

Finding Consumers More Accurately and Actionably Using Data Mining Tools

Louise Keely and Dimitar Antov

The Cambridge Group* White Paper
August 25, 2009

Presented at the Salford Systems Data Mining Conference, San Diego, CA

Abstract

Determining how to type consumers into specific groups, or segments, is an important business problem for a variety of companies, including those in the financial services, retail, and consumer packaged goods industries. The quality of these typing tools can dramatically impact the effectiveness of targeted advertising and other marketing based on these segments. Typing tools built using parametric models, like those of discriminant analysis, are often used to predict classification into a consumer segment for the purpose of targeted marketing activities. However, our experience is that these models suffer from several issues including lower predictability, inability to tolerate missing values, lack of robustness when applied to other samples, and lack of flexibility in model trade-offs between predictability, purity, and complexity. In contrast, classification trees have produced better results across these dimensions. In addition to being more accurate, classification tree construction allows for greater adaptability based on business needs, e.g. the informational requirements of the typing tool. We provide a case study in which classification trees provided superior accuracy and actionability to a financial services client.

* Contact information: Louise Keely lkeely@thecambridgegroup.com or Dimitar Antov dantov@thecambridgegroup.com; Mailing address: The Cambridge Group 227 W Monroe St, Ste 3200, Chicago, IL 60606; Website: <http://www.demandstrategy.com>

1. Introduction

Most of our clients successfully manage and grow their businesses by targeting specific consumer or customer groups. For many B2C corporations and even some B2B businesses, some type of consumer segmentation is a fundamental tool for identifying their target customers and ensuring that their products and marketing activities are aligned with the benefits sought by those targets. These segments are simply groupings of consumers into cohorts defined by distinguishing information, such as behaviors, demographics, or motivations. Being able to efficiently and effectively find those consumers is therefore fundamental to maintaining and growing these businesses.

Any business that regularly uses a consumer segmentation to make business decisions also relies on typing tools as an input into that decision making. Typing tools are algorithms for predicting consumer segments among any consumer sample using a set of pre-specified variables. Typing tools are most often used in on-going research but also can be used for targeted marketing to a customer database. Typing tools provide users with a way to quickly categorize any consumer into one of the segments. Typing tools are built using the original sample that created the consumer segments, but are intended to be applied to other existing or new datasets to produce predicted segment classifications.

A key advantage of typing tools is that they classify consumers into a segment using a much smaller set of variables, or even different variables, than were originally used to construct the segmentation. It is not uncommon for a segmentation to be based on 100 or more variables, whereas typing tool algorithms are built on a much smaller subset – as few as 5 or up to 25. The exact number of typing tool variables usually depends on the application and is tailored to the venue of data collection. For focus group recruiting, for instance, the interviewing time of respondents is limited to a few minutes so that typically no more than 10 questions can be asked. For internet surveys a larger number of predictors can be used (but there will still be a limit due to survey length constraints). Similarly, typing tools allow for the inclusion of new variables that are not part of the segmentation itself but are predictive of segment membership e.g., demographics used to predicted membership in a motivational segmentation. For managers who will be using typing tools to execute on their strategies for growth, the ability to tailor typing tools to specific needs is an important and valuable feature of any typing tool method.

There are standard techniques used by consulting and market research firms to construct these tools. The most common are based on parametric approaches such as discriminant analysis. These approaches offer a number of key benefits, such as wide industry acceptance, simplicity, and ease of construction. However, we have found that typing tools constructed using these standard parametric techniques fall short on several dimensions relative to Classification and Regression Trees (CART), a non-parametric data mining technique.

As an alternative to the standard discriminant analysis-based tools, we have developed several variations of typing tools, depending on our clients' needs, that leverage CART as the predictive algorithm. In our experience, typing tools based on CART have at least seven important advantages over the more standard methods. These include (in the order discussed below):

1. Increased customization of the tool's inputs and outputs
 - a. Toleration of missing values in the predictor variables through the use of surrogate variables
 - b. Ability to favor target segments' predictive success in the tool construction
 - c. High degree of flexibility in iteratively choosing which variables and the number of total variables to include
 - d. Flexibility to finely tune the trade-off between higher predictive success and higher purity (as they are usually in tension with one another)
2. Greater performance and usability of the completed tool
 - a. Typically higher levels of both predictive success and purity[†] with the same or fewer predictor variables
 - b. Ability to adapt the tool algorithm for use in different research settings: small-scale or large-scale typing, computer-based typing or pen-and-paper typing
 - c. Greater robustness of the results in subsequent research

The rest of the paper proceeds as follows. We review key features of standard parametric models used to build typing tools. We discuss the process for building a typing tool algorithm using CART as well as the outputs, providing more detail regarding the advantages listed above. We also point out a key disadvantage of CART. We provide a case study where CART outperformed standard tools in several dimensions.

