Risk-Based Collection Model Development and Testing

Jane Martin and Rick Stephenson, Internal Revenue Service

The IRS Strategic Plan in part calls for "increasing compliance among small business and self-employed taxpayers." In 2000, an SB/SE (Small Business/Self-Employed) Design Team report focused on the need for the IRS to do the following: adopt an integrated compliance strategy and shift the emphasis toward risk-based compliance; include profiling major customer segments; and develop multifunctional treatment strategies in order to change compliance behavior patterns. A risk-based model prioritizes collection cases by risk of nonpayment.

A key component in this Collection Reengineering process was the formation of a Collection Strategy Team to identify potential improvements in the collection process and to suggest treatments. One of the objectives developed by the team was the use of predictive models to characterize aspects of the open SB/SE collection modules. The models would indicate a higher probability of a productive closure and conversely a low probability of a negative resolution. Predictive models are used by financial institutions, underwriters, and credit card companies to assess credit risk and collectibility of accounts. The SB/SE Research staff was asked to conduct a modeling effort to develop such a system.

Model filters were identified for those modules with a balance due that are likely to be unproductive "CNC" (Currently Not Collectible) or productive "FP" (Full Pay).[1] Accounts that have been routed through both filters, but cannot be identified as either CNC or FP criteria, are designated as "Other" accounts. Several benefits can be expected by routing cases to the most effective treatment: [2] Through early intervention, cases that would otherwise digress from potential Full Pay into a CNC can be treated and cured.

- Filtering CNC's out of the mix of cases routed to ACS (Automated Collection System) and Collection Field function (CFf) ensures that less time will be spent on unproductive cases.
- A greater volume of highly productive cases can be worked and a greater total volume of cases can be worked, due to less time spent per case.

- Working more productive modules will result in more dollars collected.

Research Methods

Initial Stage of Modeling

The initial modeling effort in early 2001 used limited data and uncovered numerous shortcomings that were later leveraged to improve the data and techniques for the second phase.

- Initial project included a small sample of 40,000 IMF taxpayers.
- Used data from accounts receivable linked with return transaction file.
- Models initially developed using SPSS Answer Tree and simple regression techniques. We later acquired SPSS Clementine Data Mining Software.

Second Stage of Modeling

The second data development stage began later in 2001 and included the following:

- More complete accounts receivable data extracted from IRS internal sources for four return types.
- Additional derived variables created from capturing relevant collection case history information.
- A merging of collection history data with tax return filing information to capture current business, filer profile, and income information.
- Development of some key ratios and measures based on relationships in the data.
- The team used SPSS Clementine data mining and machine learning techniques to identify patterns in historical collections data to reveal predictors of collections outcomes.

Data mining is an interactive and iterative process to identify useful relationships in large data sets. Some common techniques include the following:

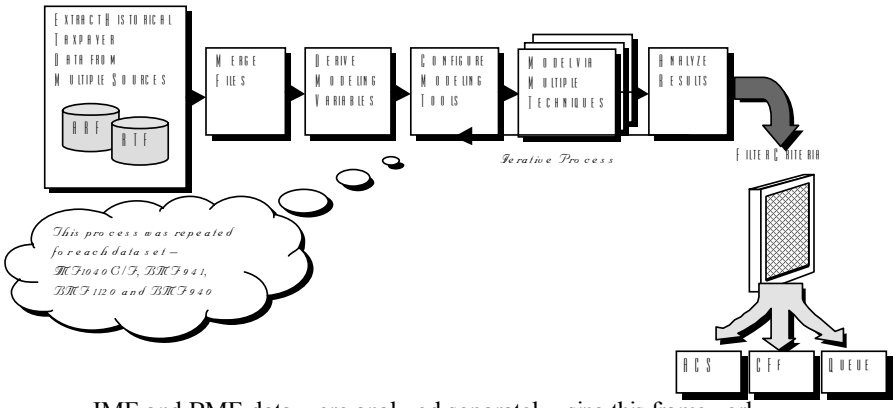
- Tree-Based Classification

- Neural Network Models
- Logistic Regression
- Cluster Analysis.

The term "machine learning" refers to the process of using historical data to generate models which can be applied to areas such as prediction, forecasting, estimation, and decision support.

