Imagery-based Multistage Area Frame Sampling for Assessing Agricultural Activity

R. Cicone [a,*], G. Koeln [b], F. Pont [a], A. Ralowicz [b,*]

[a] ISciences LLC, 2155 Jackson Ave., Ann Arbor MI 48104 USA – (rcicone, pont)@isciences.com
[b] MDA Federal, 6011 Executive Blvd., Suite 400, Rockville MD 20852 – (greg.koeln, andrew.ralowicz)@MDAFederal.com

Abstract – A multistage area frame sampling (MAFS) procedure reliant on multi-resolution remotely sensed imagery is described. MAFS concerns the process of stratification, definition of the principal sampling unit, selection and labeling of samples, and the expansion of sample attributes to the strata. MAFS, at country or continent scales, is enhanced by remote sensing techniques that produce the most accurate estimate for the least possible cost. MAFS also provides statistical measures of confidence in estimated quantities. Satellite imagery provides the basis for automated scene stratification in order to insure optimal sample allocation, for single- or double-stage sampling designs, according to a minimum variance allocation strategy. The method resolves the difficult allocation issue that the sample size (a cluster of image pixels) can overlap multiple strata.

Keywords: agriculture, stratification, sampling, sustainability

1. INTRODUCTION

Estimation of the proportion of land cover features, such as agricultural land or forest land, may be undertaken by mensuration of interpreted remote sensing imagery, for example, via human interpretation or automated imagery classification methods (Chikkara, 1984). However, proportion estimates based on total enumeration are inherently biased due to labeling error. The need to achieve a specific accuracy may drive the requirement for use of very high resolution data in order to maximize labeling performance. This is a conundrum for broad area applications, as resolution will not only drive up the cost of data, but also of processing and analysis. Hence to avoid bias in estimates, strategies are needed to reduce labeling effort and minimize labeling error within project budget constraints. Statistical sampling techniques provide a means to address this need.

The United States Department of Agriculture (USDA), for example, has a long-standing practice of performing agricultural surveys in the U.S. using area frame sampling (Cotter, 1987; Sigman, 1977). The USDA sampling frame is refined on an ongoing basis. Stratification materials include satellite imagery, imagery from the National Agricultural Imagery Program, and an assortment of other aerial photography and thematic maps. Factors considered in establishing strata include the percent of area under cultivation and historical cropland content. The NASS employs a form of double sampling that they call replicated sampling.

Allocation of a manageable number of samples that can be enumerated precisely provides the route to reduce or eliminate labeling bias. However, the trade off is the introduction of another form of error, that is, the result of sampling variance. Nothing is accomplished if error due to sampling overwhelms error due to labeling when employing total enumeration. However, the means to estimate and minimize sampling variance are well understood in the statistical literature (Cochran, 1963). An imagery-based, multistage area frame sampling (MAFS) technique is described in the paper that introduces a novel approach in establishing a sampling frame with low bias, and quantifiable minimum variance characteristics.

The paper describes the procedure and provides examples of its application in two important regions. Application of double sampling framework in China revealed China's cultivated land area was 47% greater than reported government estimates (MEDEA. 1997), a fact independently borne out by an examination of Chinese government documents by the International Institute for Applied Systems Analysis (IIASA) (Heilig, 1998). Secondly, baseline cultivation estimates in Iraq were determined for several crops using a single sample framework (Gardner, 2004).

2. METHOD

The objective of MAFS is to determine the quantity of a material of interest (MOI) in an area of regard, i.e., the study area or area frame — for example, the amount of land cultivated to a specific crop, or set of crops. MAFS provides a statistically robust procedure to estimate a material of interest using a single or double sampling strategy as illustrated in Figure 1. MAFS entails the following basic steps: 1) Define the area frame and material of interest; 2) Create the sampling frame: a) stratify the area frame, b) define the principal sampling unit (PSU), c) allocate the sample, d) label the sample; 3) Estimate the material of interest area; and 4) Validate the result and estimate error.