2. Key features of standard parametric model-based predictive algorithms

We do not exhaustively review the processes for parametric modeling of segment membership here, as that is beyond the scope of this article. Rather, we highlight some of its key features relevant to the comparison with tools built using CART.

Standard statistical software allows running typing tool algorithms based on discriminant analysis or multivariate discrete choice models (multivariate logits or probits) as soon as clustering is completed. Discriminant analysis carries the

[†] Predictive success is defined as the rate at which observations that are actually in a segment are classified into that segment by the typing tool. Purity is defined as the rate at which observations that are classified into a segment by the typing are actually from that segment.

assumption of normality in the independent variables. Discrete choice models do not assume normality of the predictor variables but have a more involved set up process. If the normality assumption approximately holds, discriminant analysis-based tools generally have higher predictive success than discrete choice-based tools. Discriminant analysis is the predominant method used. The build time for a typing tool using discriminant analysis is relatively short because much of the process can be automated. Very little user input is needed except for initiating the building process and choosing the number of variables based on the output. The output is standardized and easily integrated into a typing tool interface. In our view, the time savings of discriminant analysis-based tools is the key advantage of these tools relative to CART-based tools.

On the other hand, the distributional assumptions on which discriminant analysis relies can lead to lower performance and lower stability than non-parametric alternatives. We next discuss CART, one such alternative with which we have had significant success.

3. Review of processes to build a typing tool using CART

Building CART typing tools is more involved. The increased flexibility of CART mentioned above comes with a set of choices that need to be made prior to constructing a classification tree. There are additional choices to be made after a tree is constructed but before the optimized splitting criteria are finalized. This flexibility leads to the construction of a typing tool that has the strongest possible performance and that best serves the specific business objectives.

Among the many set-up options available in CART, we have found critical those related to the splitting method, the option to impose higher costs for particular misclassifications, and the missing value penalties. Contrary to discriminant analysis, CART-based algorithms are capable of tolerating the incidence of missing values in the predictor variables.

In the instance when a value is missing for a given data point, a surrogate variable, that is highly correlated with the variable on which there is no data, is used instead. Therefore, CART typing tools are extremely suitable for applications to data with gaps and lack of uniform quality across all respondents. Even when the pattern of missing values on a new dataset is drastically different from the build data pattern, the typing tool does not need to be adjusted. The user has the option to specify the degree of tolerance of missing values, which in turn affects the selection of splitting variables in the tree.

In contrast, typing tools based on discriminant analysis require non-missing data for every respondent. Standard (and more creative) methods to impute missing values can lead to non-robust results.

Other choices available are the selection among different splitting methods and the cost for certain misclassifications. The specific use of the tool will inform these choices. When targeting focus falls on some clusters but not others, the CART modeler can impose costs for misclassifying observations into the focus clusters and effectively increase the prediction accuracy for such clusters. Similarly, the splitting methods yield another layer of desired flexibility. Each of the six available splitting methods has its own strengths that the modeler can select from, e.g., Ordered Twoing is best used when the predictor variables are ordered levels (customer satisfaction scores from 1 to 6 vs. mode of transportation ranging from bus, train, drive to walk), while Symmetric Gini works best when imposing classification errors.

The post-building options offer additional adaptability of the typing tools to business needs. The user has the ability to vary the size of the tree, which impacts the complexity of the typing tool algorithm as well as the number of variables required by the typing tool. In doing so, the user has the ability to trade off complexity and accuracy of the tool, depending on the setting in which it is to be used. This trade off is also available using discriminant analysis, albeit in a more limited way through the choice of the number of variables used.