- Machine learning models cull through the data set to identify and analyze patterns in the records.
- These models are generated inductively by generalizing from specific examples in the data set .

The following analysis framework was developed to identify filter criteria through predictive modeling…



…IMF and BMF data were analyzed separately using this framework.

Initial second-stage models were built using data from statutory notice forward. These models were built to predict how cases closed and where cases closed by entity. For example:

- Installment Agreement in ACS
- Adjustment in CFf
- CNC in CFf
- Full Paid in Notice.

Problems:

- A variety of closure types were too small for legitimate statistical modeling (OIC's (Offer in Compromise), FP in CFf, etc.)
- Overwhelming task to model and implement each treatment stream and outcome by entity.

The final "Meta" models developed in Clementine software employed the following three tools: C5.0 Rule Induction, C&RT Rule Induction, and Neural Networks.

C5.0 Rule Induction:

- C5.0 Rule Induction generates a classification model in the form of a decision tree—built by breaking the data into subsets more homogeneous than the original sample.
- The resulting classification model should be general so that it can be used to make predictions about data sets other than those used in its construction.
- Once the model is constructed, it can be used to make predictions on other data sets containing the same variables:
  - These predictions are made by running each case through the rule sets for "1.0" and "0.0" and assigning the prediction associated with the highest confidence level.
  - If a case cannot be classified, the model will assign a default value with confidence of 0.50.

C&RT Rule Induction:

- C&RT (Classification and Regression Tree) is similar to C5.0 in that it breaks the data into subsets that tend to be more homogeneous than the original sample relative to the target field.
- C&RT and C5.0 have several key differences:
  - C5.0 requires symbolic target fields where C&RT supports both symbolic and numeric targets (i.e., C&RT has capability to produce a classification or regression tree).
  - Classifications in C5.0 are made based on the information gained at each node (derived from information theory), while C&RT

classifies according to the degree to which cases in a segment are concentrated into a single target category.

Neural Networks:

- A Neural Network teaches itself to make predictions of the target outcome based on values of the independent variables.
- Neural networks simulate the way that the human brain works.
- The model constructs a network of nodes, or "neurons."
  - Connections between the nodes enable the network to identify patterns in the data and make predictions about a specified target variable.
  - Each node acts like a small processor focused on a simple task, collecting information from adjacent nodes and passing it along through the network.
- Once the network is set up, it trains itself to make predictions about the target variable, running through the data set one case at a time. As it culls through the data, it corrects itself to improve these predictions.
- In addition to the models generated which can make predictions, the neural networks provide a list of variables that contribute to the predictions, with a numerical value ranking the contribution of each variable.

One key advantage of using various model algorithms is that they are complementary. The complementary nature of these algorithms may be leveraged to improve the accuracy of predictions by combining results of different models into one aggregate prediction, or "metamodeling." Once preliminary models of each type were built, the team applied various "metamodeling" techniques to enhance results.

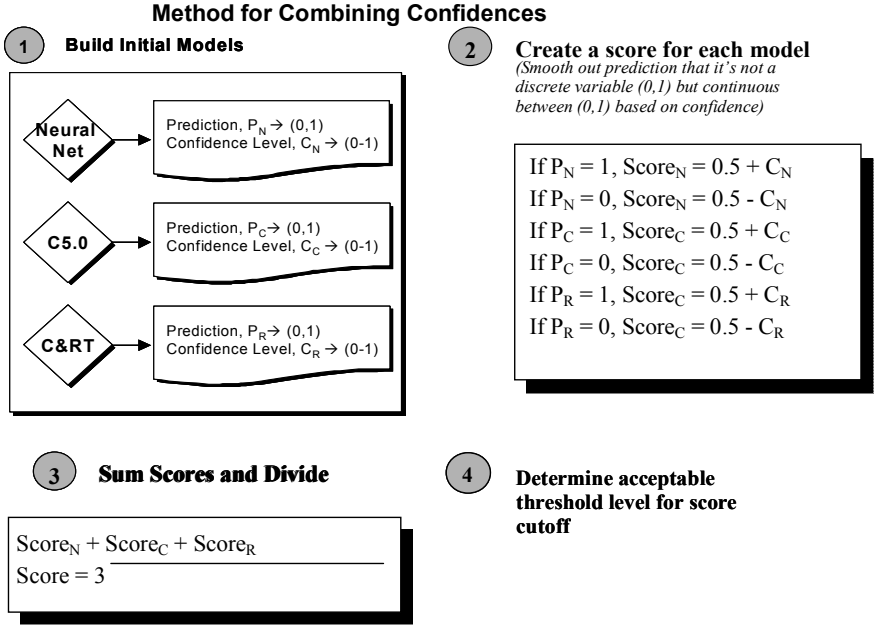
Some examples include the following:

