SHARED SYSTEM FOR ASSESSING CONSUMER OCCUPANCY AND DEMOGRAPHIC ACCURACY: ICIQ-2003 PROCEEDINGS

(Practice-Oriented)

John R. Talburt, Ph.D.
Acxiom Corporation, Little Rock, Arkansas
John.Talburt@acxiom.com

Greg Holland
Acxiom Corporation, Little Rock, Arkansas
Greg.Holland@acxiom.com

Abstract: A discussion of the design, development, and administration issues related to the implementation of an automated, multi-user tool for assessing the accuracy of consumer occupancy and demographic information in information products based upon a comparison to shared database of collected survey responses. Covers user education and acceptance issues as well as system design and quality assurance processes.

Key Words: Data Quality, Information Quality, TDQM, Information Product, Data Accuracy, Data Quality Tools, Data Quality Automation, Survey Reponses

INTRODUCTION

For commercial, consumer-based information (data) products, the accuracy of occupancy and demographic information is usually one of the most important dimensions of quality, but it is often the most difficult for organizations to measure. For the context of this paper, "occupancy" is defined as "an individual or household at an address for a given period of time." For practical purposes, "occupancy" also entails telephone number in that the primary source of reference data often comes from telephone survey responses. Aside from name, address, and telephone number, there are also a number of other important consumer and household demographics such as date of birth, length of residence, dwelling type, etc.

In the Solutions and Product Organization within Acxiom Corporation®, the TDQM initiative is called the "Data Quality Scorecard." The Data Quality Scorecard requires that all products, including consumer-based information products, will have at least one measure of accuracy.

Problem

"How can data product managers obtain regular, reliable, and consistent measures of the accuracy of consumer occupancy and demographic elements without large investments of time, money, and other resources?" The most direct way to verify the accuracy of consumer information is through personal interviews with consumers, either in-home or by phone. However, designing and conducting a consumer survey campaign can be both costly and time-consuming to do just once, much less repeatedly, as is necessary in the TDQM cycle [1].

For this reason, many data product managers often rely on less direct measurement methods such as competitor comparisons and customer feedback. However, competitor comparisons beg the question of "Who's right?" and customer feedback can be difficult to quantify and act upon. Even worse, many have substituted consistency for accuracy simply because it is easier to measure. Although competitor comparisons, customer feedback, and product consistency are important quality measures, they are poor substitutes for regular measurements of the accuracy of key product elements.

Solution

The solution for our organization was to design and build an automated system (called the Geneticx AccuCheck Tool) based upon a common repository of consumer interview responses (called the Reference Database) that allows the product teams to take accuracy measurements as part of their normal processing flow. In effect, the basic value of the system is to promote reuse--in this case, to "re-use" survey response data, i.e., "collect once, use many times." Instead of each product team conducting its own survey campaign only for its own product, a single team continuously collects survey results and makes them available in a shared repository. As a further convenience, product teams can access the survey data through a software application (called the Analysis Operator) that matches the input data to survey data, then calculates and reports the estimated accuracy of each item analyzed.

Paper Organization

The remainder of the paper is organized as follows

I. Discussion of requirement drivers for AccuCheck in terms of three types of applications
   A. Data Quality Scorecard Initiative
   B. Competitor Product Analysis
   C. Pre- and Post-Acquisition Assessment of Source Data
II. System Design
   A. Reference Database
   B. Analysis Operator
III. Conclusion
   A. Advantages
   B. Limitations
   C. Future Work

APPLICATIONS DRIVING SYSTEM REQUIREMENTS

Data Quality Scorecard Initiative

The Data Quality Scorecard Initiative is the organization's approach to TDQM for its data (information) products. Initiated and promoted by organizational leadership, it is an attempt to bring some level of statistical process control to the building of data products within the organization.