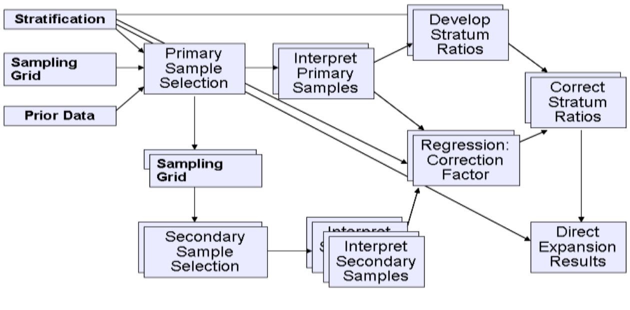

Figure 1. Process for definition of sampling frame, and area estimation using a MAFS double sampling strategy.

* Corresponding author.

MAFS is designed to report the value, any possible error resulting from random sources, and the confidence interval associated with the estimate. While this is basic sampling method, it is in the construction of the sampling frame that MAFS makes a unique contribution. Remote sensing data are employed to create the strata. For example time series MODIS or MIRIS could be used to stratify a large area of regard. Subsequent assignment of samples to each stratum is conducted in an efficient way, accounting for the possibility that the satellite-based PSU could overlap multiple strata. A single stage implementation of MAFS has been automated under ERDAS IMAGINE Frame Sampling Tools (Leica Geosystems, 2003). MAFS procedural fundamentals are described in the following.

Sample Frame.

The area frame defines the study area. The area frame is the extent of area over which the MOI will be estimated. It may be defined as an arbitrary geographic region (e.g., a rectangle bounded by latitude and longitude), a specific zone defined by natural or political boundaries, or a combination of the above. The sampling frame refers to the process of stratification; definition (shape and size) of the principal sampling unit (in the case of MAFS, the principal primary and secondary units); and selection and labeling of samples as illustrated in Figure 2.

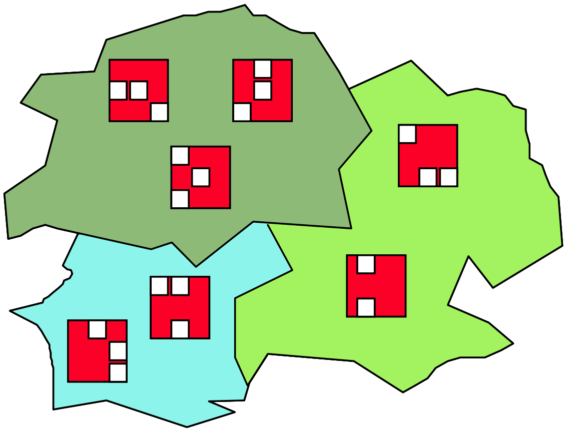

Figure 2. Elements of a sample frame in MAPS: strata (polygons), primary sample unit (red), secondary sampling unit (white)

Stratification

Stratification is a process of grouping homogeneous areas relative to the MOI. The more homogeneous each stratum, the fewer samples required to achieve the desired accuracy. Since each sample may entail expensive action, for example visiting each parcel of land in the sample, or carefully interpreting aerial photography, reducing the number of samples that are needed will reduce overall cost.

A simple, though unrealistic, example will make this clear. If the surveyor is interested in knowing the percent of forestland in an area, the ultimate stratification is to divide the area into two strata, one that contains all the forest, and one that contains all else. The surveyor would then need to sample only one parcel of land to determine how much forest is present. Clearly, such stratification would be fortuitous and unlikely. It is not necessary to stratify, if the surveyor can allocate and analyze a sufficient number of samples to achieve the statistical performance desired. However, it is most often the case that attention paid to this process pays dividends by reducing the number of samples required to achieve a desired level of statistical accuracy.

It is desirable to identify as few strata as possible, while retaining a high degree of uniformity in each stratum. We refer to non-uniformity in a stratum as spatial non-stationarity. This is the spatial analog to temporal non-stationarity, faced in sampling problems where characteristics of the targeted population change rapidly as a function of time.

Imagery based strata files can be constructed using unsupervised or supervised clustering tools. While also advisable to maintain strata of similar size, strata do not have to be contiguous. In fact, one of the advantages of MAFS is its ability to manage highly fragmented strata to the surveyor's statistical advantage. Three important considerations are: the definition of strata that are internally homogeneous, the definition of prior expectations for the MOI appearing in each stratum, and the determination of strata in which the MOI is a rare occurrence.

