# SEARCHING FOR DONORS: FINDING AN IMPUTATION STRATEGY

Michael Hogye, National Agricultural Statistics Service, USDA
3251 Old Lee Highway, Fairfax, VA 22030

KEY WORDS: imputation, donor, editing

## 1. Introduction

When responsibility for the U.S. Census of Agriculture was transferred in 1997 to the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture, the move set in motion a series of institutional and procedural changes within NASS that continue today (Atkinson, 2003). The 1997 census was conducted as a joint effort, heavily utilizing the resources of the U.S. Bureau of the Census, which had conducted the census in previous years. NASS planning for the next agriculture census in 2002 focused on realigning its organizational structure and developing resources to take full responsibility for this task (Atkinson and House, 2001). The transition was used as an opportunity to establish an ambitious framework for change, internally named the Project to Re-engineer and Integrate Statistical Methods, or PRISM (Atkinson, 2000). As final steps are being taken to complete production and dissemination of the 2002 census, reassessment and realignment of the PRISM project has begun, in part to address the 2007 Census of Agriculture. In particular, considerable attention has been focused on the role of imputation.

NASS has a long history of conducting agricultural surveys, employing a nationwide network of analysts to select samples, edit data, and report statistics. Individual training and career development cultivate a sense of personal responsibility for data integrity through all phases of the survey process. Agency culture emphasizes a survey product that encompasses "timely, accurate and useful statistics." Each survey report merits the attention of an analyst, who may make adjustments to ensure that the data record is internally consistent. Whenever possible, imputation of data to compensate for errors or omissions is done subjectively, using expertise applied to previous data from the same respondent, or from a similar agricultural operation. Since 1997, the NASS mission has been complicated by the added demands of the census, and NASS traditions have been challenged by the expanded mission. Among those traditions, the NASS approach to data editing and imputation is undergoing a transformation.

It was recognized from the beginning that the size of the census editing task would force it to be largely automated. This required not only the development of a new processing system, but a change in editing philosophy. Beyond providing better tools to improve the productivity of analysts doing data editing, the PRISM system itself also performed much of the editing and imputation. The black-box nature of the new editing facility caused concern for a staff whose relationship to the data was changed, as personal intimacy with individual data reports was replaced with detached oversight of an automated system. Internal agency acceptance of PRISM was further hindered by problems to be expected of a new system, developed under time pressure with minimal resources. In the end, staff expertise was used at a much higher rate than planned to manually resolve editing problems. PRISM's capabilities were heavily stressed by many overtime hours of traditional NASS teamwork, which successfully completed the task. Although the automated editing programs performed as planned, imputation results played a major role in some of the PRISM editing difficulties that were painstakingly rectified by analysts.

In developing its methodology for the agriculture census, NASS had no in-house precedent to follow for automated imputation. On one hand, manual imputation in NASS surveys uses the judgment of subject-matter experts, whose case-by-case thought processes are not easily reduced to computer code. On the other hand, alternative imputation procedures have been extensively programmed and successfully used for survey processing (Atkinson, 1988), but their methodology is not directly applicable to the census. In the case of imputation for NASS sample surveys, after all survey reports are gathered, cell averages or totals---sometimes scaled by relevant proportions---may be used as needed to represent missing values within a stratum. Because aggregate values are the main objective of these surveys, findings are minimally affected by such static imputation, even though their variance estimates may be understated. NASS recognized in its planning, however, that the agriculture census would require a broader approach to imputation, since data distributions rather than just aggregates would be of interest.

## 2. Transition from 1997 to 2002

Imputation for the 1997 census used a hot-deck approach. Steadily updated streams of individual values from successfully edited records were held in reserve for imputation of incomplete or inconsistent records processed later in the same batch. As preparations began for the 2002 census, NASS had little in common with the Census Bureau in regard to computer hardware, operating systems and database engines, so that a whole new editing system was required. In anticipation of major changes,

NASS considered a wide range of imputation methodologies, rather than building directly on the 1997 hot-deck experience. PRISM working groups debated what would be most desirable in a fully integrated editing and imputation system for NASS use. Since a fully automated system had been deemed necessary, specifications focused on what it should produce; its ease of use; and its comprehensive and seamless treatment of data records (Yost *et al*, 2000). While the Fellegi-Holt principles were cited in discussions as desirable attributes for the new system, few specifics were laid out regarding methodology that might implement them.