As part of the project, the teams that build each product and each internally compiled source have to demonstrate that they have a set of data quality requirements in place for their product and a plan for data quality improvement. In addition, each team has to design and implement quarterly measurements in the following five quality dimensions in a way that is relevant to its product requirements:

1. Accuracy
2. Coverage
3. Consistency
4. Access

5. Grouping Accuracy

Note that measurements of accuracy are required for all products, and for most consumer products, these are done using the AccuCheck Tool.

It is also worthwhile to note that the last dimension, grouping accuracy, is not typically listed in the literature, although implied by other quality dimensions. For this project, it was defined to be a measure of how well products bring together occupancy records (i.e., name and address information for the same individual or household). It was added to the Scorecard because of the organization's focus on Customer Data Integration (CDI), one the basic processes needed to execute a successful Customer Relationship Management (CRM) strategy. Therefore, it was deemed important to measure how well each product team was able to integrate its own data sources

Because the Scorecard for each product team was based upon the same template, all of the individual product scorecards can be "rolled-up" into an organizational scorecard that answers the question, "How's your data quality?" To help assure that appropriate attention was placed on the initiative, organizational leadership also elected to put a portion of each product manager's compensation at risk based on the annual improvement of his or her product's Scorecard.

Competitor Analysis

Just as the AccuCheck Tool helps product teams advance toward internally set quality goals, it also provides a means to objectively rank the organization's products against its competition in accuracy as well as in coverage and consistency. Data product competitor analysis has become an important facet of the organization's data quality strategy, by-and-large facilitated by the AccuCheck system.

Evaluation of Information Sources

The AccuCheck System is an integral part of the screening process for new data sources. It is also used as part of the "accuracy trend analysis" performed on regularly updated data sources. The ability to easily and objectively estimate the accuracy of consumer occupancy and demographic data has enabled the organization to make tremendous strides in improving the overall quality of its data products.

SYSTEM DESIGN

Geneticx AccuCheck comprises two primary subsystems:

1. Survey Response Database (Reference Database)
2. Analysis Operator

Figure 1 shows these sub-systems in the overall system context.