Similarly, by expanding and trimming the tree, purity and predictive success are traded off. With a business objective of identifying as many of the targeted clusters, one can select the tree with the highest predictive success. If the goal is to correctly recruit respondents of a target cluster, the modeler would instead choose a tree with the highest purity rates. With discriminant analysis such a tradeoff is not possible; predicted success rates cannot be explicitly improved at the expense of purities and vice versa.

4. Performance comparisons of typing tools based on CART technique versus those based on discriminant analysis

In our experience, CART-based typing tools deliver considerable improvements in predictability compared with those based on discriminant analysis.

Switching to CART-based algorithms is usually associated with 5-15% overall gain in the rate of predictive success. This gain in the accuracy of classifications is *not* achieved at the expense of relying on greater number of predictors. In fact, we often find that predictive success is improved while simultaneously reducing the number of predictors used in the typing tool.

The ability to reduce the number of predictors in the typing tool without significant sacrifice of predictability makes CART-based algorithms attractive especially in applications when interviewing times are critical. Using a smaller set of predictors could yield substantial cost savings in recruiting of respondents and data collection.

CART-based typing tools offer the potential to alleviate data collection processes in other unique ways as well. The nature of the tree method is such that not every respondent needs to go through every splitting criterion to be classified into a cluster. This implies that, contrary to discriminant analysis, a respondent needs to answer only a subset of the questions that the classification algorithm employs.

This capability gives rise to “adaptive” or “dynamic” classifications where the response on a classification question determines what question follows next. The highly streamlined interviewing process is time efficient, with data collection minimized to only the relevant classification elements. In contrast, tools based on standard parametric analyses are static and cannot produce cluster classifications unless there is non-empty value on all predictor questions.

The adaptive nature of tree-based methods allows for customization in another unique way. We have produced “pen and paper” versions of typing tools that are used for qualitative research recruitment, for which interviewers do not use computers to instantly classify respondents. The pen and paper typing tool can be regarded as a horizontally transposed version of the vertical tree. Figure 1 provides illustration of such a tool.

After recording the response on the first question, the interviewer follows directions about which question to ask next based on the provided response and so forth until classification is produced. There is no manual computation of probabilities required and no need for a follow up interaction with the recruiting prospects. Classifications are produced instantaneously by following relatively simple interviewing and data management process.

An additional noteworthy feature of CART typing tools is the option to classify respondents in batch mode. Batch mode scoring simultaneously classifies each one of multiple respondents into one of the defined segments. A scoring code that is available in multiple programming languages can be outputted with any optimized tree the modeler settles on. This code can be applied on readily available datasets and also adapted in user-friendly interfaces in Excel.

CART-based typing tools also offer improved robustness and stability in cluster classifications. In comparison to standard tools, we have obtained higher degree of replicability of cohort distribution when applying the CART-based tools on parallel client studies. An example of this difference is provided in the case study below. We attribute this difference to the absence of distributional assumptions underlying CART typing tools and the related fact that tree algorithms are less sensitive to the effects of data outliers.

Figure 1: Pen and paper typing tool for in/out of cluster binary classification

General Instructions		Response Scale	
1) Start at row 1 and ask the question in the "Question" column below		Disagree Completely = 1	
2) Record answer to question, matching the response to the corresponding number in the "Response Scale" to the right		Disagree Strongly = 2	
3) Look to the "Instructions" column and go to the corresponding row depending on the respondent's answer		Disagree Somewhat = 3	
4) Repeat process until you get the instruction to "classify as L1" or "TERMINATE." At this point, move onto the next respondent.		Agree Somewhat = 4	
		Agree Strongly = 5	
		Agree Completely = 6	

