REMOTE SENSORS—A NEW DATA SOURCE FOR AGRICULTURAL STATISTICS 1/

Earl E. Houseman
Statistical Reporting Service
U. S. Department of Agriculture
Washington, D. C.

Abstract

With reference to agricultural statistics, the processing of remote sensing recordings and the channels of information flow must be in accord with appropriate statistical standards and with the practical uses which develop. In the United States a comprehensive statistical program exists. Existing surveys probably can supply, or be modified to supply, "ground truth" data needed for the interpretation of sensor recordings. To the extent that remote sensing develops into a useful source of statistics, its major role is foreseen as an additional, but very important, source of data that should feed into the existing statistical system and be utilized for the improvement of accuracy and scope of some kinds of agricultural statistics. The reasoning underlying this view is briefly outlined and includes the coordinating and reconciling of data from various sources as a necessary part of the process of preparing forecasts and estimates. The same statistical policy involving matters of confidentiality and standards should apply to all sources of data and the release of estimates.

Introduction

By general nature, people are apprehensive about the collection of data and their use. This is especially true of data to which individuals attach some privacy rights. A wary public becomes increasingly wary as technological developments make possible the collection of information about an individual without his knowing it. The development of computerized data banks and the ability to easily link data from various sources about an individual has contributed to public apprehension. Now, the idea of remote sensors in air or spacecraft for surveying, monitoring, or revealing a wide variety of information is adding another dimension to public concern. The concern is more than a matter of privacy rights. There are questions regarding aggregate data; for example, who benefits most from information on crop and livestock production? Can the dissemination of data be accomplished in an objective, fair and equitable manner? This paper will not attempt to focus on such questions, but simply recognize at the beginning that confidentiality of data pertaining to individuals and fairness in the release of information are major factors bearing on statistical policy.

1/ Prepared for presentation March 3, 1970, in Annapolis, Maryland at the Earth Resources Observations and Information Systems Meeting sponsored by AIAA (American Institute of Aeronautics and Astronautics).

The scope of the author's experience limits this paper primarily to domestic agricultural data, which are collected for statistical purposes. In the United States a comprehensive agricultural statistics program exists, but even so, data from sensors for statistical purposes must be considered as a source of information. Crop and livestock reporting was conceived more than a century ago to provide farmers with information on the supply of agricultural commodities that would enable them to bargain more effectively. The reporting of agricultural statistics has grown to a point where the Statistical Reporting Service of the USDA issues more than 600 reports each year from Washington. Among these are reports covering about 175 crops. They include estimates for individual crops of acreages farmers intend to plant, acres actually planted, acres harvested, production, disposition and stocks. Forecasts of crop production are made each month during the growing season. Livestock and poultry reports include estimates of animals on hand at strategic times during the year. Many of these commodities are subject to monitoring by remote means. However, there are numerous reports where there is no potential at all for remote sensors as a source of information.

Few farmers receive crop reports directly from Washington, D. C., and only limited numbers receive reports directly from State offices. Most rely upon news media, farm publications, and public and private economic analysis services for guidance in their decisionmaking. Through such avenues of communication, farmers and others receive information often accompanied by interpretation that makes the data more meaningful.

It is important in any discussion of data dissemination to distinguish between data for individuals and other data; the latter being data which either do not relate to individuals or are aggregates of individual data so that disclosure of information about individuals is not involved. The term "individual data" is used as a generic expression referring to any data identified with an individual farm, household, business, or organization. In this context, data for a field or a feedlot collected by a remote sensor would not be regarded as "individual" unless the data are linked with a person or organization.

Following a brief discussion of statistical practice regarding the confidentiality of individual data and the release of aggregate data, some views on the amalgamation of data from remote sensors with data from existing sources will be presented.

Individual Data

Statistical reports show averages, totals, distributions, or other aggregates and do not disclose data for individuals. In fact, data are generally collected by statistical agencies in the Federal Government with a specific understanding that individual data are confidential and will not be disclosed to anyone outside of the organization sponsoring the

collection. That practice is pursuant to laws, administrative regulations, or policies of statistical agencies. In essence, statistical agencies charged with responsibility for data collection feel that confidential treatment of individual data is essential in obtaining accurate data and the cooperation of individuals. Much precaution is exercised in the collection, processing, storage and publication of statistical data to guarantee confidentiality. In the USDA the Statistical Reporting Service, for example, has never released to a law enforcement agency any information supplied by an individual without the expressed consent of the individual.