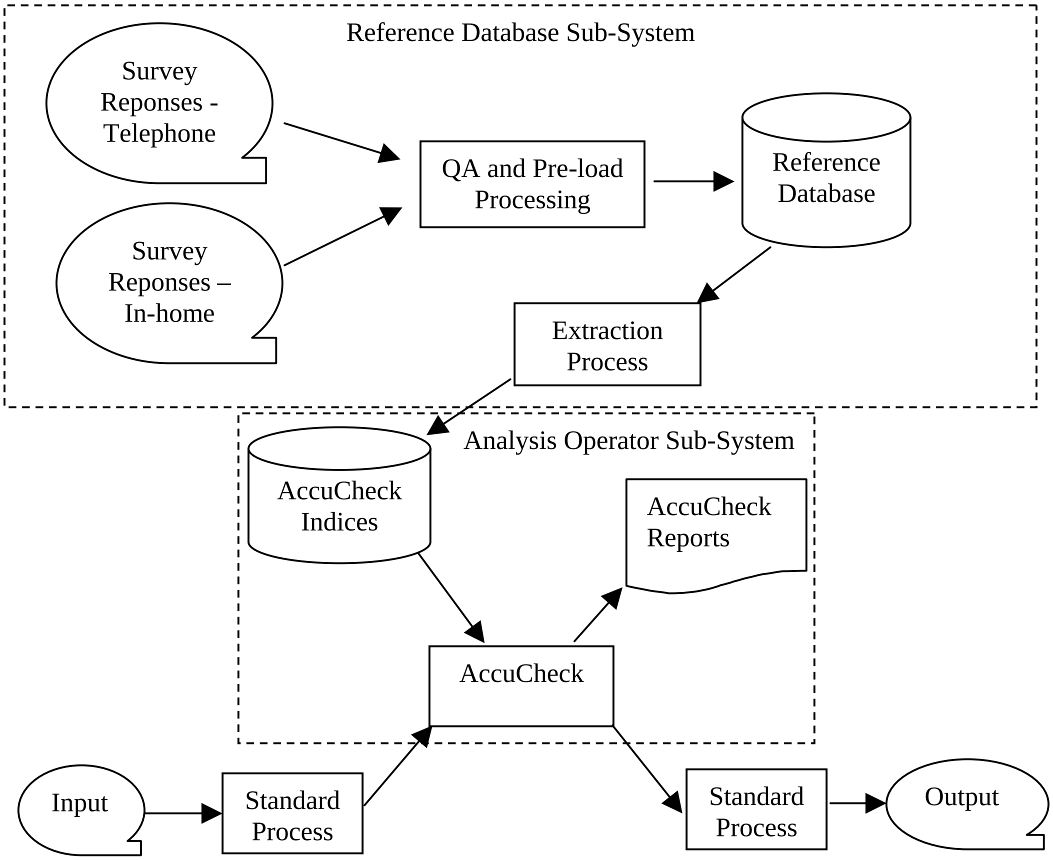

Reference Database Sub-System
Survey Reponses - Telephone
Survey Reponses – In-home
QA and Pre-load Processing
Reference Database
Extraction Process
Analysis Operator Sub-System
AccuCheck Indices
AccuCheck Reports
AccuCheck
Input
Standard Process
Standard Process
Output

Fig. 1. *Sub-systems in Overall System Context*

Survey Response Database (Reference Database)

The Reference Database is currently a MS Access Database maintained by the Geneticx Data Quality Toolkit Team. The Reference Database is multi-sourced, but the majority of the records come from weekly updates of telephone survey responses based on a calling list generated by the team. The telephone survey responses are supplemented by in-home interview responses purchased from marketing research firms; however, these are typically updated only about two to three times per year.

To increase system performance, the Analysis Operator does not directly access the Reference Database. Instead, a monthly extract is performed, which creates a set of pre-ordered indices that are used by the analysis operator for matching and for accuracy calculations. While the database retains all survey responses gathered since the inception of the system, only those gathered within the past three years and believed to still be current (correct) are extracted to build the operator indices (see discussion of "Recent versus Current Information" in "Reference Database Considerations" Section).

Analysis Operator

The Analysis Operator is a program that matches consumer records from the user input file to survey responses extracted from the Reference Database. When appropriate matches are found, the associated information from the survey is compared to the user information in order to accumulate an accuracy score for the item being tested.