Primary Sampling Unit

Definition of the PSU, its shape and size, is the next critical consideration in defining the sampling frame. Ideally, in an agricultural survey, the sampling unit would be the patch of ground over which an independent planting decision is made; for example, a farmer's field. Over large areas, randomly allocating sufficient samples of fields that are then surveyed is not practical. MAFS applications in China and Iraq, described later, rely on the use of imagery to identify and label fields, as it makes the overall survey more practical. Hence the PSU may be an image frame, or a cluster of fields.

In this manner, the surveyor is able to label a large number of field-like samples. However, because the fields are clustered together in an image, each field sample is less "independent," that is, less valuable in a statistical sense, than a single parcel allocated randomly. There is a trade-off made when one elects to analyze several clusters of samples, rather than many individual dispersed samples. MAFS accounts for the statistical cost of this trade-off upon execution of its estimation protocol. If labeling error associated with the primary sampling unit is a concern, a secondary sample may be employed to de-bias proportion estimates based on the PSU.

Sample Selection

Once the primary sampling unit is defined, MAFS requires allocation of samples. Random sampling is strictly enforced. If strata and priors are provided, samples are allocated in proportion to the size of each stratum and relative to expected proportion of the MOI in that stratum.

The allocation scheme is according the minimum variance strategy (Cochran, 1963) where the variance of the estimate is given by Equation 1:

$$Var(P_{moi}) = \sum_{s=1}^{nstrata} \left(\frac{A_s}{\sum_s A_s} \right)^2 \frac{P_s(1-P_s)}{N_s} \qquad \text{Equation 1}$$

Where A_s is the area of stratum s, P_s is the prior probability of the MOI in stratum s, and N_s is the number of samples in stratum s.

This approach will tend to allocate more samples to larger strata, and to strata with equal distributions of the material of interest and other materials, thus taking advantage of statistical relationships to provide the lowest variance result possible for a given random sample size. Figure 3 illustrates a possible sample frame configuration using remote sensing to identify strata and optimally allocate cluster samples against those strata. Note that the strata are not necessarily continuous, and samples overlay multiple strata.

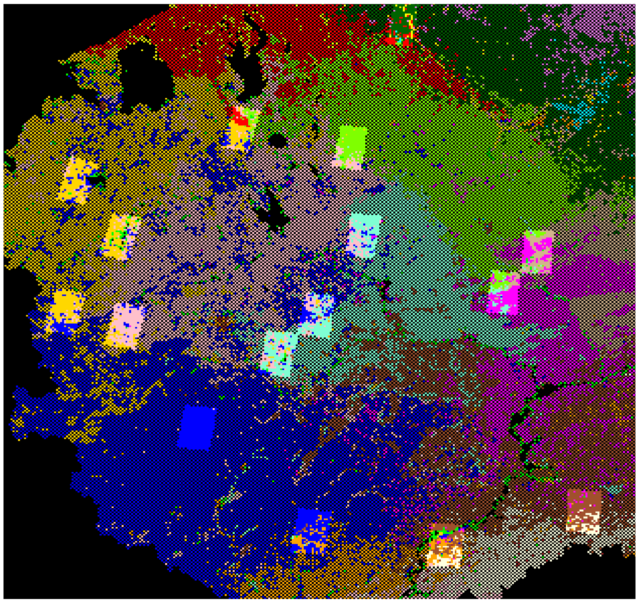

Figure 3. Depiction of a single stage sampling process in MAFS. Strata may be discontinuous reflecting natural groupings of land cover materials. PSUs, clusters of a higher resolution image, may overlap multiple stratum. Sample allocation in MAFS accounts for the fragmentation to obtain an optimum sample (see Equation 1.)

Sample Identification

Once samples are selected, labeling of the materials present in each sample is the next order of business. MAFS provides for two mechanisms, one based on grid labeling and one based on polygon labeling. Labels may be generated from field surveys, or though imagery interpretation.

