Innovative Uses of Data Mining Techniques in the Production of Official Statistics

Jaki McCarthy, Thomas Jacob, and Dale Atkinson

National Agricultural Statistics Service, US Department of Agriculture
3251 Old Lee Highway, Fairfax, VA 22030
Jaki_mccarthy @nass.usda.gov

Introduction

Data mining techniques are used to find patterns, classify records, and extract information from large data sets. These techniques, often used in the private sector for market research, fraud detection, and customer relationship management, can also be used by statistical agencies to analyze their large survey datasets. While large datasets are common in many statistical agencies, data mining techniques have not been widely used to improve the production of official statistics. However, innovative applications of these techniques can be very effective in efforts to improve survey data, processing and estimation.

Data mining is a general term which refers to a set of several different techniques. These techniques differ, but they all exploit the idea that existing data contain information that can be used in the future. With large datasets this information is often hidden, but data mining techniques can be used to distill or uncover it. Various techniques can be used to classify data into subsets, predict outcomes based on the data, cluster records into like subgroups, or assign propensity scores for some measure to records.

Several data mining based applications have been or are currently being used in NASS, and others have potential future application. To date, the most widely used data mining technique in NASS is the classification or decision tree.

Classification Trees

A classification tree model is constructed by segmenting a dataset using a series of simple rules. Each rule assigns observations to a segment based on the value of one input variable. One rule is applied after another, resulting in a hierarchy of segments within segments. The rules are chosen to maximally separate the sub-segments with respect to a chosen target variable. Algorithms such as the chi-square automatic interaction detection (CHAID) algorithm can be used to determine how to split the segments. The hierarchy is called a tree, and each segment is called a node. The original segment contains the entire dataset and is called the root node of the tree. A node with all its successors is termed a branch of the node that created it. The final nodes are called leaves. NASS used decision tree models for several applications in the 2007 Census of Agriculture. The following paragraphs describe these and other current and potential applications of decision trees in the Agency:

Census Non-response Weighting. Classification tree models were used to divide the 2007 census records into response propensity groups representing weighting adjustment cells (Cecere, 2009). This approach has been used by other survey organizations to do post- survey non-response weighting adjustments (for example, Cohen, DiGaetano and Goksel, 1999), but had not been used previously in the Census of Agriculture. Variables such as operator race and gender, farm type, and size were used to segment operations on the Census mail list (CML) into subsets with homogeneous response propensities. Non-response weights were then generated for each of these groups individually.

Census Mail List Trimming. Classification tree models were also built to identify records on the initial Census mail list (Garber, 2009) that were not likely to represent farming operations. The models used variables such as the source of the record, the length of time the record had been on the NASS list frame, the location of the operation, the previous gross receipts of the operation and other auxiliary variables to identify records for operations with lower probabilities of qualifying as farms. Since the census of agriculture targets only “farms,” as qualifying by the USDA definition, operations that would not qualify are considered out of scope. Prior to mail-out, the records for operations with the lowest probabilities of qualifying were removed from the CML in areas with larger than desired mailing lists, to reduce unnecessary data collection costs and to improve the overall efficiency of the census processing.

Analysis of Reporting Errors. Classification tree models were also used to identify characteristics of operations with specific reporting errors (McCarthy and Earp, 2008). Models were constructed separately for individual types of errors, such as subtotals not equal to their subparts and item non-response. Variables such as the location, type of commodities raised, size of operation, and operator demographics such as race, age and gender were included as possible predictors. The trees generated showed that operations with certain types of land were much more prone than others to the reporting errors in this study. This information will be used to aid in the redesign and testing of these questions. The information may also be useful in designing edits for these items.

Allocation of Survey Incentives. The Agricultural Resource Management Survey (ARMS) is one of the most complex and challenging of NASS' surveys. As such, this survey uses a variety of incentives, both monetary (in the form of Automated Teller Cards) and non-monetary, to encourage response. Classification trees were used to identify characteristics of sample units most likely to respond both with and without incentives (Earp and McCarthy, 2009). This information may then be used to target our finite pool of incentives appropriately – providing them to subsets of respondents for which they'll most likely be optimally effective.

Prediction of Survey Non-respondents. We are also currently developing models using decision trees to predict agricultural survey non-respondents based on several sets of historical data (McCarthy, Jacob and McCracken, 2009). Most of our sample survey units will have completed the mandatory Census of Agriculture, so that information is available for both survey respondents and non-respondents. In addition, we know how often NASS has contacted each sample unit in the past for other surveys and how often they have responded in those contacts. We can also link general information known about the location of the sample, such as demographic and geographic characteristics of the county (population, median income, percent of land in farmland, etc.). The resulting tree identifies characteristics of sampled units most likely to be non-respondents in a sample survey. NASS has no current plans to use this as a weighting adjustment. However, these models can be applied to individual surveys and used to tailor data collection techniques. For example, subgroups of likely non-respondents can be targeted for in person interviews rather than mailed questionnaires or for earlier non-response follow-up.