Even before PRISM was launched, a small research project had been undertaken in NASS to develop a prototype for an integrated editing and imputation system. As it progressed, hope grew in some quarters that, after testing as a survey processing tool, this work might become a pilot for 2002 census development. Alternative methodologies were considered under this initiative, invoking the spirit of Fellegi-Holt while seeking practical solutions applicable to NASS (Todaro, 1997). It ultimately was modeled after the GEIS (Generalized Edit and Imputation System) package of Statistics Canada, and became known as AGGIES, the Agricultural Generalized Imputation and Edit System (Todaro, 1999). While the goal of the AGGIES investigation was a functioning, small-scale system to integrate the mathematics and methodology of an editing philosophy into a NASS context, PRISM development focused on the logistics of a large-scale processing system. Even as AGGIES sought proof of concept, PRISM was creating complex database structures, large bodies of edit rules and metadata, and tools to manage both the development and production processes for the census.

During the stretch run leading up to the 2002 Census of Agriculture, PRISM proposals for census editing were widely debated. Pride in a long NASS tradition of labor-intensive survey work that provides high-quality information lay in the background. The 1997 census experience, using borrowed resources employing well-refined hot-deck procedures, was in the recent past. New ideas, including AGGIES, were in the air. Time pressure increased rapidly as the foundation for processing of census data was being laid, and administrative choices had to be made. Through the efforts of NASS research staff, the Fellegi-Holt editing philosophy was publicized within the agency (Perritt, 2001). PRISM teams included NASS staff who had previously served with the Bureau of the Census and had experience with hot-deck methods. These groups tried to balance practical needs for processing speed and smooth production flow with concerns about editing methodology and skepticism about the quality of machine-edited data.

AGGIES was written in SAS, using SAS/AF to develop a graphical interface for entering groups of linear edits, specifying data sets, and selecting options. SAS/IML was used for the bulk of computations. As in GEIS, edit groups were evaluated for consistency, redundancy and hidden equalities, allowing the user an opportunity to make necessary changes before continuing. Survey records were then run through the edits, determining which ones would need correction. For inconsistent or incomplete records, error localization employed the Statistics Canada formulation of the modified Chernikova algorithm, marking fields for imputation by changing their values to -1. For the next step, AGGIES allowed the user to determine a hierarchy of formulas that would create imputed values based on means, previously reported values, or such values scaled in proportion to an auxiliary variable. The various options for substitution would be considered in the specified order until a combination of values succeeded in passing all the edits. As an option of last resort, in case all versions of an imputed record failed, AGGIES derived imputed values using an idea from the Bureau of the Census' SPEER system. Extreme points of the feasible region were calculated and then used to find the midpoint of each dimension within the region, assuring a solution that passed all edits.

AGGIES was tested and reviewed within NASS (Perritt and Todaro, 2000), but agency attention increasingly focused on the anticipated demands of imminent census work. With the departure of the creator of AGGIES, the agency also lost its primary proponent of Fellegi-Holt editing principles. Work on a planned AGGIES donor imputation capability was discontinued, even as the PRISM team sought a comprehensive editing and imputation strategy for census processing. However, momentum from AGGIES research carried forward far enough for a consensus to be reached, that a donor imputation module would be included in the PRISM system. Further, it was agreed that error localization would be used to identify the fields addressed by donor imputation. Although AGGIES was laid aside, error localization and donor imputation modules were thus added to census plans, and as a result edits were to be expressed in linear form for imputation purposes.

Although AGGIES and its editing methods showed promise, the system could not be integrated into the emerging PRISM system. As a pilot project, AGGIES was not designed to be scalable to the point of handling the agriculture census, and it had been designed as a standalone system rather than a module within the PRISM framework that emerged later. Even so, the influence of AGGIES and the Canadian GEIS model proved to be substantial in PRISM's approach to imputation. The concept of imputing from a nearest-neighbor record, chosen from a pool of clean records, was adopted for PRISM use. However, events forced this idea to take on

a life of its own, divorced from the broader context of Fellegi-Holt principles and specifically from a mathematical error localization process.