Estimation method

Once samples are selected and identified, MAFS computes the proportion of the MOI in the study area, and reports statistics associated with that result using a "direct expansion" estimator. Equation 2 is the estimate as a result of direct expansion of the PSU. Equation 3 is the estimate as a result of PSU labeling bias correction using a secondary sampling unit.

The parameter of interest is P_{moi}, the proportion of MOI for the region. The region is partitioned into stratum S = 1, 2, 3, ... , nstrata. Denote the true proportion of MOI within stratum S as P_S. A good stratification can improve the error characteristics of P_{MOI} for a given sample size. P_{moi} is estimated indirectly through the stratum proportions, that is

$$P_{moi} = w_1P_1 + w_2P_2 + w_3P_3 + ... + w_{nstrata}P_{nstrata}$$

where $$w_s = \frac{A_s}{\sum_{s=1}^{nstrata} A_s} \qquad \text{Equation 2}$$

Denote the proportion of stratum S which are assign MOI labels by the classifier as Y_S. If the classification is not perfect there will be a bias which is a linear function of P_S. A second stage sample can estimate a linear bias correction, a_S and b_S, so that $P_S = E(a_S + b_SY_S)$. Denote the bias corrected MOI proportion estimate for stratum S as $X_S = a_S + b_SY_S$. The overall MOI proportion estimate is the weighted bias corrected estimates for the strata:

$$X_{moi} = \sum_{s=1}^{nstrata} w_s \cdot X_s \qquad \text{Equation 3}$$

The mean and variance of X_{moi} are

$$E(X_{moi}) = P_{moi}$$

$$Var(X_{moi}) = \sum_{s=1}^{nstrata} w_s^{\ 2} \cdot \frac{P_s(1-P_s)}{N_s}$$

In summary, the MOI area within each stratum is based on the parcels from each allocated sample that fall within each stratum; the proportion of MOI in the study area is determined by summing the overall proportion based on the weighted contribution of each stratum. Coefficients a, b are derived from secondary samples used to de-bias residual labeling error in the PSU.

Errors – Bias and Variance

MAFS estimates sampling variance using a "bootstrap" strategy (Efron, 1983). A Monte Carlo method is used to estimate variance using the empirical distribution from the allocated sample. This results in an estimate of sampling variance. Errors in the final estimate can result from random sources that affect the variance of the estimate, and structured sources that affect the bias of the final estimate. In either case, an error is an error and the result is a deviation from the "truth". The objective is to make that realized error as small as possible. Repeated application of a procedure that is affected by random errors would provide an average result that has no error. But in any one instance the estimate could deviate from the truth. A procedure with an inherent bias will generate the same error, on average, regardless of the number of times the process is applied. In MAFs project development, the surveyor should avoid designs that structure bias into the process, and attempt to minimize sources of variance. In the absence of egregious labeling errors, MAFS is devoid of structural bias. However, such error can be introduced as a result of bad design and other factors.

3. CULTIVATED LAND IN CHINA circa 2000

In 1997 a US government sponsored scientific team, MEDEA, used a remote sensing based nested area frame sampling approach to determine that China had under-reported its cultivated land base by nearly 50 percent (MEDEA, 1997). The MEDEA estimate has been validated by recent official reports stating that China's cultivated land area exceeds 130 million hectares (M Ha) versus the 90 M Ha previously reported (MEDEA, 2000). The technique, developed by the authors and used by MEDEA, forms the foundation for much of the method and theory employed in MAFS. Figure 4 illustrates the basic elements of the approach. Stratification was based on time-series NOAA AVHRR LAC data for 1992.

