

Real-World Database Marketing Experiences Using SAS Software

Robin Way
Director, Market Research
NW Natural
Portland, Oregon

Abstract

Database marketing is a proven method for improving the efficiency of targeted marketing campaigns in generating sales leads and reducing marketing costs, impacting the bottom line directly. Various elements of SAS software enable marketing analysts to design and implement powerful and highly tailored database marketing efforts unique among software systems. This paper will describe the lessons learned from several database marketing campaigns conducted by NW Natural, a natural gas distribution company serving Oregon and Washington. These insights include data quality issues, predictive modeling, target selection, and metrics used in evaluating the incremental value of database marketing techniques over more traditional methods.

Introduction

Businesses that want to build consumer awareness and preference for their products but believe that mass-market advertising doesn't offer a credible payback will be attracted to database marketing. The price of entry into the field of database marketing includes customer research, market data, and advanced analytic resources. The upside of this investment is the capability to conduct verifiable, pinpointed marketing campaigns that identify and focus on your best prospects and customers.

NW Natural, a natural gas distribution company headquartered in Portland, Oregon, has a long history in marketing and sales, and recently added database marketing to its toolset. To address objectives including gas equipment promotion and lead generation, NW Natural marketing staff designed and implemented database marketing resources and techniques.

This paper describes the data resources and modeling and target selection techniques used in an actual campaign at NW Natural. This paper also describes the unique role of SAS software in meeting a wide range of data management, analysis, and reporting needs affiliated with this project.

Business Need

Three specific business needs were addressed in the "1997 NW Home Comfort" campaign. These needs included boosting awareness of the company's new corporate image, promoting the benefits of residential gas appliances, and generating sales leads. The campaign centered on a full-color, high-value magazine titled "NW Home Comfort" edited and designed by NW Natural and its media partners. 100,000 copies of this magazine were printed, of which 76,000 were sent via direct mail to NW Natural customers and prospects in our franchise area. The direct mailing list was developed through our database marketing techniques, which will be described subsequently.

Company image

NW Natural recently decided to change its name by dropping the term "gas" from "Northwest Natural Gas." We did this to broaden consumer expectations for the kinds of markets the company might enter in the shift from regulation to deregulation of the retail energy industry. At the same time, we wanted to reinforce our core values as a local company with Northwest values.

The 1997 NW Home Comfort campaign was designed to create broader awareness of the company's new name and logo as well as solidify the message about its corporate direction. Specifically, we wanted to strengthen the

association between NW Natural and the comfort and beauty of residential gas appliances, as well as the reliability and value of natural gas itself.

Gas equipment promotion

NW Natural profitability increases as customer demand for natural gas increases, since we earn revenue on every therm of gas we ship over our distribution system. Demand for natural gas increases in turn as customers add new gas appliances in their homes and businesses. Therefore, promoting gas equipment directly ties to the bottom line and increases market share.

Conversions and added load

Another primary objective of this campaign was to generate sales leads from our non-customer base to switch their furnaces, water heaters, kitchen ranges etc. to natural gas, resulting in customer conversions. We also wanted to increase the number of gas appliances installed by current gas customers by adding appliances like barbecue grills, hearth products, and clothes dryers. Business reply cards and sales lead tracking systems were used to measure response to the campaign, and were forwarded to our sales force for follow-up calls and visits.

Project Components

A typical database marketing project consists of five components: planning and designing the campaign; collecting and managing data resources; conducting predictive modeling; selecting the best targets for the campaign; and campaign impact evaluation. Planning and design is germane to most projects (e.g., determine objectives, identify and assign resources, set timelines and deliverables), and won't be covered here in any more detail.

Available data resources are a primary constraint on the sophistication of modeling and target selection. It is important to understand the reliability and validity of data resources up front in the project planning process. Important steps in data management are collection, visual inspection, reviewing descriptive statistics, dataset merging and concatenation integrity checks, and informal hypothesis testing.

Data analysis and model development follows the data assembly process. Model development offers benefits over expert judgment-based approaches because hypotheses can be tested and verified and ROI measured. In this campaign, we selected a powerful non-parametric technique available in SAS software similar to the Chi-Square Automatic Interaction Detection (CHAID) approach. However, other modeling approaches are available that might be appropriate given project objectives, including linear and logistic regression and neural nets. CHAID-style analyses do not require strict multivariate normality assumptions, and results are easily interpretable and actionable, making this approach ideal for this campaign.