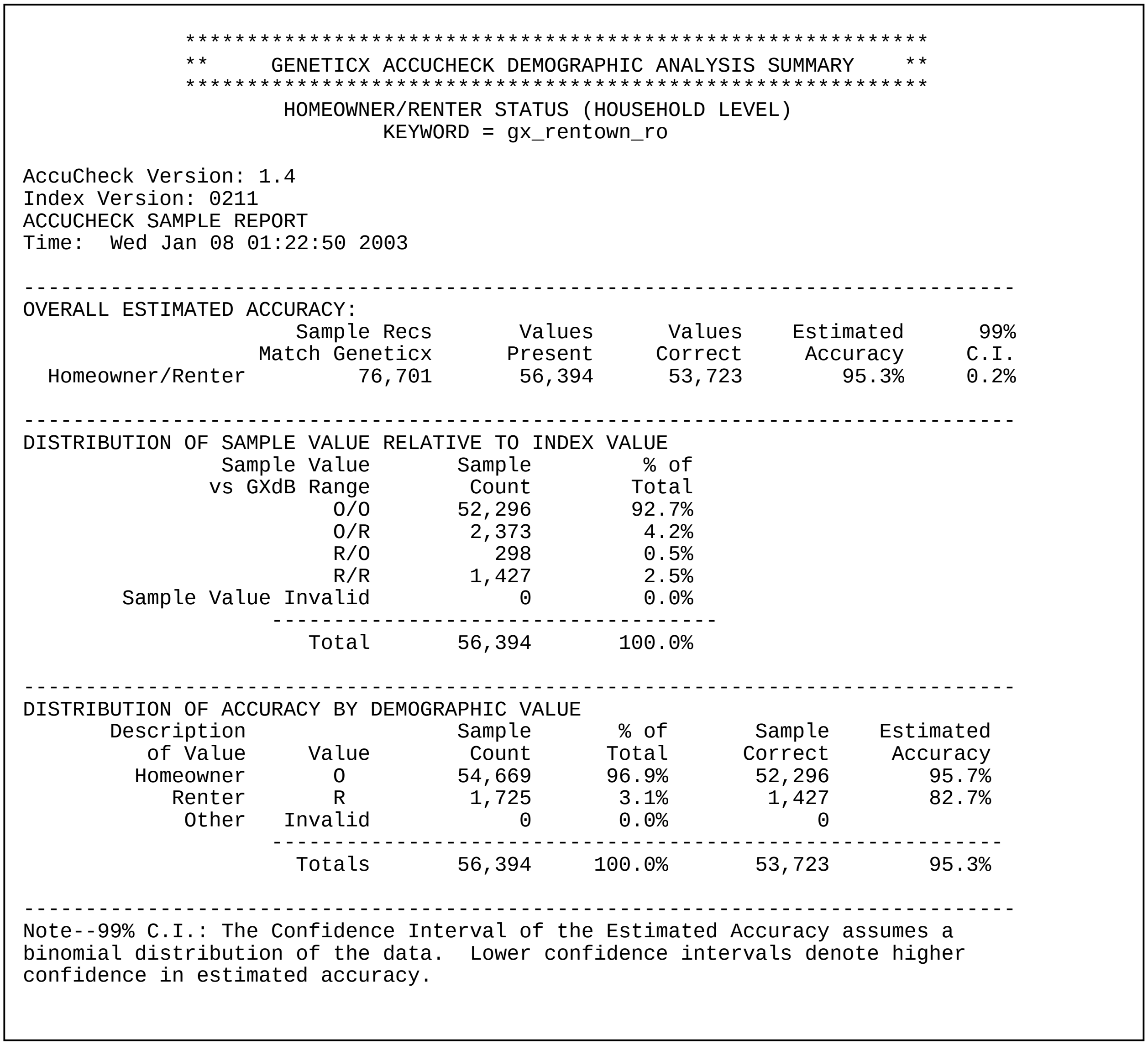
For example, if the user requests the accuracy of telephone numbers at an address, the system will attempt to match each user address to an address in the index. If a match is found, then the system will compare the user-provided telephone number to the survey response telephone number and score it as correct or incorrect. After all user records have been processed, the operator calculates an overall score of correct numbers to all numbers in the user sample and prints the results in a report for the user.

The Analysis Operator currently can produce up to 18 reports depending on which analyses the user requested. There are three occupancy accuracy reports, and fifteen demographic accuracy reports. These reports are shown in Table 1.

Report	**Type**	**Match Level**
Surname & Telephone Accuracy	Occupancy	Address
Surname & Address Accuracy	Occupancy	Telephone
Address & Telephone Accuracy	Occupancy	Individual
Date of Birth	Demographic	Individual
Marital Status	Demographic	Individual
Gender	Demographic	Individual
Individual Age in Two-year Increments	Demographic	Individual
Education Level	Demographic	Individual
Homeowner or Renter	Demographic	Household
Purchase Date of Home	Demographic	Household
Length of Residence	Demographic	Household
Number of Adults	Demographic	Household
Number of Children	Demographic	Household
Household Income	Demographic	Household
Presence of Children	Demographic	Household
Household Size	Demographic	Household
Home Market Value	Demographic	Address
Dwelling Type	Demographic	Address