### 3. Preparations for 2002

PRISM plans for data editing focused on the size and complexity of the task. In one dimension, the census project was expected to process nearly 2 million census records. In its other dimension, the long-form version of each census record contained over a thousand fields, including those derived from respondent data. Hundreds of additional flags and codes were used to track the status of each record, and to accumulate a complete history of its editing changes. As a first step, NASS editing specialists mapped out the data fields and their attributes as tables of metadata; this work alone consumed a staff year of effort (Manning and Beranek, 2004). Decision logic tables (DLTs) were then built to apply limits to data values and evaluate relationships among fields. To make this task manageable, the fields and their DLTs were broken into nearly four dozen modules.

The metadata defining data fields and the tables expressing edit logic were used by NASS IT staff to map out database structures and to outline schematics for the flow of processing. The census information itself, including a record of changes accumulated through the editing process, was stored in a RedBrick database. A large and complex body of administrative data tables, ranging from census lists to edit parameters, was housed as a Sybase database. Editing was designed to accept as input the identifiers of up to 75 census records, whose current data values were extracted from RedBrick. After running a series of preliminary checks while gathering relevant administrative data from Sybase, the program cycled through each of the 46 edit modules. The batch-mode SAS program that processed each group of records was "wrapped" around the many editing steps and diverse housekeeping chores, tying them together and properly synchronizing them for a successful edit. SAS/ACCESS provided the wrapper program with entry into RedBrick and Sybase. After a run of the Wrapper, data modifications and changes to codes and flags would first appear as direct updates to appropriate Sybase tables; a separate "loader" process would then post data updates to Redbrick.

To be included in census processing, editing steps---including error localization and donor imputation---had to be integrated into the modular batch environment of the Wrapper. A three-pronged editing strategy allowed concurrent development of required components by appropriate staff. These included a master edit program for each module, a warehouse of historical data, and a donor imputation program. For each edit module, control was passed by the Wrapper to an edit program written in SAS/SCL and designed specifically for the module. An in-house facility was created to aid in the development of these module-by-module edit programs. Original edit logic and a steady stream of modifications were entered interactively with this "authoring tool," which saved changes in DLTs and recompiled the module's edit program. During census processing, these modular edit programs applied DLT logic to each census record and determined whether any corrections were required to fields in a given module. Deterministic changes were made first; these were unique and readily apparent corrections identified as the direct outcome of edit logic. If further adjustments were required, another subsystem was consulted for any previously reported data (PRD) on the record under consideration. As deemed appropriate by the edit program of that module, such historical data would be imputed, often after adjustment in proportion to an auxiliary variable. If this second form of adjustment failed to make the record internally consistent within the module, then donor imputation was considered as a third remedy.

Donors were to be sought and data were to be borrowed from them in the context of each editing module and its relevant fields. As input to the donor imputation program, values of -1 were to indicate missing values. Additional fields requiring changes were to be flagged with -1s after error localization, preliminary to donor imputation. However, the performance of AGGIES code for error localization, even when adapted to PRISM use and further rewritten in Base SAS rather than SAS/IML, proved disappointing. After all efforts to substantially speed up error localization fell far short of targets, it was dropped from development plans, leaving a void in the PRISM editing strategy. Measures were taken to implement donor imputation alone. In place of error localization, DLT logic was expanded to assess the potential causes of edit failures and thereby to specify which fields were most reasonable to impute from a donor. Matching variables---those fields used to define similarity between donor and recipient---were also explicitly specified for each field selected for imputation. The linear form of the edits became moot, as they no longer drove mathematical processes identifying which fields to change and which fields to use for donor searches. The development burden on the data editing staff increased further, while the addition of static tables increased the scope of Sybase tasks.

