A Quantitative Model to Support Data Quality Improvement

William Haebich, Simsion Bowles & Associates, whaebich@sba.com.au

Attempts to improve database accuracy need to address both the current accuracy of data holdings and the accuracy of inputs and updates. A mathematical model is proposed which can be used to set realistic goals for data quality improvement projects. It describes how the accuracy of a data view in a database is affected by the accuracies of new, or changed, instances of the data view. In particular the model can be used to calculate the time-lag between the initiation of the improvement and its ultimate effect.

1. Introduction

Those responsible for data quality improvement need to be aware of the dynamic nature of the databases they deal with. Not only do database sizes change, but also the accuracy of the data they hold will change. Accuracy can even vary when the business processes which feed the data are quite stable. Thus symptoms of decreasing accuracy may not be due to a deterioration in quality control: they may be an intrinsic aspect of the dynamics of the database itself.

Attempts to improve accuracy need to address both the current accuracy of data holdings (data 'stocks') *and* the accuracy of inputs and updates (data 'flows'). Focussing on one at the expense of the other can be a waste of time and effort.

The model described below can be used to set realistic goals for data quality improvement projects. The results of the model can be understood without an understanding of the underlying mathematics. What the model offers over and above intuition is a means for making reasonable estimates of the size and scope of the interaction between stocks and flows.

A simpler version of the model was originally developed as part of a consultancy. The need for it arose after an sample survey had been taken to establish the prevailing levels of accuracy of

the company's major databases. The survey results led to an appraisal of a number of existing and proposed initiatives to improve data quality.

2. A Working Definition of Data Accuracy

Data is a symbolic representation of the external world. A data view is of poor quality in as much as it does not adequately represent its intended subject, according to an intended use.

It is at least theoretically possible to test whether any given instance of data is correct by locating and identifying its subject and then comparing the two. A variety of factors can be built into whether or not the representation is deemed to be satisfactory. These include combinations of exact and partial matches between 'atomic' attributes and whether the attribute values agree within a specified tolerance (including timeliness). Its all up to the user of the data.

Data accuracy can be defined as a measure of data quality by counting the number of correct instances and computing the proportion of correct instances of the view, that is,

$$\text{accuracy of an data view} = \frac{\text{number of correct instances of the data view}}{\text{the total number of instances of the data view}} \qquad (1.)$$

According to this definition, an accuracy is a number between 0 and 1. It represents the probability that an instance of the data view, drawn at random from the database, will be correct. Different accuracies will be obtained for the same data view depending on the factors and interpretation rules chosen.

The following analysis makes no assumptions about how the accuracy measure is constructed as long as it is expressed as a proportion as shown in equation (1.)

3. Database Stocks and Flows

Any database is initially empty. It is filled by a series of inserts, possibly through a bulk load from other sources. Most databases are dynamic because they are subject to a continual stream of interleaved inserts, updates and deletions.

The database *stock* is the set of instances in the database at any chosen instant. A database flow is the changes which occur to the stock: either new instances which are inserted or updates and deletions which are applied to the stock (Figure 1). (Redmond 1992) refers to stocks as 'lakes' and flows as 'streams'.

Stocks and flows are time-dependent notions. An update or insert is a flow at one instant which becomes a stock instance in the next. On the other hand, a stock instance becomes a flow as it is deleted.

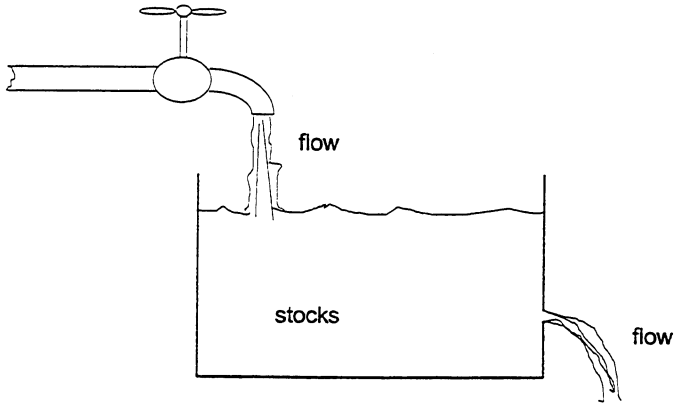

Figure 1: The concept of stocks and flows.

The accuracy of stock must be the result of the accuracies of the flows which generated it. Updates, inserts and deletions can occur at different rates and updates and inserts may have different accuracies. In order to quantify the net effect on the average accuracy of the stocks of these mixed effects it is necessary to analyse the flows process in more detail.