- Use C&RT or Neural Networks to reduce data—build C&RT or Neural Network, then generate filter to select only the variables used in that model, then build a C5.0 model using only those variables.
- Data reduction using factor analyses or principle components analysis—identify variables that naturally group together to eliminate redundancies.

- Build multiple models and select prediction with highest confidence level.
- Voting—build two models and only use prediction if they both agree.
- Error Modeling
  - Generate model using one technique, then build a second model to predict which cases will be misclassified.
  - Select cases predicted to be misclassified and build a model for those using a different technique.
  - Also useful to identify variables that cause misclassifications.
- Combine confidences of two or more models that predict the same thing into an aggregate score.

The team conducted an iterative process of review and refining models to ensure consistent results.

After applying these various techniques, the team identified a method of combining confidences that yielded the best result.



Implementation and Testing

This collection strategy of using predictive models was implemented on January 1, 2003. One Full Pay (FP) model and one Currently Not Collectible (CNC) model were implemented for each of the following types of SB/SE tax returns:

- 1040 Individual Income Tax
- 1120 Corporate Income Tax
- 941 Employer's Employment Return
- 940 Employer's FUTA Return.

The primary objective of this testing phase of the project was to determine the accuracy of IDS (Inventory Delivery System) models that were implemented to predict the outcomes of cases as Currently Not Collectible and Full Pay. The final report will measure the accuracy of both CNC and FP filters for each form type and will use three measures of model accuracy as follows: (1) Measure 1 uses only closed modules as the common denominator. It is the number of modules that closed as predicted compared to the total number of closed modules for that form type and prediction. It is expressed as a percentage of closed modules. Example: 200 F1040 FP predicted modules, 100 are closed, 75 closed as FP. Thus, 75 divided by 100 = 75 percent. Also reported is the number of misclassified modules (predicted FP but closed CNC, and vice versa) compared to the total number of closed modules. Refer to Appendix A-1 for a collective summary of module closures for each form, year, and prediction. (2) Measure 2 is a more encompassing measure of the model's accuracy and compares the number of modules closed as predicted to the total number of modules for that prediction. It is expressed as a percentage of total number of modules for that prediction. Example: 75 F1040 modules predicted FP closed as FP, divided by 200 F1040 FP predicted modules = 38 percent. This is then compared against the same standard for the model as established under the model optimization and testing guidelines.[3] Appendix A-2 has the standards and how they were derived. As might be expected, 2004 results were less favorable than 2003 for all models because a smaller percentage of all modules had closed. This was due in large part to the shorter time frame that the 2004 modules had to close. There were also more modules selected in 2004 than 2003. For the overall model findings, results for our 2 years of data are averaged and compared to the predicted standard. (3) Measure 3 is the most comprehensive attempt to determine the overall accuracy of the model. This measure captures the misclassifications made by the models in addition to the accuracy rate. The three calculations of Measure 3 are as

follows: (a) The overall accuracy rate of each filter; (b) The percentage of accurately identified CNC and FP cases; and, (c) The percentage of misclassified cases (i.e., FP cases identified as CNC and CNC cases identified as FP).

The baseline standard measure for our comparison was the overall predicted accuracy rates[4] that were generated from the original models. We were not able to duplicate precisely the overall predictive accuracy formulas due to several data limitations, including the model's inability to sample modules that were filtered but did not receive a prediction (Other modules). Alternative formulas were developed to approximate as closely as possible the predictive formulas using the data available to us. These data limitations and compensations are discussed in Appendix A-3.

