Evaluation enrichment through data warehouse mining

Stuart Turner

Stuart Turner
stuart.turner@ird.govt.nz
Inland Revenue Department, New Zealand

Paper presented at the Australasian Evaluation Society 2004 International Conference 13-15 October-Adelaide, South Australia www.aes.asn.au

Abstract

Evaluators conduct evaluations under time, budget and often data constraints. Mining of a data warehouse has the potential to enrich and add real value to an evaluation, both quantitative and qualitative, adding strength and depth to the evaluation and widening design opportunities.

This paper discusses the value of data mining in supporting and enriching an evaluation. It looks at the opportunities for profiling key population characteristics and the potential for highlighting areas of analytical interest.

Sample design and case study selection can be enhanced by the perspective offered via a data warehouse. There are possibilities for the review of respondents' contact details and much can be gained from extracting and providing background information for fieldworkers on selected cases.

This paper includes practical examples drawing from experiences in the *Compliance Costs to Small Businesses* and *The Impact of Tax Audit on Taxpayer Compliance* case studies. The risks involved will be discussed along with ethical implications of using data mining in an evaluation context.

Introduction

Inland Revenue holds a great volume of data within its data warehouse; through the efficient use of this wealth of data Inland Revenue's Evaluation Services have enhanced our evaluations. By way of a discussion of various projects and practical examples, I will highlight how utilising a data warehouse has enriched and added great value to our evaluations. "To be power laden, information must be relevant and in a form that is understandable to the users" (Patton, 1997).

What is a data warehouse?

A database, frequently very large, that stores an organisation's information. In Inland Revenue's case it is a huge electronic repository storing detailed information about all our taxpayers. Detailed information such as tax return particulars, filing dates,

addresses and phone numbers, payments made and refunds received are all captured and stored together with historical information in the data warehouse.

A data warehouse makes it much easier and more efficient to run queries, retrieve and analyse data that originally came from different sources.

Data Mining

The volume and nature of Inland Revenue's tax administration data lends itself well to "data mining". This paper uses the term "data mining" to refer to the basic digging, drilling and summarising of data to discover profiles, patterns and trends. Standard summarising, analytical and statistical methods are used. Any interface or software that allows summarising and querying of data can be used, in this context, to "data mine". Excel or Access, for example, can be used as well as more specific software like SAS's Enterprise Miner, or Teradata's warehouse miner, incorporating modern pattern recognition technologies and processes such as neural networks and regression trees.

Background - Inland Revenue and the "Data Warehouse"

Inland Revenue

The New Zealand government gets over 80% of its revenue through tax and it is Inland Revenue's job to collect it, approximately 35 billion dollars a year. Inland Revenue also administers child support, family assistance, paid parental leave, administers gaming duties and supports the NZ student loan scheme.

Some 10,000 pages of legislation direct Inland Revenue's activities notably the Tax Administration Act 1994, Income Tax Act 1994, Goods and Services Tax Act 1985 and Child Support Act 1991.

Inland Revenue interacts with large numbers of citizens through 22 delivery sites, employing over 4,600 staff. The customer base is over 5.5 million taxpaying entities.

This results in large volumes of interactions:

- 4.1 million person-to-person telephone contacts
- Processing 10 million tax returns
- 47 million financial transactions per year.

Inland Revenue's Data Warehouse

The size of the Data Warehouse reflects the size and complexity of Inland Revenue's business. It holds full information from 1997 onwards.

Currently the database is 1.4 terabytes. One terabyte, says Gregory Piatetsky-Shapiro, editor of KDnuggets.com (www.math.ccsu.edu/larose/courant), is roughly equivalent to the data in a million Yellow Pages phone directories. The Data Warehouse is supported by six technical support staff and a database administrator. There are 160 ad hoc query users from around the department and 440 web based application users

(this is set to grow to 900 by March 2005) all looking at over 300 Oracle tables (detail and summary), which draw the data from 96 operational and 16 non-core system datasets.