The nearest-neighbor donor imputation program (NNI) was also written in SAS. It was called by the Wrapper as needed within each editing module, to process each batch of records. Corresponding to each census field containing -1s, the edit program provided an imputation rule giving specifications chosen dynamically by DLTs. The rule dictated whether the imputed value had to be positive; whether it needed to be scaled; and which

auxiliary variable to use if scaling was required. When used, scaling consisted of multiplying the donor value in the imputed field by the auxiliary variable's recipient-to-donor ratio. Before beginning donor searches, the NNI program consolidated the imputation specifications relating to each census record. The union of matching variables was found for all fields to be imputed in a record. At a minimum the latitude, longitude and size of a farm were always available, so that even in cases where matching variables required imputation, there was still information for calculating distances between a recipient record and its potential donor records. The union of positive-value restrictions on donors was also determined. This included both imputed fields explicitly required to be positive and also fields specified as scaling variables. After collecting matching variables and positive restrictions, NNI was ready to search the available donor records, in order to identify those nearest to the census record being edited.

Methodology for donor searching was straightforward. Simple Euclidean distance calculations were applied to the consolidated set of matching variables. Matching fields were scaled to unit variance within their pool of donors, so that individual fields would not be weighted by the relative sizes of their measurement units. Thus the distance between a recipient record and a donor pool record was simply the sum of squared differences between matching variables, with each term adjusted by its variance within the pool of donors. Only donor records satisfying the accumulated positive restrictions were considered for imputation; up to 25 nearest neighbors were identified by the search. A special set of edits, provided by the NASS editing staff for this purpose, was applied to the completed record formed by imputing from each of the nearest neighbors. After considering donors in order of closeness, the first such composite record to pass all the edits became the edited record. During searching and imputation, the process might fail for either of two reasons. All donors in the pool might fail the positive constraints for the search; or imputed records using all the neighbors found in the search might fail one or more of the subsequent edits.

## 4. Processing for the 2002 Agriculture Census

Implementation of donor searching, both in logistics and performance, posed challenges. Usefulness of search trees, such as the k-dimensional tree algorithm used in GEIS, was likely to be limited by the batch nature of census editing. As the composition of the donor pool changed frequently during processing, the rearrangement of donors would require substantial computing resources to rebuild tree structures used to represent them. After techniques were considered for optimizing search speed, a strategy was chosen to limit the number of donors made available for each search. Instead of assembling all clean census records into a single large donor pool, PRISM created fifty smaller ones, each corresponding roughly to a state and its surrounding areas. Donors for a record were sought only in the geographically appropriate donor pool. The individual donor pools were SAS data sets of successfully edited records, extracted from the RedBrick repository and represented in a "skinny record" format emulating a database; each skinny record contained a single data item and its identifying information. Since all records in a batch were from the same state, each Wrapper process was able to make use of a single donor pool.

Final preparations for the PRISM editing system addressed concerns about system throughput. In order to make efficient use of an IBM Regatta P690 system, deployed with 32 processors and 128 gigabytes of memory running under LINUX, planning increasingly focused on ways to run multiple processes without creating bottlenecks. Grouping of records by state made it possible to run multiple batches simultaneously, with each job having dedicated use of the appropriate donor pool data set. However, batch edit jobs came from two distinct sources. First, each census record was edited as a member of a batch of up to 75 records when it entered the PRISM system. Following this initial batch edit, the record became available for interactive corrections, made online by a NASS analyst. The corrected record would then be processed individually by the batch Wrapper program as a "batch-of-one." Further analyst modifications to the same record would result in another trip through the edit program as another batch-of-one. It was a challenge to schedule jobs so that the system could run these batch-of-one jobs, in order to respond quickly to online requests. Two processing paths were made available: one for the initial, larger batches; another for the batches-of-one, with the latter having priority. To facilitate this, two copies of the entire set of donor pools were maintained, resulting in a hundred large datasets.

System scheduling of edit jobs became a key issue in expeditious processing of census forms. Each SAS process driving the Wrapper program would ideally be routed to its own processor; have its own scratch disk; and have exclusive possession of a donor pool. Balancing these considerations, with little or no time to break in the system, was a difficult learning process. An additional complication was day-to-day management of donor pools. When corrected census records became available for use as potential imputation donors, they were added to donor pools. Even with concessions in methodology to improve system performance, donor imputation proved to be a conspicuous drain on system resources, especially as individual donor pools grew in size and donor search times increased correspondingly.