Figure 2. illustrates the interaction between inserts, updates and deletions over a period. The sequence shown here is chosen to best show how the three types of flow interact. In practice the different types of flow will occur simultaneously.

For a database of any size, and for a time-scale in years, the flows may be considered to be continuous. This situation can be modelled in terms of a differential equation.

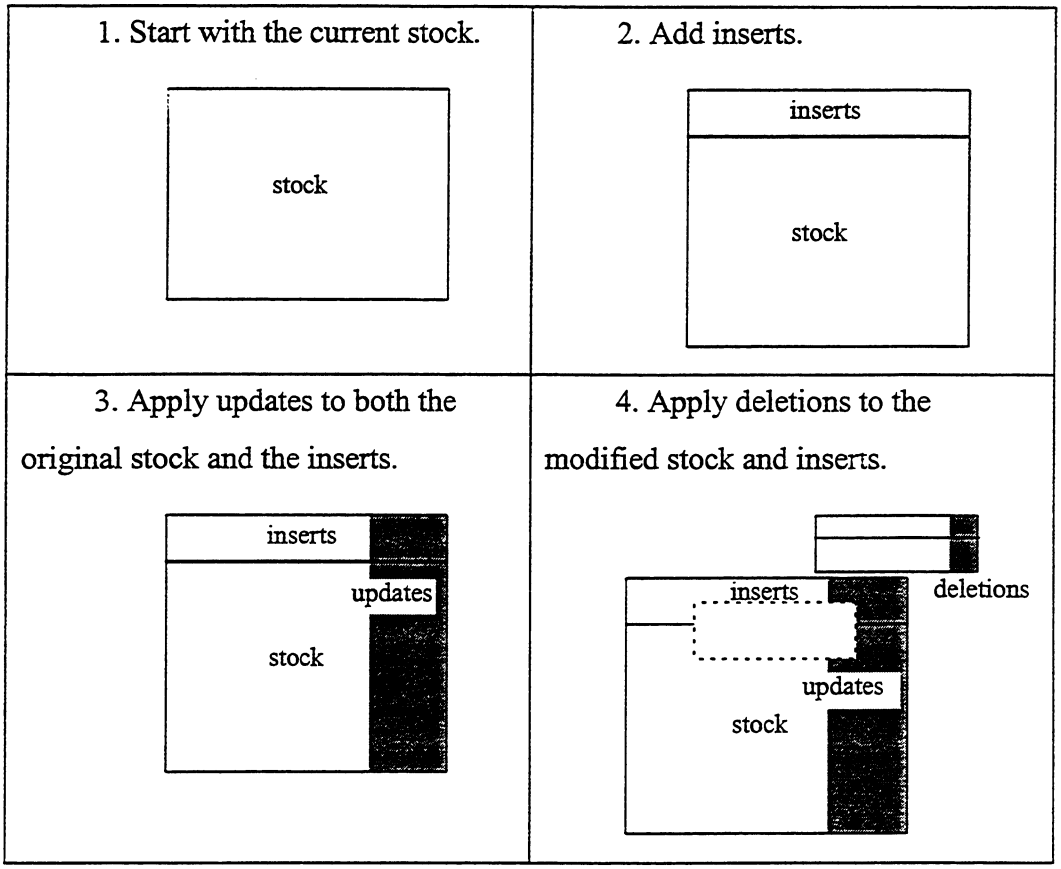

Figure 2: Interaction between the types of flow.

4. The Model

4.1 Database Size

In order to calculate accuracy, according to equation (1.), it is necessary to start with a picture of how the size of the database changes over time. Database entities typically represent some limited set of instances in the real world, a set of customers, a set of assets or a set of transactions. The logistic curve is often used to model population growth or economic growth (Montroll, 1974). Figure 3 shows a picture of logistic database growth. In this case the data is initialised on day 1 with 5,000 instances. At first the growth is roughly proportional to the current size of the database (that is, exponential) and then it slows down as the market ceiling of 1 million instances is approached, after 15 years.

Equation 2. shows an expression for logistic database growth. This is a slightly modified form of the standard logistic with an extra term ϕ_D to account for deletion.

$$n = \frac{A}{\phi_I + Be^{-At}} \quad \text{where } A = \phi_I N - \phi_D,\ B = \frac{A}{n_0} - \phi_I \text{ and} \qquad (2.)$$

t = Time in years

N = The size of the population intended to be captured in the database in a specific entity.

n_0 = The initial size of the database, at $t = 0$.