Table 1: Analysis Operator Reports

Figure 2 shows an example of a Homeowner or Renter Report.



```
             ************************************************************
             **     GENETICX ACCUCHECK DEMOGRAPHIC ANALYSIS SUMMARY    **
             ************************************************************
                     HOMEOWNER/RENTER STATUS (HOUSEHOLD LEVEL)
                             KEYWORD = gx_rentown_ro

AccuCheck Version: 1.4
Index Version: 0211
ACCUCHECK SAMPLE REPORT
Time:  Wed Jan 08 01:22:50 2003

--------------------------------------------------------------------------------
OVERALL ESTIMATED ACCURACY:
                      Sample Recs       Values      Values    Estimated      99%
                   Match Geneticx      Present     Correct     Accuracy     C.I.
  Homeowner/Renter         76,701       56,394      53,723        95.3%     0.2%

--------------------------------------------------------------------------------
DISTRIBUTION OF SAMPLE VALUE RELATIVE TO INDEX VALUE
                Sample Value       Sample         % of
               vs GXdB Range        Count        Total
                         O/O       52,296        92.7%
                         O/R        2,373         4.2%
                         R/O          298         0.5%
                         R/R        1,427         2.5%
        Sample Value Invalid            0         0.0%
                    ------------------------------------
                       Total       56,394       100.0%

--------------------------------------------------------------------------------
DISTRIBUTION OF ACCURACY BY DEMOGRAPHIC VALUE
       Description                 Sample       % of       Sample    Estimated
          of Value     Value        Count      Total      Correct     Accuracy
         Homeowner       O         54,669      96.9%       52,296        95.7%
            Renter       R          1,725       3.1%        1,427        82.7%
             Other   Invalid            0       0.0%            0
                    -----------------------------------------------------------
                      Totals       56,394     100.0%       53,723        95.3%

--------------------------------------------------------------------------------
Note--99% C.I.: The Confidence Interval of the Estimated Accuracy assumes a
binomial distribution of the data.  Lower confidence intervals denote higher
confidence in estimated accuracy.
```

Fig. 2. Example of an Analysis Operator Report

Reference Database Considerations

Although some of the considerations discussed for the Database and Analysis Operator may seem obvious for some readers, the system represents our organizations' first attempt to build a shared, accuracy-analysis tool, and we could find little guidance to pave the way. Although not intended to be comprehensive, the following considerations were the ones that we found to be most critical in the successful implementation of the system:

Data Independence between Reference Database and Products

In the early development of the system, one of the most difficult messages to get across to the Data Product Organization was the importance of maintaining complete independence between the Reference

Database and the products being measured. When data product managers discovered that we were building a high quality database of consumer information, often their first reaction was to request a copy to be incorporated into their products. "After all," they would say, "wouldn't that improve its quality?" Of course, what they were not thinking about was the fact that it would immediately invalidate any measurement of their product's accuracy based on the included data.

A subtle but no less compromising situation arose when product teams requested a copy of the survey responses to perform record-level analysis of their quality failures. Although the Data Quality Team encouraged analysis as a follow-up to measurement in the TDQM cycle, it was difficult to get teams to overcome the temptation to "correct" their data to agree with survey responses when differences were found.

In order to avoid these situations, the Data Quality Team found it necessary to require any product team requesting access to the survey response data to first sign a user agreement with strict prohibitions against these practices. The agreements must be signed by the product manager, list all of the associates who will be handling the survey data, and outline its specific use. Product teams are also responsible for securing the survey data and prohibited from sharing it with anyone else, inside or outside of the organization, who is not listed on the agreement.

Self-Reported Data Only

An equally difficult message to convey was the need to acquire only the highest quality data--in particular, self-reported, primary source data. Many stakeholders urged the Data Quality Team to consider using so called "high quality" secondary-sourced data. For example, there were arguments that deed transaction data could be used. Their argument was, "Surely it must be 100% accurate because no one would want to put incorrect information on such and important document where property ownership was at stake."