In order to further limit the length of donor searches, a cap was placed on the size of each donor pool. To carry

out these requirements, it was necessary to regularly recreate the donor pools, giving consideration to new records while restricting the size of the updated data set through systematic sampling of eligible records. This donor-pool maintenance task further complicated the flow of edit jobs, since it denied use of a donor pool while it was in progress. Management of the editing system thus required the scheduling of processes in accordance with a set of continually reassessed priorities. These priorities included the backlog of multiple-record batch jobs; batch-of-one requests generated by interactive editing; refreshing of donor pools; posting corrections to RedBrick with a loader program; balancing overall editing progress among states; and required system maintenance.

**5. 2002 Imputation in Hindsight**

Although NASS made the transition to automated editing, the overall editing task was accomplished only through intense and extended efforts by agency staff to compensate for system shortcomings. Fundamental dilemmas are being reconsidered to mitigate this in the future. One critical difficulty was in the creation of initial donor pools. For imputation, suitable information from other sources was very limited, while system design made it difficult to use early returns from the census for this purpose. Although editing and imputation were done in modules whose scope was limited to a portion of the overall record, it was not possible to edit modules independently and then make their data available for imputation. Instead, it was necessary to cumulatively edit and correct an entire record before it could be released for use as a donor. Without imputation to provide corrected values, editing would fail and require manual input. After such intervention, records were likely to fail again in subsequent modules, until they had been manually nursed through the entire edit program. This labor-intensive process to create clean starting records was done under intense time pressure, since automated editing could not begin until satisfactory donor pools became available.

To some extent, the goal of applying the three Fellegi-Holt principles fell victim to the pressures of time constraints and limited resources. The third principle in particular, urging that imputation rules be derived from edits in a manner that assures compliance with edits, was an attractive ideal that proved very hard to achieve. Once the NNI routine was invoked for a module, it was impractical to evaluate the validity of the imputed record by reusing the code that originally applied the edits to the record. Instead, a set of linear constraints was used by the imputation program to filter out unacceptable donor candidates. This segregation of the editing and imputing tasks sometimes resulted in unreasonable imputed values, which required considerable effort on the part of analysts to identify and correct. A most infamous example was the operation whose imputed expenses initially exceeded $171 billion, evoking wry comments about Pentagon budget items masquerading as farms.

The Euclidean distance measure used to express similarity between farms was applied to all donor searches in all modules, with little opportunity beforehand to investigate its applicability in diverse situations. The set of matching variables used in any given search could range from the minimal group (latitude, longitude and total acreage) to several dozen variables, according to the circumstances. When geography and farm size were the only selection criteria, it was possible that information would be imputed from a farm with radically different kinds of production. At the other extreme, there was substantial risk that the list of matching variables could be over-specified, with related variables competing to make use of the same information. Usefulness of the similarity measure was further weakened by related issues. Categorical variables were not explicitly used as part of the distance formulation, missing an opportunity to efficiently stratify donor pools. Although the continuous fields used in the distance computation were normalized to unit variance, this standardization itself depended upon the contents of the donor pool. Each time a donor pool was reconstituted, normalized values for retained records were likely to change. In effect, the relative weights of individual variables within the distance calculation could be changed by refreshing the contents of a donor pool.

In spite of extensive efforts to make it work smoothly, imputation achieved unwanted visibility for its significant contribution to lengthy edit processing times and related difficulties. Concessions had been made by eliminating error localization and by simplifying matching variable determination. To aid in processing, donor information was managed as a large collection of independent data sets. Job scheduling was done to optimize throughput, in part by keeping concurrent jobs from making use of the same donor pool. Even with these and other arrangements to speed up processing, analysts were frustrated by inefficiencies the system introduced into their work. The batch nature of edit processing forced updated records to be queued for re-editing, and then queued again for posting to the database. This cycle would generally take far too long for an online analyst to make changes and then follow up on the record during one session. By the time the record was updated and released again for online data review, it was often a distant memory. A different analyst might be assigned to consider the record, or the original analyst would have to be reacquainted with it. A final complication was the possibility that the editing program would override the analyst's efforts to correct the record, imputing an entirely new set of potentially unreasonable values. A high price was paid for slow turnaround and erratic imputation, both in terms of the analysts' time and their trust.