Row	Question	Response	Instructions
1	I look for gum and mints that are refreshing but don't have too many calories		If answer is 1 or 2 then go to row 2 otherwise go to row 13
2	I hardly ever eat candy		If answer is 1 or 2 then go to row 3 otherwise go to row 10
3	I always have gum with me no matter where I am		If answer is 1-3 then go to row 4 otherwise go to row 7
4	I feel guilty when I eat too much candy		If answer is 1-4 then classify as cluster 1 otherwise go to row 5
5	I look for candy that is satisfying but doesn't have too many calories		If answer is 1 or 2 then go to row 6 otherwise TERMINATE
6	I usually eat what I want, not what I should eat		If answer is 1-4 then TERMINATE otherwise classify as cluster 1
7	I'm always looking for gum and mints that offer new tastes and textures		If answer is 1-3 then go to row 8 otherwise go to row 9
8	Eating chocolate is one of my greatest pleasures		If answer is 1-3 then TERMINATE otherwise classify as cluster 1
9	I hardly ever chew gum		If answer is 1-3 then TERMINATE otherwise classify as cluster 1
10	Eating chocolate is one of my greatest pleasures		If answer is 1-3 then TERMINATE otherwise go to row 11
11	I feel guilty when I eat too much candy		If answer is 1-3 then go to row 12 otherwise TERMINATE
12	I hardly ever chew gum		If answer is 1-3 then TERMINATE otherwise classify as cluster 1
13	I keep track of the fat and calories I take in each day		If answer is 1 or 2 then go to row 14 otherwise go to row 31
14	I hardly ever eat candy		If answer is 1 or 2 then go to row 15 otherwise go to row 27
15	I'm always looking for gum and mints that offer new tastes and textures		If answer is 1-3 then go to row 16 otherwise go to row 21
16	I feel guilty when I eat too much candy		If answer is 1-4 then go to row 17 otherwise TERMINATE
17	Eating chocolate is one of my greatest pleasures		If answer is 1-4 then go to row 18 otherwise classify as cluster 1
18	I look for gum and mints that are refreshing but don't have too many calories		If answer is 1-3 then classify as cluster 1 otherwise go to row 19
19	I feel guilty when I eat too much candy		If answer is 1 or 2 then go to row 20 otherwise TERMINATE
20	Eating chocolate is one of my greatest pleasures		If answer is 1-3 then TERMINATE otherwise classify as cluster 1
21	I always have gum with me no matter where I am		If answer is 1-4 then go to row 22 otherwise TERMINATE
22	Sometimes I chew gum, and other times I eat mints		If answer is 1-4 then go to row 23 otherwise TERMINATE
23	I feel guilty when I eat too much candy		If answer is 1-4 then go to row 24 otherwise TERMINATE
24	I usually eat what I want, not what I should eat		If answer is 1-4 then go to row 25 otherwise classify as cluster 1
25	I feel guilty when I eat too much candy		If answer is 1-3 then go to row 26 otherwise TERMINATE
26	I always have gum with me no matter where I am		If answer is 1-3 then classify as cluster 1 otherwise TERMINATE
27	Sometimes I chew gum, and other times I eat mints		If answer is 1 or 2 then go to row 28 otherwise TERMINATE
28	Eating chocolate is one of my greatest pleasures		If answer is 1-3 then TERMINATE otherwise go to row 29
29	I usually eat what I want, not what I should eat		If answer is 1-3 then TERMINATE otherwise go to row 30
30	I always have gum with me no matter where I am		If answer is 1-3 then classify as cluster 1 otherwise TERMINATE
31	I look for gum and mints that are refreshing but don't have too many calories		If answer is 1-3 then go to row 32 otherwise TERMINATE
32	Personally I hardly ever eat chocolate or non-chocolate candy but I keep it on hand for others		If answer is 1 or 2 then go to row 33 otherwise TERMINATE
33	I feel guilty when I eat too much candy		If answer is 1-4 then go to row 34 otherwise TERMINATE
34	Sometimes I chew gum, and other times I eat mints		If answer is 1-4 then go to row 35 otherwise TERMINATE
35	Eating chocolate is one of my greatest pleasures		If answer is 1-4 then go to row 36 otherwise classify as cluster 1
36	I usually eat what I want, not what I should eat		If answer is 1-5 then TERMINATE otherwise classify as cluster 1

5. A potential disadvantage of CART techniques

CART-based typing tools offer a number of advantages, but they do come with increased time costs both in the construction of typing algorithm and in the construction of a user interface for the typing tool.

In comparison to standard typing tools, the time for choosing the CART-based algorithm for the tool is significantly higher. Optimizing the tree classifications is dependent on the characteristics of the build sample and cannot be easily standardized. It requires exploring various options and assessing how their performance is aligned with the business need. As a result, the building process is iterative.

To ensure that business needs are optimally addressed, evaluation of the modeling results needs to be discussed with project leads and implications of the selected tree should be carefully considered. Size of the tree, complexity, number and type of predictors, predictive success and purities need to be finalized before building the tool interface, which is a multistep process of itself.