Calls to use secondary-sourced data usually arose in connection with cost reduction, particularly for telephone survey data. Product managers were accustomed to paying for data that cost a few dollars per thousand, not dollars per record.

Decision to Outsource Telephone Surveys

Early in the development of the system, the decision was made to augment purchased in-home survey data with our own telephone surveys. There were several reasons: the quantity and quality of data available, the latency often being six months or more from confirmation to availability, and the need for particular items and data formats specific to our products.

At the same time, the team realized that conducting telephone surveys is not one of the core competencies of our organization; therefore, it would be necessary to outsource this function. After working with two vendors with limited success, we were able to develop a relationship with a marketing research firm that has proved to be very successful.

The key factor in the success of this relationship has been that the vendor understands what the team is trying to accomplish and is willing to work closely with the team to make changes and adjustments to the process as needed. For example, by jointly designing the survey script, the team was assured that the items that it needed were collected, while the vendor was able to assure that it would be a "reasonable" script that could be executed efficiently and in quantity.

During the initial phase of the survey development, the vendor was paid on the basis of time and resources. However, after several months of experience, there was enough information available on calling time and survey completion rates to convert the contract to a cost-per-survey basis.

Generating a Call Sample

One of the responsibilities of the Data Quality Team is to provide the telephone survey vendor with a file of telephone numbers to call. The main reasons that the team wanted to retain this responsibility were to assure a random coverage of telephones and to have the ability to seed the call sample with "quality control" records in order to estimate the rate of false confirmations in the responses returned by the vendor.

The typical process for generating the call sample starts with a large internal file of about six-million consumers that is already a random sample of records from a data product with national coverage. From this file, an n-th process is used to select approximately 200,000 records.

Because the starting file represents data from one of the products that will eventually be measured using the survey responses, an additional step is added assure that the process is not entirely "self-referencing." The additional step is to "roll" the last digit of each of the 200,000 telephone numbers from 0 to 9, creating an expanded file of 2,000,000 records.

Before delivering the two-million record file of telephone numbers to the vendor for calling, the team performs three additional steps. The first is to apply the names and addresses of the last known user of the telephone number, when available. Having a name and address to confirm makes the interview faster (and less costly) and improves the survey completion rate. The second step is to apply standard "do not call" suppression.

The third step is to seed the file with a number of "quality control" records. These are records for which the name and address have been deliberately replaced by reasonable, but incorrect (fake), values. The purpose of these records is to provide a means to assess the rate at which survey respondents (or interviewers) provide false confirmations. Because the interviewer is usually asking the respondent to confirm previously populated name and address data, there are instances where the information is confirmed, even though it is not correct. As a further control, the vendor is not told which records are quality control records. When the vendor delivers a file of survey response records, it is checked to see if any of the quality control records are present and which, if any, were returned with the original data. The observed rate of false confirmations is used as a crosscheck on the vendors' internal quality assurance processes and as an estimator for the intrinsic error rate in the survey responses.

Quality Assurance Processes

In addition to the quality control records described above, a number of other quality assurance processes are used to make the Reference Database of survey responses as accurate as possible. Perhaps one of the most important processes is that the survey vendor has agreed to record 100% of the interviews on tape. All of the interviews are played back by a second person who checks that interviewers entered the respondents' answers correctly. The ability to provide this service was an important consideration in the selection of the vendor.

As the Data Quality Team receives each weekly feed from the vendor, the responses undergo further quality assurance processing and review. First, the responses are profiled to assure that all values are within acceptable ranges on a per-record basis. In addition, the aggregate response statistics are calculated and compared to previous feeds for consistency and for assurance that service levels are being met. As an example, the Monthly Aggregate Demographic Response Rate (MADRR) is the total number of survey response values returned as a percentage of all possible values that could have been returned. By service level agreement, the MADDR value must not fall by more than 5% in any two consecutive months. If it does, then there are prescribed financial penalties in the contract.