To provide the most comprehensive review possible on the available data, we calculated the Measure 3 in two different ways: first, using only the closed data from Scenario One, and, second, using the additional open data from Scenario Two. The "Best Case" (Scenario One) scenario considers the outcome for only closed cases, and the "Worst Case" (Scenario Two) scenario considers all open modules as well.[5] In addition, Scenario Two considers all open modules as incorrectly predicted.

For Measure 3, it is important to note that modules go through the CNC filter first. For those modules not receiving a CNC prediction, they move on to the FP filter. Those modules that move on to the FP filter are considered as receiving a Not CNC prediction for the CNC coincidence matrix purposes. Conversely, those modules that did receive a CNC prediction are considered to be predicted as Not FP for the FP filter matrix. Our coincidence matrix is a combination of both filter predictions.

Using the F1040 FP model as an example, the overall accuracy in Scenario One is the number of correctly predicted Not FP modules plus the number of correctly predicted FP modules divided by the total number of modules closed: (9+545)/729. The misclassification of FP is the number of actual FP modules predicted as Not FP divided by the total number of actual FP closures: 3/548. The percentage of accurately predicted FP is the number of correctly predicted FP modules divided by the total number of closed FP modules: 545/548. See the following table for Scenario One F1040 FP.

The overall accuracy for the F1040 FP model for Scenario Two is the number of correctly predicted Not FP modules plus the number of correctly predicted FP modules divided by the total number of modules (open and closed): (9+545)/1547. Open modules are considered as Not FP. The misclassification of FP is the number of actual FP modules predicted as Not FP divided by the total number of actual FP closures: 3/548. The percentage of accurately predicted FP is the number of FP modules correctly predicted by the model divided by the total number of closed FP modules: 545/548.

This "Worst Case" measure uses both open and closed modules. As expected, the open unresolved cases provide a much more conservative measure of the overall accuracy.

Scenario Two: 2003 Model Performance for 1040 FP

		Predicted Outcomes		
		Not FP	**FP**	**Total**
Actual Outcomes	**Not FP**	9	990	999
	FP	3	545	548
	Total	12	1535	1547

Measure 3 evaluates the model in terms of "successful," "undetermined," or "unsuccessful" for each year. Our definition of "successful," for each scenario, was met when our confidence intervals overlapped with those of BAH (Booz Allen Hamilton) or were superior to those of BAH for the variable "Overall Accuracy Rate." Our definition of "not successful" was applied when our confidence intervals were inferior to those of BAH for the variable Overall Accuracy Rate.

In a few instances, the best case scenario is successful, and the worst case is not successful, in which the result is considered undetermined. In the instances where the result is "undetermined," Measures 1 and 2 were given more weight in making a decision on the overall model performance. All three measures have results for both 2003 and 2004. An overall accuracy determination including both years is shown in the overall findings for each form and model. Those few situations where the results cannot be determined will be identified.

Direct interpretation across years is difficult as those modules filtered in 2003 have had a minimum of 12 months and a maximum of 24 months to close, while those filtered in 2004 have had a minimum of 1 week and a maximum of 12 months to close after filtering. Consequently, a higher percentage of modules filtered in 2003 have closed simply because they have had more time to do so. Certain trends observed between the 2 years will be noted.

Sample Design

The population for this project was SB/SE balance due modules that passed through the IDS CNC and FP filters. A sample design was previously developed and implemented to identify a sample of the modules passing through and selected by the CNC and FP filters. The modules to be tested were sampled between January 1, 2003, and December 12, 2004, and designated as monitored cases. The population was segmented into four market segments, based on tax return type—1040, 1120, 941, and 940.

Monitored cases for 2003 were projected based on FY 2000 closures and subsequently revised for 2004 based on 2003 actual incoming inventory of modules that qualified for modeling.

Sample sizes for the 2 years were quite disparate. The confidence levels of the sample sizes were computed at 95-percent confidence for both years. The precision, or error percentage, of the sampling was poor for the CNC model in 2003 for all four form types. The following error rates resulted in 2003:

- F1040 CNC had an error rate of 17 percent.
- The F941 CNC had an error rate of 20 percent.
- The F1120 CNC had an error rate of 34 percent.
- The F940 CNC had an error rate of 60 percent.

Conversely, the FP model precision ranged from 3 percent to 7 percent. Due to the poor precisions for the 2003 CNC samples those results should be interpreted cautiously.