Inland Revenue's Data Warehouse grew from the:

- strategic requirement to better understand customer behaviour in order to improve compliance
- need to analyse and report on data held in information systems
- need for flexibility and responsiveness to information analysis and reporting requests.

The warehouse contains the following applications, for the use of managers, analysts and operational staff:

- ad hoc query and reporting
- reporting applications
- analytical applications.

Some of the overall benefits of the Data Warehouse for Inland Revenue are:

- targeting and analysis of customer characteristics
- improved tax form design
- better case selection for enforcement activity
- input to project, policy and legislation design
- more efficient use of resources and flexible reports
- reduced need to interact with core business systems hence reduced mainframe load
- longitudinal tracking of compliance behaviour
- ability to rapidly and flexibly generate reports.

Enrichment and benefits of using a data warehouse for evaluation

In many cases the ability to base evaluation decisions on facts supported by a depth of actual data improves the validity and robustness of those decisions. Some of the many opportunities for adding value and enriching an evaluation through data warehouse mining are summarised below, using practical examples from our recent projects.

Profiling population characteristics

Data mining in order to detail specific population characteristics (for example age, income, time use distributions and patterns) helps define information requirements and proves invaluable in the early stages of an evaluation. For example our profiling of specific businesses alerted us to the number of organisations and enterprises with tax exemptions. The discovery of this group and ability to analyse its characteristics factually supported further work in defining a target population.

Highlight areas of interest (and their relative size)

The ability to highlight and investigate specific groups of a target population allows the evaluator to focus attention on the make up and characteristic of a group. What is different or special about this group? Do we need to exclude them from our target population? How can we achieve sufficient representation? Why are they behaving

the way they do? In the case of our compliance cost research, a surprisingly large and quite unexpected group - taxpayers registered for GST but having a zero turnover - made us question the value of including this specific group in our target population. Having highlighted this group, we were able to dissect the group by business type, discovering similarities and trends within. With this extra knowledge we were better placed to decide whether to exclude the whole group or parts of it.

Sample design
Knowing the population size of specific groups allows a better understanding of sample implications and helps with sample stratification, strata delineation, determining sample size and weighting results.

Assist questionnaire design (what information do we already have?)
The number of questions asked in a questionnaire may be reduced by replacing a survey question with information provided from a data warehouse. The reduction in questionnaire length may translate directly into a corresponding improvement in response rate. A shorter questionnaire can do no harm to response rates, the shorter length reduces respondent burden. "Common sense suggests that the shorter the questionnaire, the more likely a high response rate, …" discusses Karen Bogen (1996). Michael Bamberger et al (2004) make similar comment, "considerable saving in cost and time can often be achieved through reducing the length and complexity of the survey instrument. A ruthless pruning of survey instruments to eliminate non-essential information can often significantly reduce the length of the survey".

We were able to do this "ruthless pruning" as a part of the compliance cost research. By removing questions that can be provided for from Inland Revenue's Data Warehouse we were able to shorten the questionnaire considerably. Examples of information that was provided by the Data Warehouse and which is then linked to survey data are: Taxes paid, how often taxpayers file a return, the way in which they file a return – electronically or traditionally And weather they use a tax agent.

Review survey respondent's contact details
The worth of being able to check the correctness of contact details is quickly apparent. Where addresses or phone numbers are incorrect or appear "odd" posted information or phone contacts may not find their target. By appropriate digging in a data warehouse, alternatives and recent changes can be looked into, all improving the chances that information gets to whom it is intended.

Assist relevant pilot / case selection
The selection of participants to take part in a pilot or case studies is based on the need to gather reliable data to help the researcher understand any issues influencing their evaluation questions. The ability to select pilot members to fit a certain criteria is improved by having the population characteristic at your fingertips via a data warehouse. The fine tuning of selected cases reveals precisely the information needed. The provision of adequate and appropriate information not only improves the value of the pilot, but a successful pilot adds to the overall validity of the evaluation.