ϕ_I = A growth constant such that $\phi_I n$ represents the probability of capturing a new instance from the remaining population of size $N - n$.

ϕ_D = The proportion of the data deleted each year.

The effect of deletions means that the size of the database approaches a limit of $N - \frac{\phi_D}{\phi_I}$ as t becomes large. In this case the limit is 998,571.

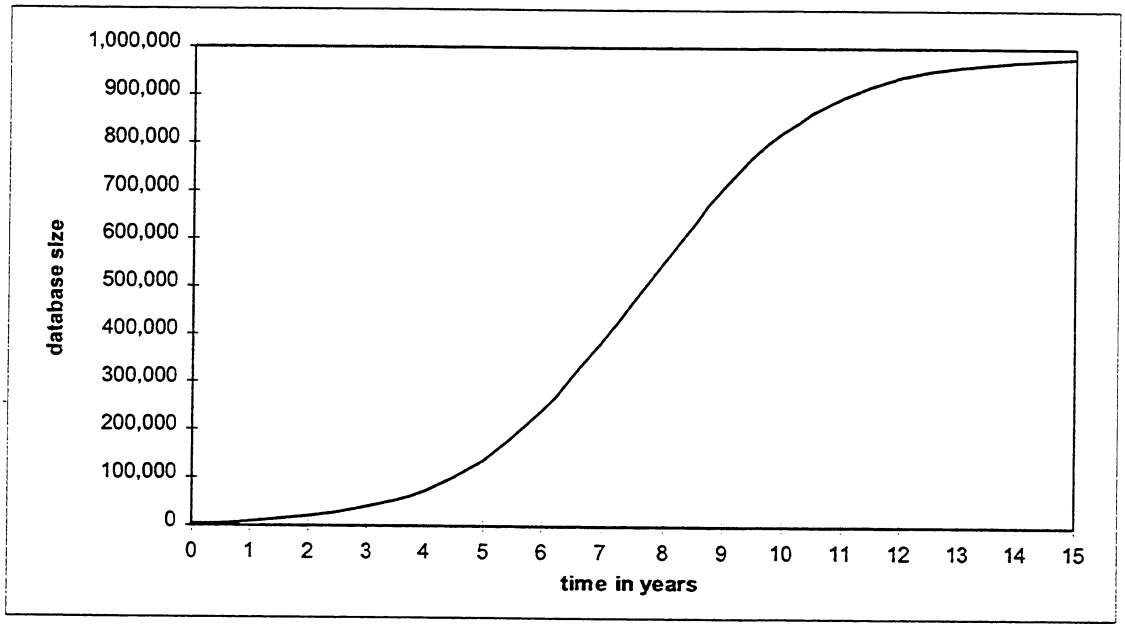

Figure 3: Logistic growth in database size for:

$$N = 1{,}000{,}000; n_0 = 5{,}000; \phi_I = 7 \times 10^{-7}; \phi_D = 0.01$$

See 9, Appendix: Sketch of the Model Derivation, for more detail.

The logistic may be inappropriate for some applications. The author is currently dealing with a forecasting database. The database contains time series which are used for decision

support. The forecasts relate to a fixed population of items which is unlikely to increase significantly but there are regular batches of new forecasts. The growth in the data is roughly linear with no deletion, that is $n = p_I t + n_0$ where p_I is the constant proportion of inserts per annum. Similar models to the one developed below 3. can be established for different growth patterns using the same techniques (see equation (7.) section 9.4).

4.2 Database Accuracy

The analysis in section 3 can be used to set up a differential equation for the accuracy of the database in terms of a set of parameters, including size. The equation can be solved (partially) to produce the expression:

$$a = a_I + \frac{(a_U - a_I)p_U}{ne^{(p_U+\phi_D)t}} \int_0^t ne^{(p_U+\phi_D)t}dt + \frac{(a_0 - a_I)n_0}{ne^{(p_U+\phi_D)t}} \qquad (3.)$$

where

a = The accuracy of the database at time t

a_0 = The initial accuracy of the database (at $t = 0$).
a_I = The accuracy of the data being inserted into the database.
a_U = The accuracy of the data being updated.
p_U = The proportion of data updated per annum.