For both FP and CNC models in 2004, the sampling precision or error ranged from 3 percent for F1040 FP to 11 percent for F940 CNC. Consequently, the results for all the models in 2004 can be interpreted and used with a high level of confidence. See Appendix B-1 for actual sampling numbers. The sampling design attempted to achieve a 95-percent confidence level for dichotomous variables. The estimated precision varies by form type. Data extracts were performed at 6-month intervals beginning in June 2003 and ending in January 2005. The analysis represents all of the modules that flowed through the FP and CNC filters during 2003 and 2004.

Analysis Issues

The assessment of results is complicated by factors related to time. The models were designed using cases that were allowed up to a 4-year resolution period, which is much longer than the average cycle time. Average cycle time for resolution of cases in the field for 2002 was approximately 40 weeks.

Therefore, the comparison of actual case outcomes to the previously specified performance measures should be considered tentative.

Results

An overall assessment indicates that the FP models for all form types are performing well in making accurate outcome predictions. All meet or exceed our baseline overall predictive accuracy rates except the F1040 model which has a neutral outcome. The F940 has the highest accuracy rates, followed by the F941 and F1120 models for Measure 3. These three form types also perform very well for Measures 1 and 2. The F1040 FP model overall accuracy for Measure 3 is neutral, but the model performs well on Measures 1 and 2 and is therefore considered successful as well.

The CNC models have mixed results in accurately predicting outcomes but overall are less successful at this time than the FP models. The F1040 CNC and F941 CNC are performing the best of the four CNC models. Some of this can be attributed to the smaller numbers of sampled modules, especially in 2003. In 2004, the modules counts are higher, but the majority of these were selected later in the year resulting in fewer closures because they have had less time to be worked and closed. CNC modules also generally take longer to close than FP modules. Given additional time, the 2004 modules closures may improve the overall accuracy of the models. The F1040 and F941 CNC modules are the subject of additional tracking reporting that will look at them over an additional year.

Did It Work?

Results of Collection Strategy and Model Implementation

- Yield from categories other than first notice has increased by nearly $1.8 billion or 8.4 percent over FY 03.
- The single largest component is Taxpayer Delinquent Account (TDA), and TDA yield increased by over 8 percent, from $9.6 billion in FY 03 to $10.4 billion in FY 04.
- These results reflect increasing effectiveness in collecting tax revenue.

% Improvement			
	FY 04	**FY 03**	**over FY 03**
Average Hours per ACS TDA Closure	3.21	3.51	8.55%
Average Hours per ACS TDI Closure	1.84	3.07	40.07%
Average Hours per CFf TDA Closure	32.53	34.07	4.52%
Average Hours per CFf TDI Closure	56.54	93.58	39.58%

Endnotes

[1] CNC is defined as those accounts that have been removed from active inventory for a variety of reasons, including undue hardship, inability to locate the taxpayer or assets, etc. For this project, cases are classified as FP when 95 percent of the initial module balance has been paid.

[2] Booz Allen Hamilton SB/SE "Collection Strategy Findings" (1/31/02).

[3] Derived from the Coincidence Matrix for each model in Booz Allen Hamilton, User Guide, SB/SE Collection Strategy Filter Maintenance and Testing, Section III, 1/31/2003.

[4] Booz Allen Hamilton, User Guide, SB/SE Collection Strategy Filter Maintenance and Testing, Section III, 1/31/2003.

[5] Detailed results are available from the authors at: jane.e.martin@irs.gov and rick.w.stephenson@irs.gov.

Appendices

Appendix A-1

Measure 1

2003 and 2004
Resolved/Closed Module Summary

FP Predictions

Form	Total Closed	Resolved: Closed as Predicted (FP)	Percent of Closed	Misclassified (CNC)	Percent of Closed	Neutral (Tolerance)	Percent of Closed
1040							
2003	763	545	71.4%	172	22.5%	46	6.0%
2004	376	242	64.4%	42	11.2%	92	24.5%
941							
2003	290	264	91.0%	16	5.5%	10	3.4%
2004	328	299	91.2%	5	1.5%	24	7.3%
1120							
2003	722	595	82.4%	104	14.4%	23	3.2%
2004	469	398	84.9%	36	7.7%	35	7.5%
940							
2003	159	152	95.6%	5	3.1%	2	1.3%
2004	239	234	97.9%	1	0.4%	4	1.7%