Next, targets are selected based on the modeling results, with highest priority given to the targets with the greatest likelihood of desired response to the marketing campaign. In many campaigns, the modeling and target selection are based on the same dataset. In the NW Home Comfort campaign, we used one data source for modeling and a related but different data source for target selection. Care needs to be taken in this circumstance so that modeling results are applicable and appropriate across data sources.

After targets have been selected and the campaign is in the field, it is important to measure the impacts of the database marketing effort, so that return on investment can be calculated and the predictive modeling tuned if necessary.

The remainder of this paper addresses each of these database marketing project components in more detail.

Data Resources

NW Natural has assembled a sizable marketing data library to address a number of strategic marketing initiatives, among them database marketing campaigns. In the 1997 NW Home Comfort campaign, five major data sources were used in the predictive modeling and target selection project components.

Residential market tracking studies

Since 1981, NW Natural annually conducts a residential market tracking study to stay in touch with trends and dynamic market forces. The data collected from these studies also makes an excellent resource for analytic projects, including database marketing analysis and modeling. The sample for this survey includes NW Natural customers and non-customers, and collects information on respondent demographics, housing characteristics, saturation of energy-using appliances, appliance purchase intentions, advertising awareness, and attitudes and opinions about energy companies and issues.

In the 1997 NW Home Comfort campaign, household response data (i.e., whether or not the household is likely to convert to natural gas or add new gas appliances) were not available in the target selection (i.e., "population") dataset. We surveyed a stratified random sample of the households in our franchise area and generated our model based on the collected survey data. Next, from the population of households in our franchise area, we selected those ranking in the top tiers of our model-based profile of households most likely to respond to the marketing campaign. In this way, the predictive modeling provides a bridge between the sample and the population.

Customer Information System

NW Natural maintains a central customer information system (CIS) for billing, customer service, and transaction record-keeping purposes. An affiliated CIS data mart is used for ad hoc queries and canned report generation. The 1997 NW Home Comfort campaign used fields from this data mart including customer names, mailing and site addresses, billing histories, and gas equipment profiles.

Census data

The federal decennial census of population and households is a rich source of consumer statistics aggregated at the level of zip codes and census block groups (i.e., each typically contains between 200 and 500 households). In addition to the core demographic statistics such as income,

age, education, and household size, the census collects housing characteristics and equipment saturation useful in validating our residential market tracking studies. Third-party marketing vendors have appended purchasing behavior data to the census and applied sophisticated marketing models to this dataset, with results useful to database marketers.

Property tax assessor data

Many counties make their tax assessment records available to the public. Although each county maintains different data collection and maintenance standards, it is possible and generally worthwhile to acquire this data either directly from the county tax assessor or through third party vendors. Extensive data cleansing, validation, and standardization are required to compile reliable data across counties. Useful data fields include property owner names, site and mailing addresses, and housing characteristics (square footage, number of stories and rooms, sales price and date, owner occupation).

Marketing Geographic Information System

NW Natural is developing a geographic information system (GIS) tailored to the needs of its marketing and sales division. GIS provides a unique approach to data visualization and analysis, and also presents inherently unique spatial functionality not available in tabular software packages. By overlaying census block group boundaries, gas main maps, and geocoded customer and property tax assessor locations, NW Natural can determine the proximity of customers and non-customers to existing gas mains, which determines availability of service.

Predictive modeling

Developing a predictive model consists of three steps: specifying dependent and predictor variables, specifying and refining models, and analyzing results.

Dependent variables

For the 1997 NW Home Comfort campaign, we actually developed distinct models for two dependent variables. First, we wanted to develop a profile of non-customers who were likely to

convert their major energy appliances to natural gas. This generally includes converting a furnace and water heater. Second, we also wanted to develop a profile of existing customers who were likely to add new gas appliances to their existing mix. This might include addition of a gas hearth product, range, barbecue grill, clothes dryer, or spa or pool heater.

The NW Natural residential market tracking studies explicitly identify actual purchasing behavior on the part of customers and non-customers. Based on these survey variables, we developed two ordinal scale dependent variables to capture the range of behavior we felt appropriately modeled the available alternatives.