Provide fieldworkers with selected case background information
Providing qualitative researchers with relevant information about the participant can help their better understanding of the topic, setting, or participant's characteristics. A

reliable source of background information can enrich the fieldworker's interpretation of qualitative research. For example, does the business employ a seasonally fluctuating workforce? For example – fruit pickers. Or is it a new business – "having recently set up shop". Is it too new to have a full compliance cost experience post business set up?

Case Study 1 – Measuring Compliance Costs of Small and Medium Sized Business

Research Objectives

Inland Revenue commissioned a project to measure the tax compliance costs of small and medium sized businesses. Its purpose was to better inform the department and the Government of the compliance cost impacts of current and future tax policy changes for the purposes of developing and evaluating tax policy proposals.

The starting point, and core of this research assignment, was a major benchmark survey quantifying compliance costs for small and medium businesses, in dollar terms and time spent.

- the survey covered all major tax types and a range of business sizes: micro (0 employees), small (1-5 employees) and medium (6-19 employees)
- a pilot survey (350) was conducted to help with the question refinement
- 5600 letters were sent to businesses offering them the opportunity to withdraw from the research and identifying businesses no longer trading
- full questionnaire was mailed out to approximately 4000 businesses.

Population Definition

The use of data warehouse type information enabled us to profile the overall taxpayer population and make more informed decisions on whether particular groups should be included or excluded in the target population. The research focus was small and medium businesses so we needed to know about the size and the nature of the businesses.

Table 1 summarises some of the challenges and decisions we made in defining our final target population. Initial warehouse mining gave rise to many additional questions and made us think really hard about, for example, the definition of businesses size – is 'number of employees', 'turnover', or 'PAYE (Pay As You Earn) paid' a better representation of business size? The power of the Data Warehouse was demonstrated, identifying where the boundaries overlapped and where not, suggesting categories to eliminate, and introducing greater flexibility into our analysis. We were able to capture size based both on employee numbers and on annual turnover, meaning future analyses could be based on the one most appropriate to the present policy issue.

Table 1: Summary of key decisions in defining our target populations

Decision
How small is "small", how big are "medium" businesses?
Non-GST (Goods and Services Tax) registered but still a business and trading, e.g. turnover <$40,000: can we identify them to include in sampling?

How do we ensure that we get enough companies with employees to reliably assess PAYE (Pay As You Earn) and FBT (Fringe Benefit Tax) compliance costs?
How do we deal with overseas taxpayers? and non-profit organisations?
How do we determine the number of employees a business employs?

Equally important, the data mining analysis quickly posed questions about how we should define and handle "non-active" businesses and non-profit organisations.

Table 2 shows the stepped improvements in population definition starting with all current businesses and entities.

Table 2: Improvements to target population "small and medium businesses"

Step	**Population**
Starting whole population (all taxpaying entities)	7,536,000
Selecting "active taxpayers"	5,824,000
Selecting appropriate businesses "legal form": *Excluding clubs and societies, unit trusts, government. Including self employed individuals*	1,355,000
Excluding non-NZ residents & non-valid or overseas contact details	1,245,000
Excluding not-for-profit businesses	1,210,000
Excluding very small and "non-active"	366,000
Final main target population	**366,000**

Sample Design

The target of the research was small and medium-sized businesses, including self-employed taxpayers. The use of the Data Warehouse and the depth of information stored allowed development of a complex sample design with 32 strata which ensured the research objectives were met while minimising respondent burden and field resources.

To meet the objectives of the research "*to better inform the department and Government of the compliance cost impacts of current and future tax policy changes*" various types of tax are of interest. Careful sample design was required in order to relate to and discriminate between all major taxes. The sample design and selection had to cover:

- businesses with and without employees; and business size as relating to numbers of employees and annual turnover
- the key tax types - income tax, PAYE (Pay As Your Earn), GST (Goods and Services Tax) and FBT (Fringe Benefit Tax)
- businesses with and without tax agents
- in particular businesses which were both registered for GST and income tax as a great number of current policy initiates are focused on simplification of tax for this group.