## 6. Looking Ahead: Reaching Back

With completion of the 2002 effort, NASS focus has shifted toward improvements for 2007. A high priority has been assigned to improving editing speed, for truly interactive responses. This consideration harks back to the Census Bureau system that was used to edit the 1997 Census of Agriculture. Some characteristics of that system, that were set aside in favor of new ideas for 2002, are being reconsidered in light of their contributions to speed achieved during 1997 processing. With the 1997 memory-based hot deck, donor pool data sets were unnecessary. Imputation calculations were fully integrated into the edit modules as steps within their DLTs. 1997 imputation began with hard-coded "cold deck" values, while imputation transactions were simply changes to individual memory locations. These could be loaded, checked and updated without any disk processing.

Different software development platforms also contributed significantly to differences in processing speed. Object modules compiled from FORTRAN code were used for 1997 data, while editing was done with SAS for 2002, much of it using SAS Component Language (SCL). On one hand, SAS provided a wide range of capabilities that allowed development to be accomplished by NASS staff. Fundamental processes could be written in Base SAS, making efficient use of PROC SQL and seamless database management provided by SAS/ACCESS. SAS/AF was used to develop the Authoring Tool, a user-friendly system that was highly successful in providing tools for analysts to write and test edit rules. The Data Review program was also developed by integrating many SAS resources into a comprehensive, interactive system for data editing, correction and analysis. Unfortunately, the magnitude of accomplishments and the repertoire of capabilities were overshadowed by inherent speed limitations; interpreted SAS code was no match for executable modules already in machine language.

While there is debate over migration of the 2002 SAS-based systems to faster software, there is unanimity about the need to improve imputation. The impetus toward a sophisticated edit and imputation system for 2002 has given way to a new motto for 2007: "Speed, Stability and Simplicity" (NASS Senior Executive Team, 2004). As they did in preparations for 2002, plans have focused on laying out principles and goals, with a continuing theme of making efficient use of staff resources. Previous guidelines were directed at providing analysts with a comprehensive set of integrated tools; current guidance emphasizes speed, with extensions to ease of use. Absent from both planning cycles has been the outline of an underlying editing and imputation strategy. As 2002 development proceeded, imputation methodology fell into place as a series of pragmatic compromises that lasted until well after production runs had begun. With preliminaries for the next development cycle already under way, the current proposal calls for grafting 1997 hot deck methods into the 2002 edits.

Even though an explicit methodology has not been pursued, there may be some intuitive appeal in the ideas of Statistics Canada's New/Nearest-Neighbour Imputation Methodology (NIM), especially considering the current NASS mood. A strategy that bases its actions on the information available for imputation seems attractive, since it bypasses the error localization step. It may also be useful to think, in NIM terms, of the imputation step as a compromise between an established, acceptable record and a failed record needing correction. Although such an imputation strategy does concentrate on minimally changing the recipient to satisfy edits, its work is predicated on available donor data. While there is reluctance to pursue any solution for census use which is at all computing intensive, some of the logic implied by working group discussions seems consistent with NIM philosophy. The NASS desire to satisfy edits with reasonable imputed data, at a low computing cost, using a transparently simple process, has been dubbed "NIM Lite."

## 7. Suggestions

Some recommendations follow, in harmony with agency priorities on fast, practical and reasonable imputation. First, imputation data should be organized and managed to suit a wide range of options. As a general rule, such information should be stratified into categories that are meaningful in expressing similarity between farms. Since stratification is a critical aspect of NASS surveys, it should be a natural step to partition imputation data into relatively homogeneous cells. For disk storage, stratification variables would be indexed for rapid data retrieval; for volatile memory, stratification variables would define the cells of an array. In either case, categorical fields helping to establish similarity between operations could be efficiently evaluated, quickly narrowing the search for imputation data to similar operations. An important option could further structure the imputation repository, allowing for diverse imputation strategies. Imputation data could be stored and retrieved at the record level or the individual field level. To dispense data in record units, an identification field would either be added as a database index (for data on disk), or as an array dimension (for data in memory). Finally, it should be possible to choose the depth of the data repository, in terms of how many values of the same field or how many entire records are allowed to accumulate from clean data for imputation purposes during the editing process.