CNC Predictions

Form	Total Closed	Closed as Predicted (CNC)	Percent of Closed	Misclassified (FP)	Percent of Closed	Neutral (Tolerance)	Percent of Closed
1040							
2003	12	9	75.0%	3	25.0%	0	0.0%
2004	78	54	69.2%	15	19.2%	9	11.5%
941							
2003	13	10	76.9%	3	23.1%	0	0.0%
2004	60	33	55.0%	20	33.3%	7	11.7%
1120							
2003	4	2	50.0%	2	50.0%	0	0.0%
2004	18	3	16.7%	12	66.7%	3	16.7%
940							
2003	1	0	0.0%	1	100.0%	0	0.0%
2004	16	4	25.0%	4	25.0%	8	50.0%

Appendix A-2

Measure 2

Our Partial Accuracy of Module Predictions compared to those of BAH's

	2003	2004	Average of Both Years	BAH	Rate of change	Accuracy Rating
1040 FP	35%	20%	28%	44%	-36%	Fair[1]
941 FP	72%	44%	58%	70%	-17%	Good[2]
1120 FP	68%	46%	***57%***	51%	12%	Good
940 FP	92%	91%	***92%***	41%	123%	Good
1040 CNC	28%	13%	***21%***	21%	-2%	Good
941 CNC	40%	13%	27%	52%	-49%	Not Accurate[3]
1120 CNC	25%	2%	14%	42%	-68%	Not Accurate
940 CNC	0% *	7%	7%	49%	-86%	Insufficient Data[4]

* There were no closures at time of data extraction.

[1] If the Rate of Change, R, between our result and BAH partial accuracy is -40% = R = -20%, we consider the model prediction as fair.

[2] If the Rate of Change between our result and BAH partial accuracy is R > -20%, we consider the model prediction as good.

[3] If the Rate of Change between our result and BAH partial accuracy is R < -40%, we consider the model prediction as not accurate.

[4] We do not have enough cases in our sample to draw any conclusions.

Appendix A-3

Measure 3: Accuracy rates

In the tracking project, we attempted to use as our baseline the overall accuracy rates that were predicted by BAH in their filter maintenance and testing documents. We were unable to duplicate their methods of analysis for several reasons previously mentioned due to our data limitations. Our method of measurements consisted of considering the CNC and FP module predictions only. We considered only those modules that were modeled and sampled between January 3, 2004, and December 31, 2004. We had originally planned to analyze those that had been assigned to ACS and the field collection a

minimum of 1 month. This was not practical due to the data constraints. In our test, we analyzed modules if they were filtered anytime within our data extract cycle 200301 to 200451. Therefore, we had some modules that were filtered up to 24 months from the last extract cycle and some up to 1 week from the last extract cycle.

One aspect of the characteristics of the modules is that FP modules historically close faster than CNC closures. Since the test was looking, first, at closed modules only, we had more FP module closures to analyze than CNC module closures.

Another barrier that prevented us from fully complying with the BAH methodology, and subsequently with the plan, was that the model was not designed to monitor the Other modules. The results of these closures were an integral part of the performance evaluation of each model. Indeed, for BAH methodology, modules predicted to be CNC that were still open or that closed in a way other than CNC were considered as "Not CNC," and, conversely everything predicted to be FP that was open or closed other than FP was considered as "Not FP." The BAH study was cross-sectional in time.

Our prominent comparison with BAH methodology was based on the Overall Accuracy Rate. This rate is defined by the diagonal and the Total cells (see chart below). The Type I error means that we predicted NOT CNC, and it was an actual CNC. Type II error means that we predicted CNC, and it was an actual NOT CNC. The misclassification plays a major role in the Overall Accuracy Rate. The higher the misclassification, then the lower the overall accuracy rate, and vice-versa. An example of this is in the charts below: Based on a Reject-Support testing (RS testing) in which the null hypothesis reject favors the model claim.