These last 3 parameters are assumed to be constants. The derivation of this expression is sketched in Appendix, 9. A more detailed derivation can be supplied on request. The integral in equation (3.) cannot be evaluated analytically but it can be computed using power series or by standard numerical methods. The graphs displayed below were based on a numerical approximation and were generated in a spreadsheet. The spreadsheet is also available on request.

The model in equation (3.) shows that:

1. The contribution of the initial accuracy a_0 attenuates over the life of the database as the new entrants begin to swamp the initial instances, i.e.

$\frac{n_0}{ne^{(p_U+\phi_D)t}}$ approaches 0 as t increases.

2. There is a dynamic tension between the insert and the update accuracies such that a_I dominates during the early, fast growth of the database, but,

3. the accuracy approaches a weighted average of the insert accuracy and the update accuracy $\frac{a_U p_U + a_I \phi_D}{p_U + \phi_D}$ as the database becomes older, i.e. $\frac{p_U}{ne^{(p_U+\phi_D)t}} \int_0^t ne^{(p_U+\phi_D)t} dt$ approaches $\frac{p_U}{p_U + \phi_D}$ as t increases.

The model is based on a range of assumptions including:

- All variables are non-random. In practice the values of most of the variables described in the model will be subject to random fluctuations. We are interested in the broad relationship between stocks and flows and, at this level, such randomness can be ignored.
- The number of updates and deletions, in any period, is directly proportional to the size of the database.
- The rate of updates and deletions to records is independent of how long the records have been held in the database.
- Inserts, updates and deletions take place continuously over the year.

5. Patterns of Accuracy

The model shows that the accuracy varies over time if the initial, insert and update accuracies are different from each other. For example Figure 4 shows how it will change as the database size changes where $a_I > a_0 > a_U$.

It is reasonable to expect that these accuracies will differ because they are often generated via different processes.

In this example the organisation will experience a much higher level of complaints, foul-ups and re-work from poor data quality from year 8 onwards without any business practices in relation to the data having changed.

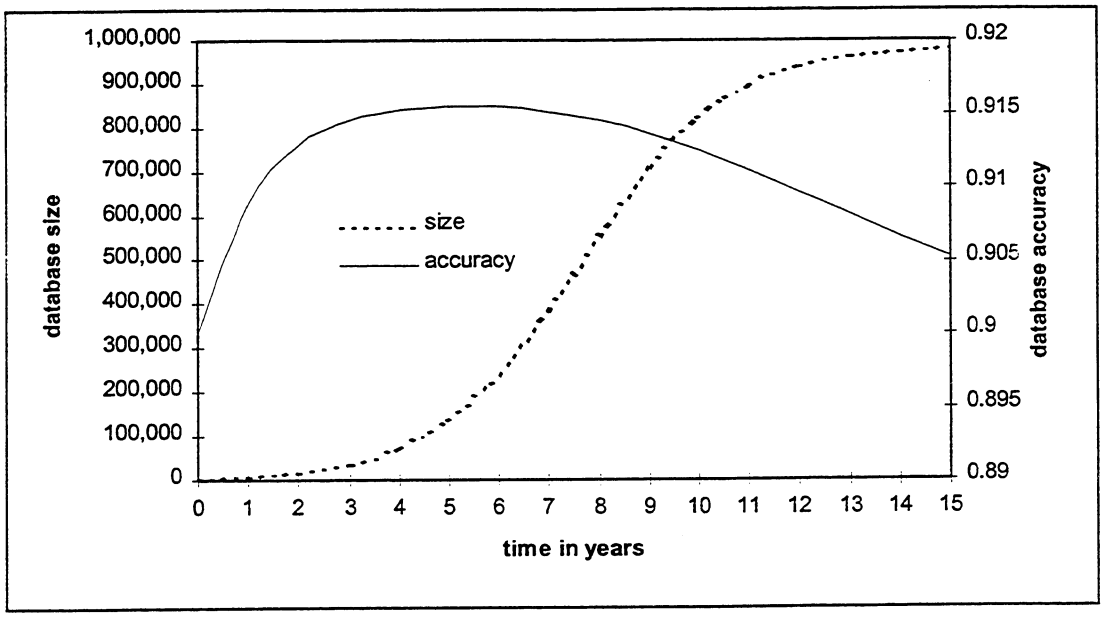

Figure 4: Change in accuracy relative to change in size where,

$a_I = 0.92, a_0 = 0.9, a_U = 0.89, p_U = 0.1.$

Figure 5 shows the above curve (A) together with 5 other possibilities, derived by permuting the same numbers (0.89, 0.9, 0.92) amongst the 3 variables (a_I, a_0, a_U). Its clear that the history of the database accuracy is very sensitive to these parameters and that it is very difficult to determine the cause of a trend in accuracy by observing the database in isolation.