Other measures are for timeliness of confirmation, for quality of addresses as measured by their ability to be zip+4 coded, and for the filtering and assessment of seeded quality control records as described above.

In a final step, all records in the Reference Database, both old and new, are processed through an internal "change-of-address" product. If the process indicates that a change of address has been reported for an occupancy record since the date of its last survey confirmation, then the survey record is flagged as no longer current. In keeping with the principle of self-reported, primary source data, the new address provided by the change-of-address product is not used or assumed to be correct. These "downgraded" records are kept in the database but are not extracted to the indices used by the Analysis Operator for estimating the accuracy of occupancy.

Recent Versus Current Information

In this regard, the Data Quality Team differentiates between "recent" occupancy versus a "current" occupancy. A recent occupancy is one for which relatively little time has elapsed since it was confirmed. On the other hand, a current occupancy is one that is still believed to be accurate. An occupancy record can be current even though it is not recent, and *vice versa*.

As described earlier, confirmed occupancy records are tested each month against a change-of-address file under the assumption that if there is not match (i.e., no indication of a move), then the occupancy is still current. However, it is known that the change-of-address file being use is not 100% complete and that some individual and household moves are not represented in the file. Therefore, a time related "degradation" of individual occupancy records exists in the database. For each occupancy record flagged as current in the database, there is a small probability that it is not current because of an unreported move; furthermore, that probability increases over time. Consequently, as a matter of policy, survey responses with confirmation dates older than three years are not extracted for use by the Analysis Operator in estimating the accuracy of occupancy.

Analysis Operator Considerations

In addition to the considerations for the Reference Database, there are also a number of issues related to the Analysis Operator itself.

Obtaining Meaningful Match Rates

Using the 2000 Census estimate of housing units as an estimator for occupancy, the total is around 118 million [3]. Because the survey sample is relatively small (250,000), the expected probability that a particular name and address in a consumer product will match an occupancy record in the Reference Database is relatively small. In practice, it has been found to be roughly 0.1%, or about one match per thousand input records processed.

On the other hand, the statistical model for the accuracy estimate assumes a simple binomial distribution of the data. Therefore in order for the user to obtain reasonable confidence intervals for the estimates of accuracy, the target should be to have at least 100 sample matches per measurement, preferably on the order of 1,000. At the random rate, sample sizes need to be between 100,000 and 1,000,000 records to achieve these targets. However, by providing product teams with profiles of reference database zip codes, address links, consumer links, and phone numbers,, the users are able to pull samples that produce much higher match rates than expected by random selection.

Data Formats and Comparability

Although name and address elements of occupancy are somewhat standard, other demographic elements can be presented in a variety of formats and value ranges in different products. In the initial design of the system, the formats were selected that corresponded to the formats used in the organization's leading consumer data product. However, both Reference Database and Analysis Operator were designed in such a way that new formats and elements could be added with minimal modification to the system.

Simple format changes such as date (MMDDYYYY to YYYYMMDD) and one-to-one mappings (like

Y/N to 1/0) are relatively simple to implement. However, ranged data presents more of a challenge. For example, one product may present home value as a series of letter codes where each letter represents a range of values (for example, B = $25,000 to $49,000). However, another product, especially a competitor product, may use not only different codes but codes that represent different, and possibly overlapping, value ranges.

In general, it is the user's responsibility to convert non-supported formats and values to supported formats and values. The Data Quality Team periodically adds support for new formats and values based on user demands and internal analysis needs.

User Access Control and Non-Disclosure

From a policy standpoint, there are two important non-disclosure issues related to the Geneticx AccuCheck System. The first relates to survey information stored in the Reference Database. Because the survey data represents an important investment of the organization and was collected under assurances to the respondent of non-product, non-marketing use, great care must be taken that individually identifiable information is protected and used only as intended. The second relates to the use of the measurements produced by the Analysis Operator, particularly with respect to accuracy measurements of the organization's data products, considered proprietary information.