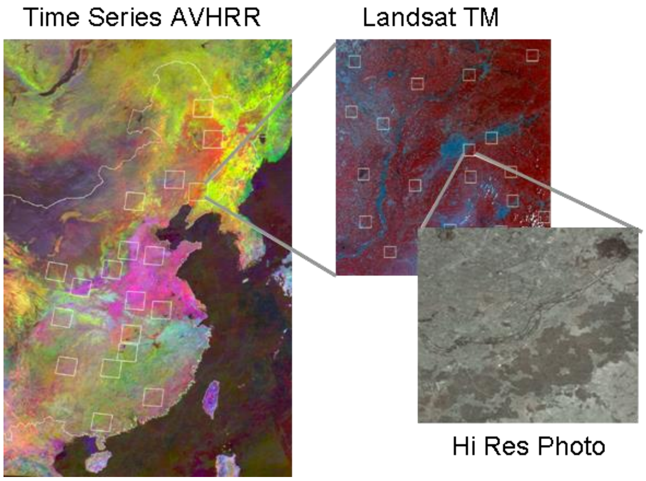

Figure 4. Nested Area Frame Sampling. AVHRR was used to define the area frame. Colors relate to "strata" derived from time-series data. About 22 Landsat "cluster" samples were allocated. Data were categorized as cultivated land and other, using unsupervised clustering and image interpretation techniques. Then about twenty secondary samples were used to correct error in the categorization of each primary sample.

The primary sampling unit was a 125x125 mile Landsat frame. MAFS overcame a difficult problem that the primary sampling unit, a Landsat frame, was composed of multiple strata. A secondary sampling frame based on approximately 10x10 kilometer square optical imagery was used to "correct" classification error in the primary sample. A post-sampling stratification process was used to refine the area frame by eliminating area that did not contain the MOI. MEDEA demonstrated the ability to estimate the statistical characteristics of its result, reporting "application of the estimation procedure resulted in an estimate of 143.4 M Ha circa 1992 to within +/- 5.6% accuracy (at the 0.95 level of statistical confidence)." Later analysis by MEDEA in 2000 adjusted that estimate to 137 M Ha as a result of land degradation and conversion over the intervening five year period.

4. CULTIVATED LAND IN IRAQ circa 2003

The first Gulf War in Iraq, 2001, raised concern over the need to monitor the productivity of the arable land in post-war Iraq. In preceding years, wheat and barley supplied approximately 70% of the calories and 66% of the protein to the Iraqi population daily. In a post-conflict, food-security assessment, these grains have the greatest impact on the population's nutritional and caloric needs and potential social stability. A study utilizing commercial satellite imagery with the Frame Sampling Tools suite in ERDAS IMAGINE was conducted to establish baseline cultivation estimates in post-war Iraq for several crops.

For the Iraq assessment, a single stage approach was employed. The MAFS sample frame for this analysis is depicted in Figure 5. Orthorectified spring 2003 Landsat 7 ETM+ multispectral images were interpreted for agricultural activity. An independent thematic classification of the multispectral imagery was used to stratify the agricultural areas. Orthorectified, pan-sharpened, spring 2003 SPOT 5 imagery of the selected sample sites was used to delineate the following types of agricultural activity: Irrigated Grains; Non-irrigated Grains; Rice; Orchard; Date Palm; Vineyard; Other Agriculture; Fallow; and Abandoned Agricultural Land. The direct expansion of the crop ratios observed in the samples to the country-wide strata produced results that could serve as a baseline for comparing future cultivation activity.

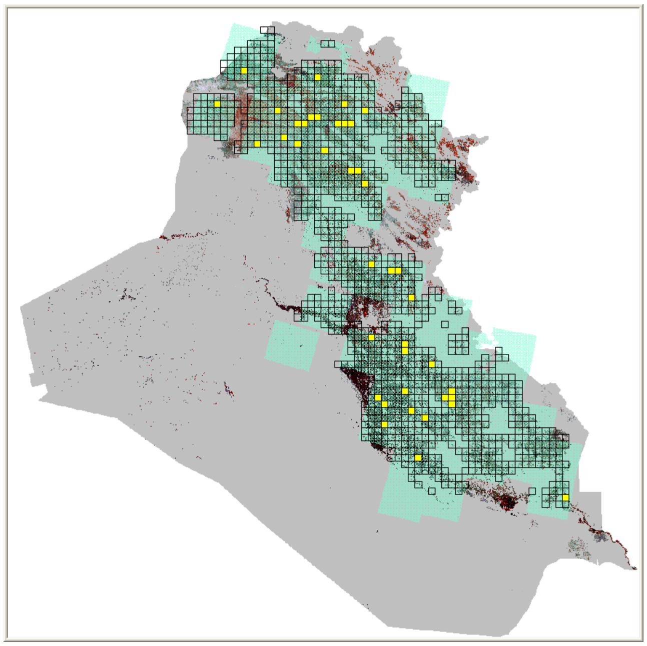

Figure 5. The teal footprints show SPOT 5 images over the 99-Agricultural strata within Iraq, the grid cells represent potential 10km x 10km samples and yellow squares show the 40 sample locations of the SPOT 5 images.