The scale for the non-customer conversion variable was 1= have converted within the last 12 months, 2= intend to convert in the next 12 months, 3= prefer to convert but do not intend to do so, and 4= neither prefer nor intend to convert.

The scale for the customer add load variable was 1=have added load within the last 12 months, 2=intend to add load within the next 12 months, 3=prefer to add load but do not intend to do so, and 4=neither prefer nor intend to add load.

Predictor variables

A range of predictor variables was available in the residential market tracking study database, but we faced the constraint of using only variables that are available in a reasonably similar fashion in the household population dataset. Some candidate variables include: building type (single-family, duplex/triplex, condo, or apartment), year home built, home size (square footage), own a fireplace, home ownership, years in home, plan to move in next year, age of head of household, income category, number contributing to household income, number of children, and average monthly natural gas bill.

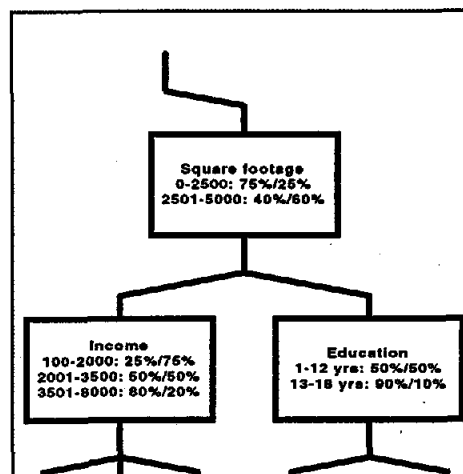
Running a CHAID-style analysis with the %treedisc macro

Overview of CHAID and %treedisc

The chi-square automatic interaction detection (CHAID) technique is a powerful non-parametric

method of building a response model whose output resembles a decision tree. This decision tree is built via recursive partitioning of a dataset into two or more subsets based on categories of predictor variables.

Figure 1: Hypothetical CHAID Tree Output



The predictor variable chosen for partitioning is the one most significantly associated with the dependent variable based on a chi-squared test of independence in a contingency table. The number of subsets made at a single node ranges from two to the number of categories in the predictor variable

Once initial subsets are created, each subset is further partitioned by the same algorithm based on full set of predictor variables, without regard to other subsets. This process continues until a stopping criterion—the p-value on the chi-squared test of independence—is met. Small p-values indicate the observed association is unlikely to have occurred solely through sampling variability (i.e., random chance).

The SAS %treedisc macro is similar to CHAID but is not strictly the same as published algorithm. SAS Institute applied several improvements and adjustments to improve performance and avoid endless feedback loops, but an option in the macro is available to specify the original CHAID algorithm if desired. You must license SAS/IML in order to run the %treedisc macro. %treedisc is available on the SAS Institute web site at this URL (current to 10/27/97):
<http://www.sas.com/techsup/download/stat/treedisc.sas>.

In order to run the %treedisc macro, you need to also download the %xmacro macro, available at the same URL. %xmacro is needed to perform procedure-like activities in the context of a SAS macro, including error checking, setting defaults, using variable lists, and doing BY processing.

The code for the %treedisc macro is typically straightforward, but many options are available for special circumstances and for controlling output. The options we use the most are to specify the input dataset, dependent variable, and predictor variables, and to indicate the predictor variable scale. Predictor variable scales can be nominal, ordinal and ordinal with a float value. Floats are used with missing or miscellaneous values to specify which end of the scale is associated with these values (e.g., is a joker high or low in a poker hand). Predictor variables with continuous scale are acceptable in the model but seriously impact performance; consequently it is recommended that you convert these into ordinal variables by enforcing an appropriate categorization.

Figure 1: Sample code to run the %treedisc macro

```
%treedisc(
  data=rts,
  outtree=tdout.&depvar,
  depvar=&depvar,
  nominal=&pvarsnom,
  ordinal=&pvarsord,
  options=list,
  trace=none);
```

Various options also exist to manage calculations, tree structure constraints, printing, etc. A highly graphical decision tree output can be created via PROC NETDRAW if your site licenses SAS/OR; otherwise tabular and list output are available by default. List output is customizable but harder to interpret quickly due to the amount of paper generated with a large tree.

