographic area might signify one specific

operation.

14. The larger reason for limiting the geo-coding in the public use file was statistical defensibility. State level sample sizes were relatively small, often only 100 per State. Resultantly, primary agriculture districts within a State would have sufficient sample sizes and quite representative data relationships. However, some districts within nearly every State and for each crop would have five or fewer reports, including zero in some cases. Thus, including even district level coding would result in some districts which have actual crop areas would end up with no corresponding cropping practices data at all and other distracts would likely have unrepresentative results based on small sample sizes. If NASS was creating cropping practices estimates, as it does with county yield estimates, it would be a straightforward process to utilize current data, previous year district-to-district relationships, and year-to-year changes at the State level to estimate for each district. However, NASS has not yet come up with a standard approach to use for this public use file. For the present, people using the data sets have been reasonably satisfied to have State level summaries or averages, with the capability to look at variations in chemicals used, fertilizer rates, etc., within the State. If NASS would change the policy on geo-coding within the State, the policy would need to be easy to implement and standard across all crop/State comparisons. The only candidate approach that seems straightforward would be to retain district coding with n or more (with n = 15, 20, or some number determined through sensitivity analyses) reports and code all other reports as "other." This solution would not be a panacea since a specific district might be codeable one year and not the next but it would allow creation of some additional geographically located data for analyses.

15. In the case of being categorized into crop types,

III. STATISTICAL DEFENSIBILITY OF GIS PRODUCTS PROVIDED TO THE PUBLIC

16. The next policy issue topic area is the statistical defensibility of the GIS derived products that are to be released to the public. The products created with GIS need to be informative, protect individually reported data and also be statistically defensible. Thus, the policy issues surrounding statistical defensibility are: defining acceptable statistical measures of confidence for each product, providing good quality metadata about each data layer or source which includes strengths and limitations, and having the management infrastructure to assure policy compliance.

17. As far as defining acceptable statistical measures of reliability to GIS derived products, the picture is mixed. The vast majority of NASS published data are supported by large statistically designed sample survey reports and the Census of Agriculture. Documentation for Ag Census procedures is readily available and many of the sample survey procedures are described in the official releases. Thus, if GIS is used to display such information graphically, metadata is available that describes the statistical procedures used for the underlying data. NASS has established target coefficients of variation and monitors the survey results for acceptable precision levels for its major probability based surveys. One desirable feature would be to design more GIS derived products where at least some of the statistical properties are displayed as well as the variable(s) of interest. Dr. Dan Carr of George Mason University has done some

interesting development in this area for several Federal agencies that may be the prototype for some future NASS graphics.

18. The role and value of good quality metadata plays an important role in GIS and for that matter, an increasing role for Federal statistics products in general. Defining procedures used, sources of the data, strengths and limitations of the data, dates associated with the data, and accuracy or precision levels of the data are helpful for data users or potential data users as they decide whether a given data set is appropriate for their application.

19. The management and systems infrastructure to follow and monitor standards associated with statistical defensibility of the products and high quality metadata are essential.