Recently we were engaged in a small research project in a Western State to test aerial photos as a means of counting livestock on rangeland. A farm magazine heard about this project and published an article. The first sentence was, "No hiding cattle and sheep from the assessor with the aerial surveying technique being tested by the USDA's Statistical Reporting Service." This sentence is subject to multiple interpretation and many readers would get the idea of an exchange of information about individual ranch operators between tax assessors and the Statistical Reporting Service. That interpretation is erroneous and is an illustration of publicity that can create a serious problem for a statistical agency which depends upon voluntary cooperation of farmers to supply accurate data.

If information from a remote sensor were received by a statistical agency and linked to individuals, the data after linking would be accorded the same confidentiality that applies to data obtained directly from individuals. The guarantee of confidentiality includes aggregating data sufficiently so that individual data are not disclosed. A statistical agency would not report, for example, the number of cattle on feed in a particular county if it were known that practically all of the cattle on feed in the county were in one feedlot, or even two or three feedlots. Until further development of remote sensing technology, it is probably sufficient to be aware of the practice calling for confidential treatment of individual data.

Policy and Standards for the Release of Aggregate Data

One important factor that influences policy on the release of aggregated data, or estimates, is the immediacy of opportunity for converting the information into economic gain. Statistical reports for some commodities may have an important impact on prices. A person with advance or superior knowledge may have an unfair advantage in acquiring financial gain. Regulations governing the possession and distribution of information should maintain fair and equitable competitive relationships among

individuals. Thus, special policies apply to preparation and release of information that may influence commodity markets and futures trading. A schedule setting forth the dates and hours of release of all regular crop, livestock, and price reports is published each year before January 1 of the ensuing calendar year. This schedule is rigorously adhered to.

The USDA regulations which apply to statistical reports specifically state that ". . . every part of the information utilized in the preparation of such reports, shall be withheld from publication until the day and hour provided for the issuance of the reports . . ." The statutes include penalties applying to employees for release of information prior to the designated time and also for knowingly issuing false statistics. Data classified as "speculative," (i.e., data leading to estimates or forecasts to which markets are highly sensitive) receive special handling. The last phases of the process of generating such estimates are performed by employees in locked quarters who are unable to communicate with anyone on the outside until after the prescribed moment when the report is released. After agricultural statistical reports are released there are no limits as to the availability of the released data.

Persons associated with the development of remote sensing technology and the processing of sensor information should recognize, aside from the privacy problem, that handling of some kinds of data require special arrangements, which are necessary to prevent advance disclosure or unequal access and to avoid putting persons working with "hot" information under unnecessary duress.

In addition to the release of information under conditions assuring fair and equal access, there are other statistical standards that should be recognized. Your attention is called to Budget Bureau Circular No. A-46, Statistical Procedures. Exhibit C of that Circular relates to Standards for the Publication of Statistical Data and provides guidelines on the labeling of data, description of methods used, accuracy of the data, the definition of terms, and other items so the statistics may be fully understood by users.

Thus, good statistical practice includes making information available that defines the parameters being estimated and describes the basis and accuracy of the estimates. The problem of measuring accuracy is a large and complex subject that is beyond the scope of this paper. But, the importance of measures of accuracy should be emphasized because an estimate is of little value unless something is known about the possible range of error to which it is subject. Indeed, one of the generally accepted goals in statistical practice is the achievement of unbiased estimates having known levels of precision. This principle, unbiased estimates with known precision, has a substantial impact on the collection and reduction of data to estimates and should apply to statistics derived from sensor recordings as well as other sources. In addition to an impact on the statistical methodology involved in the reduction of remote sensor recordings, the above principle also affects the design of systems for collecting ground truth which might be needed in the interpretation and processing of sensor recordings.