Figure 2: Sample of %treedisc List Output

```
D_INCOME value(s): Refused $25,000 or
less $25,001-$35,000 $35,001-$45,000
$45,001-$55,000 DV counts: 731 679
Best p-value(s): 0.0001 0.0001

D_BLDTYP value(s): Refused
Single family dw DV counts: 565 631
Best p-value(s): 0.0001 0.0001

YRBLT value(s): . DV counts:
87 79 Best p-value(s): 0.2251 0.2701

YRBLT value(s): 1990 & Newer
DV counts: 67 103 Best p-value(s):
0.0077 0.0157
```

A few technical notes on the use of %treedisc: Predictor variables are treated by default in the order of their formatted values. Also, missing values are treated as a valid category, which may create difficulty in interpretation if results from a sample dataset are applied to a different (i.e., "population") dataset.

Analyzing results

Our experience with these techniques has yielded a number of useful rules of thumb for interpreting the decision tree-like output from CHAID-style analysis. Interpretation is not typically difficult or complex; it just takes time to digest and thoughtfully review the results. Sometimes it will be desirable to re-run the analysis with some data changes or macro options, and other times the structure of the results are easily apparent.

- First, if you have SAS/OR to create a graphical tree, by all means use it! This style of output is easier to interpret quickly than its tabular or narrative format.
- It is usually desirable to test alternative specifications of the dependent variable, particularly uncertainty exists over how to collapse or split levels of the dependent variable. A sensitivity analysis can be conducted by running the %treedisc macro for various specifications of the dependent variable and comparing tree structures. Similarly, predictor variables can be added or dropped from the analysis to test for this type of sensitivity.

- As part of the initial analysis of a new model, I always find it useful to simplify the macro's printed list output into a single-page summary of results.
- Scan the analysis output right off the bat to identify segments that really drive response on the dependent variable. For each node on the tree, I review the ratio between observations on the "good" or desirable side of the dependent variable to the observations on the "bad" or undesirable side. Nodes with high ratios, particularly when compared with the ratio at parent nodes and root nodes, indicate potential paydirt.
- Before concluding that a particular node is high-value (i.e., based only on the ratio of desirable to undesirable observations), make sure that the number of observations on the desirable side of the ratio is sufficiently large enough relative to the size of the analysis dataset. The meaning of "sufficient" depends on the objective of your marketing campaign and on gut feel—this will change based on whether you're selling groceries (high volume) or pianos (low volume). Presuming that the analysis dataset is statistically representative of the population to which the model will be applied, sufficient coverage in sample generally signals there will be sufficient coverage in the population dataset. Of course, you should always verify this hypothesis in the target selection phase.

Selecting Targets

After an appropriate predictive model has been developed, the model results are used to drive target selection from a larger market mailing list. NW Natural's experience suggests three issues are of key importance in this phase of the project: making the transition from the sample to the population dataset; merging data sources; and transforming predictive modeling results into selected targets.

Transition from survey data to real-world data

As stated above, the 1997 NW Home Comfort campaign used a sample dataset for model development that was different from the

population dataset from which targets were selected. Making the transition required care in design and interpretation of the predictive model.

- First, dependent variable categories in the sample dataset need to contain a large enough number of observations to permit several rounds of partitioning. If you categorize the dependent variable such that you only have 10% of your sample left in the "desired" category, you may run out of sufficient observations to make robust projections from the model to the population. Instead, consider alternative specifications for dependent variable categories when applicable, or you may need to collect more data.
- Drop all predictor variables from the sample that can't be found in the population. Even though it may be appealing to include a powerful predictor in the model in hopes of generating a superior estimator, it will get you nowhere when it comes to target selection. Remember that the goal is to apply what you know to what you don't know in hopes of improving your hit rate with the marketing campaign.
- Missing values among your predictor variables can create some tricky situations. Many survey samples will include a lot of missing data, and it's fortunate that the %treedisc macro treats missing values as valid, or else the useful dataset might be too small to use in many instances. Careful model inspection may reveal a branch under a missing value parent node that closely resembles the structure of a branch under a different parent from the same level of the tree. This indicates that you should consider changing the variable's structure to ordinal with a float, or bundling the missing value with another non-missing value by applying a SAS format.

Merging data sources

The NW Home Comfort campaign population dataset was created by merging several population datasets, each of which had a different origin. Here are some findings we developed during this process.