Translating the tree to produce either an adaptive typing tool or a pen and paper tool, as discussed above, entails substantial programming modifications of the code produced by CART. For tree-based algorithms, the output is a list of splitting criteria and rules for variable substitution in the event of missing values on the next-in-line splitting element. Depending on the size of the tree and the number of splitting elements, the rules that determine which cluster a given observation falls into could be quite elaborate and complex. Adaption of these rules to create an easy-to-use interface can also take significant time.

However, our experience indicates that it is important not to be short-sighted or “penny wise and pound foolish”. If a typing tool is to be used on an on-going basis by a business, then the costs of having a poor tool or one that is not suited to its application can ultimately far outweigh the upfront cost of constructing an optimal tool in the first place.

5. Case Study

The consumer segmentation we created for a financial services client was part of broader strategy development that we developed with them. A key strategic objective of our engagement was to align the portfolio of product and service offerings with targeted consumer groups. For the purposes of product positioning and feature optimization, we were going to conduct qualitative focus groups with both current customers and potential prospects among the target consumer segments. A typing tool was required to identify which subjects to include in our qualitative research.

Following the completion of the segmentation, we had a typing tool developed based on the standard discriminant analysis. A hold-out testing on a 20% random sample was used to assure that the typing tool would not capture too much sample-specific variation of the sample used to build the model. The tool was characterized with an overall predictive success of 81% and strong test sample results. Table 1 contains specifics on the performance of this typing tool. Given that the level of predictability was above a pre-established threshold of accuracy agreed upon with the client, we proceeded with applying the tool on additional datasets.

Table 1: Discriminant analysis typing tool performance summary

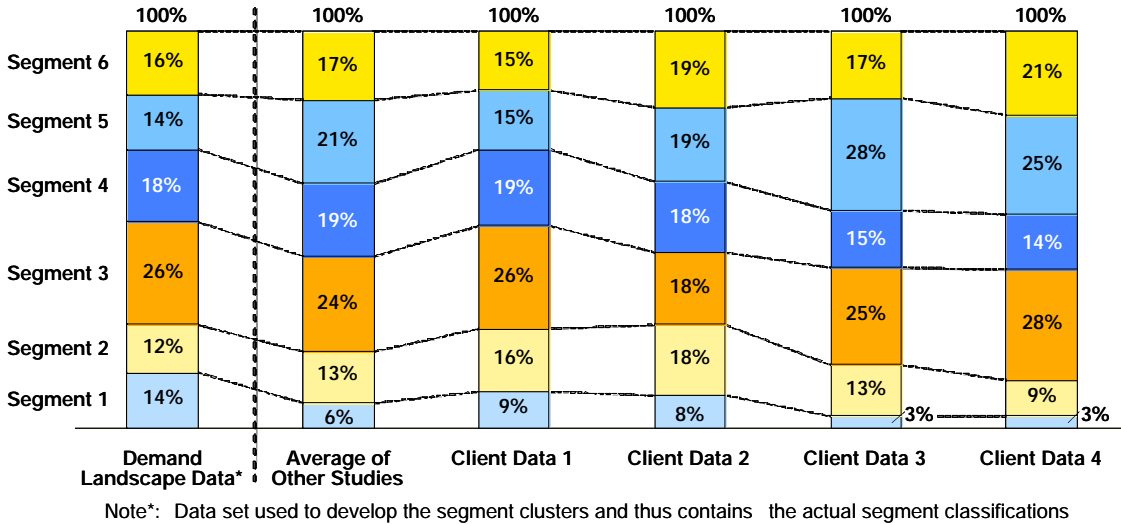
	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6
Predictive Success	73%	81%	86%	81%	79%	81%
Purity Rate	78	85	82	79	81	82
Predicted in Segment	11	13	28	18	14	16
Actual in Segment	12	14	26	18	14	16

We had identified two new and vibrant target consumer groups for our client (these are Segments 1 and 2 in Table 1). We needed a highly precise tool for use beyond this initial study to find current and prospective customers in these segments. That precision would be accomplished with a stable and precise typing tool. A typing tool would be used for ongoing research, including qualitative focus groups. In addition, a number of targeted marketing and promotion campaigns that our client was about to carry out were dependent on a typing tool.