To accommodate these requirements in the final design a disproportionate stratified random sample was used. There are 32 groups/strata reflecting different levels of turnover, employee numbers, and whether FBT is paid.

The major advantage of this survey design was that the sample data could be used to generalise to the entire population. A proportional approach would have left some important groups so small as to make drawing statistically valid findings from the survey results impossible. The large sample size and disproportionate stratified random sample allowed results for small groups to be reported on. This is particularly important for policy initiative requirements. A relevant example is FBT, where proportional sampling would result in only around 78 FBT respondents. Results from so few respondents would not be sufficiently reliable to report separately. The stratified design in **Table 3** provided a much healthier 207.

Table 3: Sample Design - Respondents for FBT tax type

FBT		Size: Turnover				Total
		No turnover	Micro <100k	Small 100k -1.3m	Medium 1.3m+	
Size: Employee numbers	No Employees	-	13	14	2	29
	Small 1-5	1	8	41	10	60
	Medium 6-19	1	1	29	43	74
	Large 20+	0	0	4	40	45
Total		**2**	**22**	**88**	**95**	**207**

The planning and actual sample selection was made possible by the Data Warehouse. As well as defining and sizing each stratum, the way the data was stored allowed each of the many strata to be grouped efficiently and then easily sampled. Without an effective data warehouse this design and selection process would have proved very time consuming and difficult, if not impossible.

Improving response rate

A risk of mail questionnaires is a poor response rate, a response rate that is so low that reliable analysis of the returned data is inappropriate. "*A survey program is only as effective as its weakest link. In general, this tends to be the low percentage of returns frequently found in mail surveys. The response number has a direct effect on the conclusions you can draw from the data"* (Pearson NCS, 1997). The questionnaire designed for the compliance cost research is no different and the results and overall success of the project depends on the overall response rate.

By defining our target population, and excluding where possible businesses and enterprises which fell outside the target or with uncertain eligibility or dubious contact information, we focus our attention and scarce resources on those enterprises for which the research is aimed. The focused definition of potential questionnaire recipients increases the chances of an acceptable response rate and hence the value of the results we do collect.

Without the ability to profile our current population we could contaminate our target population with "businesses" which are outside the scope of the research and hence potentially waste resources by sending them a questionnaire and then either the process in following up on their non-response or having to disregarded their response as irrelevant.

Case Study 2 – Measuring the impact of audit on taxpayer compliance

Research Objectives

Creating an environment that enhances compliance is a strategic direction for Inland Revenue. Audit is the keystone of enforcement efforts, to foster voluntary compliance in the willing and to detect and deter the unwilling.

The evaluation "Measuring the impact of audit on taxpayer compliance" aimed to answer the fundamental question "What impact does audit have on the compliance of those audited and on the wider community?" The evaluation was focused on the outcome of audit in compliance terms rather than on the audit process itself.

Our premise is that audited businesses and individual taxpayers, will be more compliant following an audit. As a result, their assessed and paid tax will rise, they will file more promptly than before, and as a result of the audit they may change or tighten their business practices with flow-on benefits for tax compliance.

The evaluation was made up of a number of individual projects and provides information on behaviour and changes, awareness and attitudes. The focus of the evaluation was:

- the change in behaviour of those audited contrasted to those not audited, using filing and payment behaviour records and analysis of income and revenue declaration
- the change in compliance behaviour for a sample of recently audited taxpayers through a post audit check task
- the most significant changes resulting from audits for different types of audited taxpayer (qualitative).

Measuring Behavioural Change – evaluation design and analysis

One project measures the impact of an audit on a taxpayer's compliance behaviour. The three aspects of compliance studied are filing, paying and income declaration. This project is an example of using warehouse data in a complex before-after, intervention group-matched control group study. The analysis compares the compliance of the audited taxpayers pre- and post-audit which, in turn, is compared with the matched, non-audited taxpayers. The data warehouse makes this possible. In particular:

- reconstructing past behaviour, retrospectively
- monitoring future behaviour at different points in time
- matching the intervention and control group on up to six criteria
- using the total qualifying population for the intervention group rather than a sample
- efficient extraction of data.