- Most importantly, merging datasets from different origins requires knowledge of the underlying data quality, and familiarity with compatibility issues. For example, examine key fields to make sure they will merge properly. Conduct descriptive statistical analyses to examine minima, maxima, distributions, and modes. Test the merging process with subsets of each data source.
- When using street address as a by-variable in a SAS merge, standardize the address field to US Postal Service (**Error! Reference source not found.**) standards. Software from companies like Group1 (**Error! Reference source not found.**) will help you in the address standardization process. Make sure the USPS has awarded the software package with CASS certification. In addition, your mailing house for the marketing campaign will need to perform this standardization anyway to take advantage of better pricing. In the near future, all bulk mailings will require at least this level of standardization, in addition to carrier routing and bar coding. Group1 also offers mailing standardization for Canadian addresses.
- Census data are now available from a variety of vendors at reasonable cost and make an excellent addition to your geodemographic database. Look for packages that offer not only the 1990 data (which is available nearly for free from the Census Bureau, **Error! Reference source not found.**) but also offer updates to the current year and a five-year projection. Since census boundaries are often included in many different data sources, ensure that census block group identifiers follow a standard convention. It's also a good idea to compare census-based results with other market indicators for consistency.
- Property tax assessor datasets provide a broad perspective on the consumer market, but special treatment of the data in these files is required. Each county builds their files following county-specific conventions, requiring cleansing and checks for data availability specific to each county. Data often arrive in a variety of database formats, from a variety of computing platforms, and on a variety of media. Some data formats are actually digital versions of formatted print files. The administrative burden of ordering and refreshing this data can be substantial (i.e., simply making contact with county officials, ordering data files, cutting checks, and waiting the 4-6 weeks for the files to be prepared and sent). Third-party vendors are beginning to offer solutions to the data collection and standardization issues on a county-by-county basis.
- Data recency is another issue requiring careful planning and cost-benefit calculations. One estimate states that 15% of all consumer listings and up to 40% of all business listings turn over each year. The impact of market turnover transforms any significant database marketing initiative into a process, rather than a project.
- Above all, I believe perfect database consistency and reliability is not the goal of database marketing efforts. The time and effort needed to create a perfect database are not likely to offer as good a payback as an 80/20 solution. Focus needs to be placed instead on reducing (to a reasonable and acceptable level, and in a timely fashion) the uncertainty created by mixed data sources and conventions, with the overall objective of creating sales leads and impacting the corporate bottom line.

Applying predictive modeling to target selection

Predictive models like CHAID create scores of target segments, each scored by their likely response rate in the form of a observations in each category of the dependent variable. These measurements need to be converted into a ranking scheme across all segments, which when applied to the population file will select the best target segments. This involves artfulness as well as science, especially when the model sample is from a different source than the population file, as we had in the 1997 NW Home Comfort campaign.

- One method for identifying target segments is to calculate the "response ratio" as the ratio of likely responders to non-responders, and then standardize this ratio around an index of 100. This makes it fairly easy to identify the highest-ranking segments.

- Care needs to be taken when working with segments with few observations—this happens near the bottom of the tree or for “small” branches. Even though a segment’s response ratio may look high, it may not be a robust predictor if too few observations contribute to the score. Since the CHAID approach is non-parametric, there is not a hard and fast rule to apply, but experience suggests a minimum of 30 observations and 5% of the model sample dataset are good standards with which to start.
- Once all response ratios have been calculated, they need to be appended to the appropriate segments in the population file. This is a good time to compare the relative frequency of targeted population segments to targeted segments from the model sample to ensure consistency.
- Next, tabulate the number of targets in each segment ranked by response ratio, and determine the campaign size (i.e., number of targets selected). This will depend on your budget, the number of waves in the campaign, the number of targets mailed per wave, shipping costs, and cost of materials and production for the marketing message itself. Another variant in this mix arises if you want to vary the type of message by segment.
- If time permits, it is usually preferable to start with a test mailing on a sizable subset of your population dataset. By tracking and comparing response to the test mailing with predictive modeling results, you can refine subsequent segment rankings, or even build a new model based on test results. However the number of responses from the test mailing (hence the size of the test mailing itself) must be large enough to get robust estimators. Note you will also not be capturing information on prospects that are interested but didn’t respond in test wave. Remember that repeat waves often are required to convert an interested target into a qualified lead.