Error of an estimate is usually measured by statisticians in terms of standard error and bias or a combination of the two. Information about error has many uses and I would like to interject at this point one particular use. There has been some speculation that continuous, perhaps I should say frequent, remote sensor recordings might provide a basis for making crop forecasts earlier in the season or on a more timely basis than is provided in the present agricultural statistics program. The problem of setting an operating time schedule for release of forecasts is complex. Accuracy must play an important part. For example, "When should the first forecast of production of a particular crop be made?" It is possible to make a forecast of yield per acre even before a crop is planted by using historical data on yield adjusted for trend. A forecast should not be made, even if the cost is negligible, until after the date when its accuracy exceeds that of an historical average adjusted for trend. Doing so involves an expense and a disservice. Clearly, one should not release a forecast on July 1 rather than July 15 solely because of a physical capability to do it. Information about accuracy and accuracy standards has an important bearing on the timing of forecasts.

Relationship of Remote Sensing Data to Other Data

Assuming remote sensing develops into a useful source for agricultural statistics, the relationship between it and existing systems must be very close. It is my opinion that sensor recordings should feed into existing statistical systems simply as another source of information for the improvement of some kinds of agricultural statistics. The reason for this opinion is twofold:

(1) For an unforeseeable period of time I believe that remote sensor information cannot be reduced successfully to useful statistics on crop and livestock production with known levels of accuracy without the aid of simultaneous ground information. At least that is the view which should be taken until proven otherwise.

The estimation of crop acreages, yields per acre, and livestock numbers each present very different problems. For crop acreages, simultaneous ground observations from a probability area sample are necessary for the development of the statistical or mathematical models for reducing sensor recordings. We can anticipate that the parameters in the models will require frequent updating and might vary from one set of conditions to another at a given point in time. Determining the models and their parameters for reducing sensor recordings is analagous to calibrating a complicated instrument. Improvements in sensing equipment, the introduction of new crop varieties, and changes in farm practices are obvious reasons for expecting a continuing need for simultaneous ground information on land use for "calibration" purposes. If the sensor recordings are to be used for the preparation of statistically unbiased

estimates of crop acreages having known levels of precision, appropriate ground information from a probability area sample is a necessity that in my opinion will continue indefinitely.

The nature of the crop yield estimating and forecasting problem is very different and appears much more difficult than estimating crop acreages. Clearly, a model for forecasting crop yields cannot be developed without having accurate exogenous data on yields. Moreover, the nature of the statistical inference means several years' data are required to develop and test a model. Again, if useful models are developed there is a continuing problem of keeping the parameters in the models updated and hence dependence on exogenous information such as farmers' reports on yields or a ground system for making counts and measurements in sample fields similar to the one the Statistical Reporting Service now has for some crops.

With regard to livestock statistics, sensors might provide some useful information such as indicating the number and location of feedlots or the number of cattle and sheep on open ranges. It appears that sensor recordings cannot provide complete counts of livestock. Even if it were possible, information on total number only is very inadequate in relation to the needs for livestock statistics.

As the application of remote sensing technology develops, ground observations for individual fields in probability area samples will become increasingly important in the reduction of sensor recordings to useful statistics. Also, such ground surveys are necessary to obtain information on accuracy such as the probabilities of correct crop identification or other measures of reliability. For a period of time the amount of survey data needed is likely to increase rather than decrease in order that full use of the vast amount of information that sensors are capable of recording can be utilized. Eventually the degree of dependence on ground observations might decrease. But, even if remote sensing develops to the point where it can provide estimates independently, the need for sample surveys on crop acreages and yields will not be completely eliminated.

(2) Secondly, sensor data should flow directly into existing statistical systems because statistical information from various sources must be coordinated and differences reconciled. The same policies relating to the collection, handling, and release of information should apply to all statistical data regardless of source.

Remember that the frame of reference is agricultural statistics in the United States. In that context, a foreseeable potential offered by remote sensing technology is the supplying of information that could help establish highly accurate "control" totals. Such collateral information could be very valuable in improving the accuracy of crop statistics. For example, let's suppose that in the interpretation of sensor recordings there is a very low error of misclassification as between corn and other crops so an accurate determination of the total amount of land under corn is available. But, the needs for statistics require much more detail such as the acreage harvested for grain, the acreage of white corn versus yellow corn, the acreage of popcorn, sweetcorn, etc. Many similar illustrations could be cited such as the breakdown of oranges into early, midseason, and late varieties, or tobacco by types. No one knows the full extent of the discriminations that might become possible from sensor recordings. However, for the time being perhaps the most that an agricultural statistician should hope for is an accurate measure of the total land under a crop such as corn which could serve as a control total when utilizing data from existing sources.