Individually identifiable information and the report information are addressed through two separate user agreements. The execution of a Reference Database User Agreement is a prerequisite to any product team obtaining survey response records from the database. The agreement prohibits not only disclosure of the information to a third party but also prohibits adding the information to a data product or correcting a data product to agree with the survey information. It also requires the user to secure access to the information from unauthorized users and limits the time and purpose for which the information can be used.

Users of the Analysis Operator do not have access to individual survey responses from the Reference Database, only aggregate report information. Therefore, they are required to sign a different agreement, the AccuCheck User Agreement. This agreement limits the disclosure of measurement results to third parties without specific permission from data product leadership. In addition, the receipt of a signed agreement is what triggers a change in the Geneticx User Access Control System that allows authorized users to call the Analysis Operator.

User Education and Acceptance

AccuCheck user education entails not only the mechanics of using the operator, but it also involves some basic training in data quality, especially the concept of quality dimensions [2]. Product teams often confuse, or at least substitute, other quality dimensions for accuracy. Most often, it is substituting consistency for accuracy. The argument is that if product is already acceptable to the marketplace, and it isn't changing much (i.e., is consistent from build to build), then it is okay. Most of these obstacles have now been overcome, and there is now more openness to "knowing the truth" about data accuracy, no matter how ugly it might be in some cases.

Another important factor in the acceptance of the system revolves around trust--in particular, trust in the reference data. This was achieved through education, not just about how the system works but also about the overall process such as how the survey is done and how the reference data is maintained.

Other education issues deal with how to read and interpret the analysis reports and the limitations on the precision of the accuracy and coverage estimates they present, especially as a function of sample size.

CONCLUSION

Although it has proved to be a relatively difficult undertaking, the development of a shared survey response database and operator for accuracy measurements has clearly increased our organization’s TDQM capability and maturity. The savings realized by reuse and sharing of the survey information has more than compensated for the cost of the continuous telephone survey campaign.

Advantages

- No one team has to bear the entire cost of conducting a survey; therefore, the cost of accuracy measurement per product is reduced.
- Because all products teams contribute to the cost of the survey, it is possible to build a larger pool of shared interview data than could be gathered by any one team, resulting in higher analysis samples.
- Product teams do not have to manage the survey, Reference Database, and Analysis Operator. The Geneticx Data Quality Team, a team independent of the product teams, has the responsibility for managing the survey.
- Because the Analysis Operator is deployed as part of the organization’s standard processing system that is familiar to all product teams, there is universal access to the tool (with proper authorization) and little training required for user setup.
- Accuracy measurements are consistent from product to product and from measurement to measurement of the same product.

Limitations

- Product teams can only measure the accuracy of the demographic items that were pre-selected to be on the “standard” survey. Although additional items can be added to the survey, they increase the survey’s cost and require a relatively long lead-time between their inclusion and the time at which there are enough responses to obtain meaningful measurements. The current set of three occupancy responses and fifteen demographic responses is based upon a consensus of product team “care abouts” during the initial design of the system.
- Because of privacy considerations, the Analysis Operator provides printed reports only based on aggregate data. Therefore, product teams that wish to analyze record-level results must perform a separate analysis using the raw response data extracted from the Reference Database.

Future Work

Future work calls for refining the statistical model for the system, primarily based on empirical data gathered during the past two years of system operation. Model elements of particular interest are estimates for survey error, estimates for “leakage” in the change-of-address processing, and weights for the reference data.

REFERENCES

[1] Huang, K., Y. Lee and R. Wang, *Quality Information and Knowledge*. Prentice Hall, Upper Saddle River: N.J., 1999.

[2] Talburt, John R., Data Quality, Acxiom Corporation White Paper, February 18, 1998.

[3] US Census Bureau website, http://www.census.gov/, Table of Time Series of Housing Unit Estimates