H_0: Not CNC H_a: CNC

		Predicted Outcomes	
		H_0	**H_a**
Actual Outcomes	**H_0**	**Correct Acceptance**	**Type II Error**
	H_a	**Type I Error**	**Correct Rejection**

		Predicted Outcomes		
		Not CNC	**CNC**	
Actual Outcomes	**Not CNC**	Correct	**Misclassified**	Marginal Total
	CNC	**Misclassified**	Correct	Marginal Total
		Marginal Total	Marginal Total	Total

This test was based on closed monitored modules (open monitored modules were disregarded in Scenario One (see below), and on closed and open monitored modules (Scenario Two (see below)). Therefore, it was longitudinal in time.

In our project, we used two scenarios: The Scenario One considered closed modules only and was to assume that, because of the logistics, everything that was not CNC was automatically considered as FP and vice-versa. This was the major compromise that we had to make in order to be able to compare our accuracy to a benchmark. This scenario was the most optimistic in regard to the Overall Accuracy Rate, and it inflated the rate. See chart below for Scenario One FP.

2003 Scenario One: Model Performance for 1040 FP

		Predicted Outcomes		
		Not FP	**FP**	**Total**
Actual Outcomes	**Not FP**	9	172	181
	FP	3	545	548
	Total	12	717	729

The Scenario Two was to consider, additionally, the open modules CNC's/FP's as Not CNC's/Not FP's to match the BAH methodology. This alternative generated other uncertainties due to the unknown actual closures of the open modules. This was the second major compromise that we had to make in order to be able to compare our accuracy to a benchmark. This scenario was the most pessimistic in regard to the Overall Accuracy Rate, and it reduced the rate. See chart for Scenario Two FP.

2003 Scenario Two: Model Performance for 1040 FP

		Predicted Outcomes		
		Not FP	**FP**	**Total**
Actual Outcomes	**Not FP**	9	990	999
	FP	3	545	548
	Total	12	1535	1547

The most appropriate benchmark was, of course, the BAH accuracy rate, but we had to compare one longitudinal study to its equivalent cross-sectional one due to data limitations and logistics constraints. We compared our accuracy rates (95-percent confidence level and various precisions intervals) with those

of BAH (95-percent confidence level and 95-percent precision intervals). We used the confidence intervals (95-percent confidence level) instead of z-tests that were planned before for reason of practicality. The variables used were: 1. Overall Accuracy Rate. 2. Percentage of Actual CNC's/FP's Correctly Identified. 3. Percentage of FP's/CNC's cases Identified as CNC's/FP's.

Our definition of "successful," for each Scenario, was when our confidence intervals overlapped with those of BAH or were superior[1] to those of BAH for the two variables: Overall Accuracy Rate and Percentage of Actual CNC's/FP's Correctly Identified. "Successful" was when our confidence intervals overlapped with those of BAH or were inferior[2] of those of BAH for the variable: Percentage of FP's/CNC's cases Identified as CNC's/FP's.

Our definition of "not successful," for each Scenario, was when our confidence intervals were inferior to those of BAH for the two variables: Overall Accuracy Rate and Percentage of Actual CNC's/FP's Correctly Identified. "Not successful" was when our confidence intervals were superior to those of BAH for the variable: Percentage of FP's/CNC's cases Identified as CNC's/FP's.

Our Overall Accuracy Rate for Scenario One was the same for both CNC's and FP's for each form and was inflated due to the major compromise discussed above. We considered this Scenario as the upper limit of the Overall Accuracy Rate range.

Our Overall Accuracy Rate for Scenario Two was different for CNC's and FP's for each form. We considered this Scenario as the lower limit of the Overall Accuracy Rate range.

The Percentage of Actual CNC's/FP's Correctly Identified and Percentage of FP's/CNC's cases Identified as CNC's/FP's remained unchanged for both Scenarios.

The Overall Accuracy Rate measurement for comparison was defined as follows.

If the Overall Accuracy Rate in Scenario Two (lower limit), for a particular form and module type, was higher or equal to BAH's, then the model is successful for that particular form and module type. Indeed, if the worst case is successful, then each case is successful.

If the Overall Accuracy Rate in Scenario One (higher limit), for a particular form and module type, was lower, then the model is not successful for that particular form and module type. Indeed, if the best case is not successful, then no case is successful.