We initially tested the robustness and stability of the tool it with other existing datasets that were specific to that client. Since the observations in these new datasets did not have actual segment assignments, we could not infer the predictive success or purity of the tool. The measure of “goodness of fit” of the typing tool we could use was the deviation of the distribution of predicted segments in the new data to the distribution of actual segments in the original data. We applied the typing tool to these new datasets and classified these other datasets’ observations into our six consumer segments. Once the scoring was completed, we compared the predicted distribution of segments across these different datasets to establish the degree of alignment with the original segment distribution.

The results we obtained were not satisfactory. See figure 2 for details. The size of the two target segments differed significantly between the original sample and the samples of the new datasets. The smaller size of Segment 1 was of particular concern. Furthermore, the over predicted size of segment 5 cast additional doubts on the overall robustness of the tool. The predictions for segment 5 were in some cases 2x the actual size of this cluster in the original dataset.

Figure 2: Segment incidence by study based on discriminant analysis typing tool



We therefore constructed a CART-based typing tool using the same original sample. We also restricted the tool to include the same number of predictive variables. The tree we selected was based on the symmetric Gini splitting method with high penalties for missing values and disproportionately higher costs of misclassification for the two target segments 1 and 2 because they were the target consumer segments.

The overall predictive success of the tree-based typing tool was 86%, 5 percentage points higher than the discriminant analysis tool. For our two target segments in particular, the predicted success and purity rates were substantially improved by 18% and 8%. With the exception of segment 3, where predictive success dropped from 86% to 82%, all other segments' classifications improved, with both predicted success and purity rates higher in the revised typing tool. See table 2 for details.

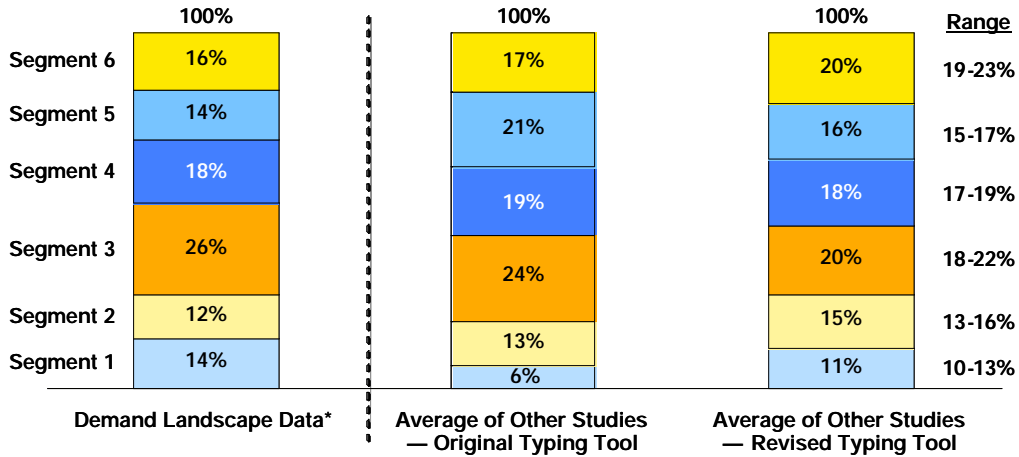
Table 2: CART-based typing tool performance summary

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6
Predictive Success	91%	89%	82%	82%	91%	87%
Purity Rate	80	88	92	85	81	87
Predicted in Segment	13	14	23	17	16	17
Actual in Segment	12	14	26	18	14	16

The improved performance in accuracy of the revised typing tool translated in higher replicability of the segment incidence in the other client datasets. This tool had significantly less variation in results across studies than did the discriminant analysis-based tool, and that built confidence in the CART-based tool. This CART-based typing tool also retained the Segment 1 membership at 10% or

above, which was of critical importance for the execution of the client's marketing strategy. See figure 3 for details.

Figure 3: Segment incidence of CART-based typing tool and comparison to discriminant analysis tool



The CART-based typing tool was clearly superior and was adopted as the on-going typing tool by the client.

5. Conclusion

We recommend using a CART-based typing tool when the tool's robustness is critical to successful strategy execution. If a segmentation is being used to manage a business, then that criterion is fulfilled. The extra time required to construct the tool, because it is not a completely automated process, is – in our experience – well worth the investment. The resulting tool is almost always significantly more accurate, more precise, and more robust.