5. CONCLUSIONS

A multistage area frame sampling (MAFS) strategy for wide area mensuration of land covers is described. The strategy is based on classic sampling theory, but takes advantage of the value inherent in overhead imagery to cluster like spectral features into strata that can be exploited to reduce samples needed to achieve a desired accuracy and improve their allocation. Application of MAFS in the two cases described demonstrates its utility. In the first case, analysts were able to determine the area of cultivated land in China and found that it significantly exceeded national reports. In a second case, Iraq agricultural activity post the 2001 Gulf War provided insight into the state of that vital economic resource.

A single stage MAFS has been imbedded in ERDAS IMAGINE Frame Sampling Tools. This toolkit provides a mechanism to apply formal sampling strategies in a variety of mensuration applications. While the focus here has been agriculture, any land cover mensuration effort over broad areas may benefit.

REFERENCES

Chhikara, R.S., C.. Hallum, T.G. Lycthuan-Lee, "Sampling designs for Landsat crop surveys," Journal Communications in Statistics - Theory and Methods, Volume 13, Issue 23 1984.

Cochran, W.G., “Sampling Techniques, second edition,” A Wiley publication in applied statistics, John Wiley & Sons Inc., 1963.

Cotter, J., J. Nealon, “Area Frame Sampling for Agricultural Surveys,” The National Agricultural Statistical Service, U.S. Department of Agriculture, August 1987.

Efron, B., The jackknife, the bootstrap, and other resampling plans, In Society of Industrial and Applied Mathematics CBMS-NSF Monographs, 38, 1983.

Gardner, R., A. Ralowicz, G. Koeln, Imagery-Based Area Frame Sampling for Assessing Post-War Agricultural Activity in Iraq, Proceedings American Society for Photogrammetry and Remote Sensing, Denver, Colorado, 24-28 May, 2004.

Leica Geosystems, 2003. ERDAS IMAGINE® Frame Sampling Tools, Part No. FSTwhitepaper, cc 01/03. http://gi.leica-geosystems.com/documentcenter/ERDASIMAGINE/Frame_Sampling_Tools_White_Paper.pdf

Heilig, G., “Can China Feed Itself: New Evidence from Recent Data on Cultivated Land”, Meetings of the Population Association of America, New York, New York, 1998.

MEDEA, China Agriculture: Cultivated Land Area, Grain Projections, and Implications, Summary Report, Washington, D.C., November, 1997.

MEDEA, China: Agricultural Productivity and Food Security Outlook, December 2000.

Sigman, R., C. P. Gleason, G. A. Hanuschak, and R. R. Starbuck, “Stratified Acreage Estimates in the Illinois Crop-Acreage Experiment,” The Fourth Annual Symposium on Machine Processing Of Remotely Sensed Data, The Laboratory for Applications of Remote Sensing. Purdue University, West Lafayette, Indiana. June 21-23, 1977.
http://www.lars.purdue.edu/home/references/sym_1977/1977_3.2-80.pdf

ACKNOWLEDGEMENTS

The authors wish to acknowledge the following individuals for their contributions toward the development of the MAFS method: Mr. Christopher Chiesa, Pacific Disaster Center, Maui, Hawaii, USA; Dr. Lawrence Chrisman, Griffith University, Brisbane, AU; Mr. David Cunningham, MDA-Federal, Rockville MD; Dr. Michael Goodchild, University of California, Santa Barbara, USA; Mr. Peter Kolasch, Iowa Geological Survey, Iowa City, IA, USA; Mr. Kurt Schwoppe and staff, ERDAS, Atlanta, GA, USA, Dr. Douglas Way, MDA-Federal.