In the only critical case: best case successful and worst case not successful, the result is undetermined**.**

How we measure any accuracy in cases of undetermined result

In these cases, we use the best comparable measurements that we have considering the data limitations. These measurements are the Partial Accuracy Rates defined in Measure 1 and Measure 2.

When all three measures are used together:

Measure 1 and Measure 2 were also used to determine the strength of success after we used Measure 3. Final results based on the three measurements of accuracy below placed in order of importance.

	Measure 3	Measure 2	Measure 1	Result by year	Final Result
2003 1040 FP	undetermined	Fair	71%	successful	successful
2004 1040 FP	undetermined	Fair	64%	successful	
2003 941 FP	successful	Good	91%	successful	successful
2004 941 FP	successful	Good	91%	successful	
2003 1120 FP	successful	Good	83%	successful	successful
2004 1120 FP	undetermined	Good	85%	successful	
2003 940 FP	successful	Good	96%	successful	successful
2004 940 FP	successful	Good	98%	successful	
2003 1040 CNC	successful	Good	75%	successful	successful
2004 1040 CNC	undetermined	Good	69%	successful	
2003 941 CNC	successful	Not accurate	77%	successful	undetermined
2004 941 CNC	undetermined	Not accurate	55%	undetermined	
2003 1120 CNC	successful	Not accurate	50%	successful	undetermined
2004 1120 CNC	undetermined	Not accurate	17%	Not successful	
2003 940 FP CNC	Not enough data	Not enough	0%	undetermined	undetermined
2004 940 CNC	successful	Not enough	25%	successful	

Footnotes

[1] Interval A is superior to interval B means, here, that each element of A is superior to each element of B.

[2] Interval A is inferior to interval B means, here, that each element of A is inferior to each element of B.

Appendix B-Sampling Information

Table B-1:

2003 Actual Sample Sizes, Confidence Levels and Precisions

Form	Period of	Population	Actual Sample Size	Confidence Level	Precision (±)
1040 FP	01/2003-12/2003	542,688	1,585	95%	3%
941 FP	01/2003-12/2003	8,507	367	95%	5%
1120 FP	01/2003-12/2003	10,569	871	95%	3%
940 FP	01/2003-12/2003	633	166	95%	7%
1040 CNC	01/2003-12/2003	24,056	32	95%	17%
941 CNC	01/2003-12/2003	3,394	25	95%	20%
1120 CNC	01/2003-12/2003	102	8	95%	34%
940 CNC	01/2003-12/2003	115	3	95%	60%

Source: IDS SAP Processing Report Jan. 1 - Dec. 31, 2003

Table B-2:

2004 Actual Sample Sizes, Confidence Levels and Precisions

Form	Sampling Period	Population	Actual Sample Size	Confidence Level	Precision (±)
1040 FP	1/1/04 - 8/22/04	191,529	691		
	8/23/04 - 12/31/04	132,712	493		
	2004 Total	**324,241**	**1,184**	95%	3%
941 FP	1/1/04 - 8/22/04	15,219	213		
	8/23/04 - 12/31/04	11,005	463		
	2004 Total	**26,224**	**676**	95%	4%
1120 FP	1/1/04 - 8/22/04	8,045	562		
	8/23/04 - 12/31/04	3,852	309		
	2004 Total	**11,897**	**871**	95%	3%
940 FP	1/1/04 - 8/22/04	631	120		
	8/23/04 - 12/31/04	181	138		
	2004 Total	**812**	**258**	95%	5%
1040 CNC	1/1/04 - 8/22/04	12,667	14		
	8/23/04 - 12/31/04	9,375	406		
	2004 Total	**22,042**	**420**	95%	5%
941 CNC	1/1/04 - 8/22/04	10,052	39		
	8/23/04 - 12/31/04	6,096	549		
	2004 Total	**16,148**	**588**	95%	4%
1120 CNC	1/1/04 - 8/22/04	253	11		
	8/23/04 - 12/31/04	177	160		
	2004 Total	**430**	**171**	95%	6%
940 CNC	1/1/04 - 8/22/04	164	*		
	8/23/04 - 12/31/04	59	*		
	2004 Total	**223**	**55**	95%	11%

* Not disclosed to protect taxpayer confidentiality.

Source: IDS SAP Processing Report Jan. 1 - Dec. 31, 2004