Pitfalls of data warehouse mining

No discussion of a data warehouse would be complete without highlighting some of the pitfalls inherit in using data warehouse information.

Risks

Data warehouses are often built from administrative databases where the system was originally designed and created for the recording and management of data for operational use. As such they are not always well set up for research. Some of the following issues may be present in the data, making it not immediately useable and requiring cleansing or careful modification:

- information may not always be complete
- information may not be entered in a way suited to the evaluators' purpose
- some critical fields may be null
- uncertainty whether null fields represent 'zero' or 'not applicable'
- important text fields might range considerably in length and be difficult to analyse automatically
- information is not correctly updated
- field descriptions can be misleading
- not all operational data fields are carried over to a data warehouse.

Technical limitations

Also intertwined in information gathering and use of warehouse data, are the system constraints and technical limitations that "data miners" must work within. The following is a brief list of issues affecting the smooth operation of a data warehouse as it aims to support a range of needs:

- The large size of the database leads to incompatible run times
- The resources needed to support the load process and keeping a database up to date with current information. Depending on the size and setup of the system this may impact on and limit the time available for analysis.
- The risk of incorrect or inappropriate analysis. Those using the data need experience and knowledge of the base data and evaluation and research skills to prevent incorrect or inappropriate interpretation of the information - context free interrogation is not helpful.

Time specific warehouse data

One of the enrichments a data warehouse can bring is its dynamic potential, with information relating to and being stored for many points in time. Taxpayers' details, for example, are constantly changing, so when limiting or extracting data from the Data Warehouse we need to decide at what point in time we "freeze" the information – the current date or at a specific date in the past. By selecting a specified point in time we will be able to consistently base future queries on information held at the same time. Care must be taken when defining data needs, specifically how to best control the constantly changing warehouse data.

Confidentiality and privacy issues

A data warehouse increases data availability and unlocks the potential to link data with other databases and to drill down so individuals can be identified. Ethics, confidentiality and privacy issues come quickly to the fore. As well as the normal ethical guidelines for evaluations, Inland Revenue is governed by law (the Tax Administration Act) which places strict statutory restrictions on information disclosure. Given the highly confidential nature of tax administration data, careful attention to ethical protocols is needed when using data warehouse information internally for evaluation purposes and when we contract external research agencies to perform some part of the evaluation. Much of the information held is also commercially sensitive.

Conclusion

With the right technology, business know-how, in-depth knowledge of the data and evaluation and analytical skill, most evaluations can benefit greatly from having access to data warehouse based information. It can enrich population definition and profiling, sample design, response rates, understanding qualitative sample and responses, and enable complex analysis of causal questions. This potential needs to be tempered with consideration of the potential risks – in our case the complex and confidential nature of the information.

References

Bamberger, M, Rugh, J, Church, M & Fort, L, 2004, "Shoestring Evaluation", *American Journal of Evaluation*, vol. 25,no.1, pp.5-37.

Bogen, K, (1996), *The Effect Of Questionnaire Length On Response Rates -A Review of The Literature*, US Bureau of the Census (www.census.gov/srd/papers/pdf/kb9601)

Diem K.G, 2002, Maximizing Response Rate and Controlling Non-response Error in *Survey Research*, The State University Of New Jersey, New Brunswick

Frahm, RA, December 2003, *Sinking A Mine Shaft Into Data: Procedure Reveals Hidden Patterns*, Hartford Courant Staff Writer (www.math.ccsu.edu/larose/courant)

Inland Revenue, 2003, *Annual Report 2003*

Patton, M, 1997, *Utilization – Focused Evaluation*, Sage Publications, California.

Pearson NCS, 1997, *Increasing Response Rates*, (www.pearsonncs.com)

Zaima, A, 2003, The Five Myths of Data Mining, What Works: vol. 15, May 2003 (www.dw-institute.com/research/display.asp?id=6674)