Arranging imputation data into categories should yield a number of benefits. The search for similar farm operations can be streamlined by the addition of an initial step that limits the search to an appropriate subset of data. The similarity of the selected cell's imputation data to the recipient record is likely to be better than it was when the use of categorical information for this purpose was very limited. In particular, the similarity measure might take into account whether or not

information has been used for imputation, and whether or not the imputation information itself has been wholly or partially imputed. In addition, the stratification grid might serve as an outline for assembling cold deck or starting donor pool values. Criteria could be established to specify how much of the framework must be populated with data suitable for imputation, before the imputation function of an editing process would be invoked.

The imposition of a categorical structure, with options to further lay out imputation data characteristics, could result in a generalized system offering the full range of imputation opportunities. Editing would begin by accumulating initial imputation data on disk until thresholds for starting imputation were reached. To begin a hot deck matrix, individual values would be loaded into memory from the cold deck of values on disk, as soon as a value became available for each cell. To instead begin use of a donor pool, records would accumulate on disk until a satisfactory number of them appeared in each category. The depth of the hot deck might be limited to an initial cold deck value, replaced by a single acceptable value; or it might be a group of such values. The depth of a donor pool (of complete records) could also be controlled, to include all acceptable records or only a limited number within each category. Complete records from a starting donor pool of shallow depth might be loaded into memory and updated there, to create a RAM-based donor pool. On the other hand, if all acceptable records were retained on disk for donor pool use, then all records in a stratum could be used to calculate means, in the tradition of NASS survey imputation.

For a specific situation, the choice of options would balance priorities in processing speed and imputation methodology. In situations where the depth of imputation information is allowed to exceed a single value or record, a number of imputation strategies would become feasible. In general, one of the multiple values (or records) might be chosen for imputation at random. As an alternative, individual values might be considered by the edit program one at a time, until an acceptable value was reached. Whole records might instead be evaluated to find nearest neighbors among the available records, using continuous measures as matching variables. As another option, all the available information might be aggregated in some fashion, such as taking the mean. These alternatives would have to be weighed for any advantages in satisfying edits, or for data distributions more faithful to the original.

A technique making use of multiple-record information has been suggested by Tim Keller of NASS. A number of potential donor records are taken from the appropriate stratum of imputation data. As clean records, these donor candidates each satisfy all relevant edits. As members of the recipient's stratum, they have some similarity to the recipient. In their roles as neighbors to the recipient, they must in a broad sense reside in the same region of the overall data distribution as the recipient. It is not known precisely where in the distribution the recipient must lie, because imputation is required for some of its fields. The objective is to make optimal use of what is known about the recipient, to put forth a robust guess about the portions of the recipient that are not known. As before, fields whose recipient values are known, and which in general are known to be correlated with the missing fields, may be used as matching variables to relate each neighbor to the recipient. Although the matching variables are used as before to calculate a similarity to the recipient, the similarity values are not used to select an individual donor record from among the neighbors. Instead, the similarities are used as weights in the derivation of a single composite record from the group of neighbors, employing a formulation known historically as the Fermat-Weber problem.

The Fermat-Weber solution gives the point from which the weighted total of distances to all the neighbor points is minimized. The distances are in terms of the fields to be imputed, which are known for the donors but not for the recipient. The weights are in terms of matching variables, which are known both for the recipient and the donors. Each donor's similarity, in terms of what is known about the recipient, exerts a corresponding influence while establishing a central point in the space among the donors, to estimate what is not known about the recipient. This result should identify a point in the data distribution where the missing values might reasonably be found. A portion of the distribution, in terms of the imputed fields, is roughly sketched out by the nearest neighbors' values in those fields. A different set of imputed values within that portion of the distribution may be derived from the same set of neighbors, according to the matching variable values of a given recipient.