Metrics for estimating value of database marketing techniques

There are several ways to estimate the value of database marketing. The first four are available directly from the results of predictive modeling and target selection activities, and thus are available immediately as measures of likely response to the campaign. The remaining metrics are based on post hoc comparison of model predictions to campaign performance, and also involve rigorous examination of strategic and customer-based factors outside the control of the modeler.

Coverage

As each node becomes partitioned, the number of observations in each successive node decreases inevitably. The ratio of observations in a child branch relative to its parent branch is a metric of the absolute size of the branch in the population. Branches that are too small relative to the root of the tree may not offer enough coverage in the population to result in a substantial number of leads. Coverage, therefore, identifies which branches of the tree contain greatest coverage of potential targets, and is also a useful metric to avoid moving too far down in the tree.

Response ratio

As mentioned above, the response ratio is the ratio of likely responders to non-responders at each node. This is similar to what epidemiologists call “relative risk.” The response ratio identifies the branches that if selected are most likely to respond to the campaign, which in turn will maximize the yield of a marketing campaign with constrained size.

Lift

This metric is commonly used in the field of marketing tactics, and definitions depend on the context of the campaign. I use the term “lift” to refer to the relative improvement of a response ratio for a given node over the response ratio for one of its parent nodes, or relative to the root node. This is conceptually similar to the odds ratio statistic (which uses proportions rather than

raw values). Lift identifies how much of an improvement is generated by further segmenting a particular branch. Lift also provides balance with the coverage metric, by guiding the analyst to move further into the tree structure to leverage better response rates at the expense of potentially smaller coverage.

Cost savings for reduction in send-out

While coverage, response ratio, and lift are focused on the added revenue of the campaign, this metric focuses on how much the analytic efforts reduce costs relative to a campaign that didn't rely on database marketing (i.e., random selection). Assuming that the absolute number of responses remains the same and costs are reduced by trimming out large groups not likely to respond, this metric is a primary component of predicted ROI. Cost savings are calculated as the ratio of number of targets in selected branches of the tree to the number of targets in entire tree.

Assuming campaign budgets have been set in advance of launching the database marketing effort, cost savings through avoidance of non-responding targets can be parlayed into an increased send-out to high-response targets. This budgetary "snapback" effect can either be converted into repeat contact with the same targets, or into expanded coverage of targets with more moderate response potential. If you choose the latter tactic, be sure to assess how fast lift decreases for increase in number of targets selected.

Evaluate database marketing return on investment

The final short-run test for the campaign's success is whether your company makes enough of a return on its incremental investment in database marketing efforts to pass internal hurdle rates. The business problem can be stated this way:

Maximize:

Return = percentage increase in leads /
percentage increase in investment

Subject to:

1. Campaign total leads must be large enough to address strategic objectives
2. Increase in staff time demand cannot exceed available reserve

Where:

Incremental investment = staff analysis costs +
database development costs + time to market

When the return on the database marketing campaign exceeds your company's internal rate of return (usually around 12 to 15 percent), then you've got an unqualified success—Marketing and Sales will be happy with the increased leads and Finance will be happy with the returns.

Evaluate model performance

While the short-run test for the database marketing efforts are externally-focused on financial return, one of the long-term metrics for success is internal improvement in database marketing performance. This includes:

- reducing analysis costs by building a sense for what works and what doesn't
- reducing time to market by automating and streamlining the process
- reducing the risk of these techniques by refining and tuning your predictive models.

There may be no tangible metrics for the first two long-term objectives, but there are a variety of ways to measure success toward the third objective. First, track the actual response ratio per segment against what was predicted in the model. If responses are smaller than expected but the actual response share per segment mirrors what was predicted, then the model is working the way it should and perhaps the marketing message itself is what needs to be improved. (Side note: Always rejoice that you've got the data on your side, as opposed to advertising and

public relations types who've got nothing more than their credibility and charisma on which to rely.)

A second technique is to evaluate the post hoc stability of the predictive model by running a second model using actual response and non-response data. Differences between the actual response tree and the predicted response tree highlight areas for model improvement.

Conclusions

Database marketing has become a useful tool for NW Natural marketers by improving the effectiveness of marketing campaigns and increasing sales leads. It offers cost savings over mass market campaigns, identifies target market segments, and inserts a feedback loop for marketing tactics improvement. SAS software is an indispensable tool for managing, analyzing, and presenting the results of these efforts.