Cluster Analysis

Another data mining technique that has been applied in NASS is cluster analysis. Cluster analysis is the classification of observations into clusters of like records. The clusters are based on information about all variables identified -- not with respect to a single target variable (as is the case with decision trees). There are numerous methods of deciding whether records are related and forming clusters, but all cluster analysis is based on the principle that the cluster solution will lead to observations within each group being more similar to each other than to observations in other groups. Cluster analysis has been used in the following applications in the Agency:

2007 Census Donor Pool Screening. The 2007 Census of Agriculture editing process required creating and seeding a donor pool of records for imputation. The initial donor pool was seeded with 2002 Census of Agriculture data. For the 2002 Census, questionnaires were scanned, and data were captured through Optical Character Recognition (OCR). The process produced scanned images of the questionnaire as well as capturing cell content as numbers and/or letters. However, the OCR introduced an unacceptable number of errors which analysts had to manually correct. For example, the OCR program generated a "1" or "11" when respondents crossed out an entire section in the census questionnaire with a vertical line. Sections that were crossed out with an "X" often resulted in numerous erroneous "7's." Also, many respondents entered a "zero" as a "0" with a slash through it. OCR always interpreted these as the value "8", not "0". It was important that these records with OCR errors not be included in the donor pool. A manual review of these questionable data was impossible given the number of records involved. Clustering techniques applied to outlier detection were used to segment the data into many clusters, one of which consisted of records with OCR errors. These records could then be screened out of the donor pools used for the imputation of missing data.

Questionnaire Design and Construction. Many of the questionnaires that NASS uses to collect survey data are produced in versions tailored to the agriculture produced in individual states and to the state and National estimates required. Different commodities may be estimated in each state and the frequency in which estimates are published for the crops may also vary across states. This has lead to the use of 50 state versions of the questionnaire for the Quarterly Agricultural Survey, used in estimating crop acreage and production and grain stock inventories.

Hierarchical clustering techniques were used to propose how individual states' questionnaires could be combined to reduce the number of versions necessary (Earp, Cox, McDaniel and Crouse, 2008). Solutions were generated for 20 clusters, and also with a further reduction to 5 clusters. Implementing "regional" versions of the questionnaires based on these clusters could lead to a reduction in the resources necessary to produce questionnaires and to coordinate data collection, while minimizing the amount of unnecessary questions administered in a state.

In the census of agriculture, all operations in a census region receive the same report form. Regions consist of states that have similar types of crops, but beyond the list of common field crops, the report forms are not tailored to specific types of operations. Cluster analysis could be used to identify clusters of farm operations based on the items they report on the census report form. These clusters could be used to develop different forms with appropriate items for subgroups of respondents without using a simplistic geographic grouping.

Identifying Subtypes of Records Missing from the Census Mail List. For the 2007 Census of Agriculture, estimates of the number of farms missing from the mail list were made based on farm operations that were found on a separate area frame sample survey. For each of the farms in that sample, data were collected to assist in determining whether the operation qualified as a farm. Cluster analysis can be used to identify whether there are distinct subgroups within this sample. This information may be used in future list building efforts to find additional sources of these types of operations to add to the 2012 Census mail list.

Association Analysis

Association analysis is yet another data mining technique with potential application in survey organizations. This type of analysis, a form of "market basket" analysis, generates association rules which describe which items within records tend to occur together. For market research this is used to determine what items appear together in customers' "market baskets." Item associations are generated based on the strength of the association, the frequency of occurrence, and the predictive utility of the relationship.

Survey Data Edit Design. Association analysis can be applied to records in an existing survey data set, treating each record as a basket and individual data in each record as the items in the basket. This will generate many known relationships between items. For example, in NASS surveys we would find that when dairy cattle are reported, milk production is also reported. But in addition, it may uncover previously unknown associations that may in turn suggest other edit rules for these data. It may also identify unusual or rare associations which may be the result of a survey edit. This analysis has not as yet been done in NASS, but could be a promising application of data mining in the future.

Concluding Remarks

Many data mining analysis techniques have been developed for use in commercial applications where large datasets are generated. These techniques help turn voluminous raw data into actionable information for decision making. To date, these techniques have found limited application in the production of official statistics. Limitations in commercially available software and computing power were practical obstacles to data mining in the past, but data mining software and platforms on which to run it are now readily available to statistical agencies. Federal statistical agencies often have large datasets at their disposal well suited to the application of data mining techniques. NASS has started several promising avenues exploring how these techniques can be applied to make improvements throughout the various survey processes involved in producing our official statistics. Further application of these techniques may help improve operational production efficiencies and the quality of official statistics.

References

Cecere, W. (2009) “2007 Census of Agriculture Non-Response Methodology” US Department of Agriculture, National Agricultural Statistics Service, RDD Report in preparation, Fairfax, VA.

Cohen, S.B., DiGaetano, R., and Goksel, H. (1999). "Estimation Procedures in the 1996 Medical Expenditure Panel Survey Household Component," Agency for Health Care Policy and Research, MEPS Methodology Report No. 5, AHCPR Publication No. 99-0027, Rockville, MD.

Earp, M., Cox, S., McDaniel, J., and Crouse, C. (2008) “Exploring Quarterly Agricultural Survey Questionnaire Version Reduction Scenarios,” US Department of Agriculture, National Agricultural Statistics Service, RDD Report 08-11, Fairfax, VA.

Garber, S. C. (2009) “Census Mail List Trimming using SAS Data Mining” US Department of Agriculture, National Agricultural Statistics Service, RDD Report 09-02, Fairfax, VA.

McCarthy, J. and Earp, M. (2009) “Who Makes Mistakes? Using Data Mining Techniques to Analyze Reporting Errors in Total Acres Operated,” US Department of Agriculture, National Agricultural Statistics Service, RDD report 09-05.

McCarthy, J. Jacob, T. and McCracken, A. (2009) Modeling NASS Survey Non-response using Classification Trees. US Department of Agriculture, National Agricultural Statistics Service, RDD Report in preparation, Fairfax, VA.