While categories to structure imputation data and options to make flexible use of the information may increase processing speed and improve quality of the results, the fundamental problem of satisfying the edits will remain. For NIM Lite to work, authors of DLTs must have control over raw data obtained from donors, to "shoehorn" values to fit the edit constraints. An attractive feature of the Fermat-Weber solution is that it will also satisfy linear edits to which the selected neighbors conform. However, the basic dilemma remains. Various strategies for overlaying values from one farm or a group of farms, onto the record of another farm, may appear reasonable in terms of estimating the recipient record's location within a complex distribution. However, there are likely to be matters of scale and many other adjustments that cannot be anticipated. There can be no guaranteed cure for such ills without the formulation and construction of a generalized editing and imputation system, as prescribed by Fellegi and Holt. The alternative is to arm the edit logic with an arsenal of tools to adjust values appropriately and to efficiently evaluate their suitability.

The greater the depth of imputation data, the more choices there will be among donor values as candidates to satisfy the edits. However, a trial-and-error strategy adds

unpredictably to the computing cost, without assurance of success. Techniques such as those used for 2002, including positive-only constraints and scaling against an auxiliary variable, are valuable. The former requires multiple donor choices, while the latter requires whole-record imputation. Editing may employ a univariate imputation strategy, in the sense of imputing and shoehorning values one field at a time, until the final imputation is deterministically implied. This method may benefit from allowing a different donor for each imputed value. In order to preserve scaling information when needed in a univariate imputation framework, the system should store the donor's ratio between an imputation field and an auxiliary field, rather than just the raw value. As reflected in the widespread use of ratios in 1997 to define univariate hot decks, and also in the prominence of ratios in proposals for 2007, it becomes clear that the key information to be imputed is often the relative size of two fields, rather than the actual value of either one. Finally, a field with strong empirically-estimated correlation to matching variables may suggest a more direct imputation strategy, eliminating the need for direct substitution of donor data. In a generalization of the ratio-variable scaling idea, relevant information from similar records might be used to establish a regression equation, or other "Blue Book" model, to calculate imputed values from auxiliary fields whose values are known for the recipient. The key will be to provide developers with simple but plentiful strategies for shoehorning values into place.

**References**

Atkinson, D. (1988), "The Scope and Effect of Imputation in Quarterly Agricultural Surveys," *NASS Staff Report No. SSB804*.

Atkinson, D. (2000), "Developing a State-of-the-Art Editing and Imputation System for NASS' Agricultural Census and Sample Surveys," *UN/ECE Work Session on Statistical Data Editing, Cardiff*, October 18-20, 2000.

Atkinson, D. (2003), "The Development and Implementation of a New Processing System for the 2002 Census of Agriculture," *UN/ECE Work Session on Statistical Data Editing*, Madrid, October 20-22, 2003.

Atkinson, D. and House, C. (2001), "A Generalized Edit and Analysis System for Agricultural Data," *Conference on Agricultural and Environmental Statistical Applications*, Rome, June 5-7, 2001.

Manning, A. and Beranek, J. (2004), "Data Edit and Review," *Internal NASS Presentation for Russian Federation*.

NASS Senior Executive Team (2004), "PRISM II System Redesign," *Internal Presentation*.

Perritt, K. (2001), "Fellegi-Holt Editing," *Internal Presentation to NASS Staff*.

Perritt, K. and Todaro, T. (2000), "Overview and Evaluation of AGGIES, an Automated Edit and Imputation System," *NASS Research Report No. RDD-00-03*.

Todaro, T., (1997), "Evaluation of the SPEER Automatic Edit and Imputation System," *NASS Research Report No. RD-97-04*.

Todaro, T. (1999), "Evaluation of the AGGIES Automated Edit and Imputation System," *NASS Research Report No. RD-99-01*.

Yost, M., Atkinson, D., Miller, J., Parsons, J., Pense, R., and Swaim, N. (2000), "Developing A State of the Art Editing, Imputation and Analysis System for the 2002 Agricultural Census and Beyond," *Processing Methodology Sub-Team Report, National Agricultural Statistics Service.*