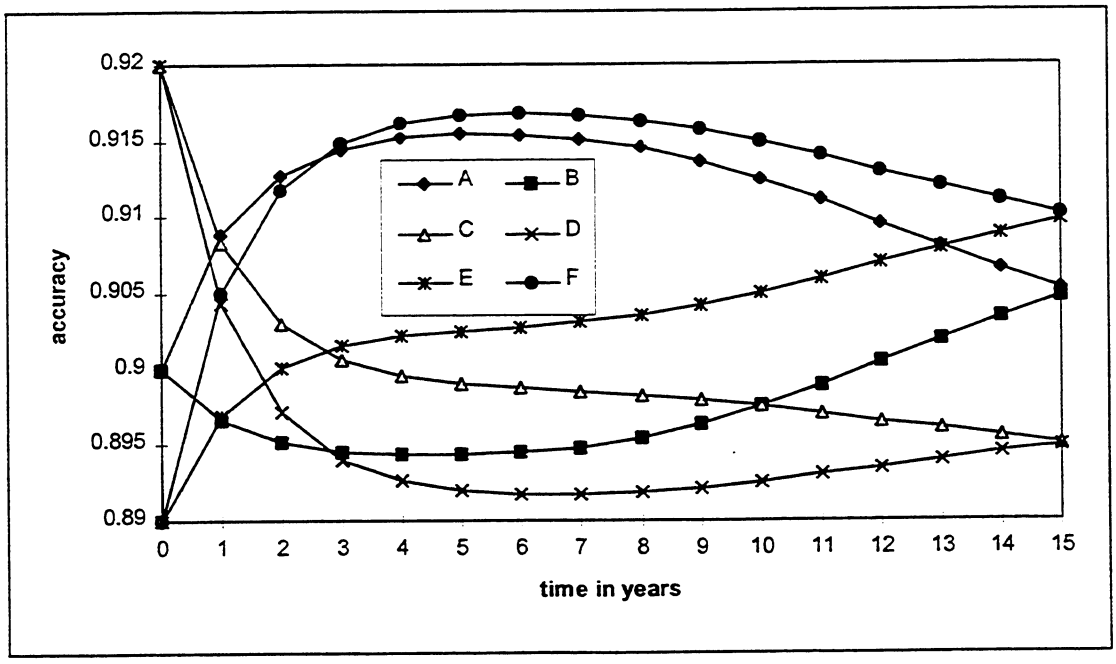

Figure 5: Six possible patterns of accuracy

	A	B	C	D	E	F
a_0	0.9	0.9	0.92	0.92	0.89	0.89
a_I	0.92	0.89	0.9	0.89	0.9	0.92
a_U	0.89	0.92	0.89	0.9	0.92	0.9

Key for Figure 5

6. The Results of Improvement Efforts

A common approach taken to data quality improvement is to apply data scrubbing or data alignment techniques to increase the stocks accuracy. If such a clean up is attempted without a concurrent attempt to improve the flows accuracy then the effect will die away fairly quickly as illustrated by the model in Figure 6. The dotted curve shows a sudden increase in stocks accuracy from 0.895 to 0.92 in year 5. This was applied in a situation corresponding to pattern (B) in Figure 5. The good work has largely been dissipated after 3 years because the poorer flow accuracy re-pollutes the stock.

Conversely the dashed curve in Figure 6 corresponds to a boost in insert accuracy to 0.92 in year 5. Although it takes about the same period of 3 years for the most of the effect to appear, this improvement is permanent. In this scenario the purer flow thins out the pollution in the stock and then keeps the accuracy high because the stocks are continually being replenished by new, cleaner data. The stocks accuracy cannot exceed 0.92 however because this is also the update accuracy.

Similar patterns will be observed for the other 5 cases in Figure 5. That is, a stimulus to stocks accuracy alone will decay back to the underlying curve after a few years but a jump in the flows accuracy will be maintained although it will take a comparable time to reach its full effect.

There are also differences in the effect of insert and update accuracies. In section 4.2 it was shown that database accuracy approaches a limit of $\frac{a_U p_U + a_I \phi_D}{p_U + \phi_D}$. Figure 7 depicts how the different patterns in Figure 5 yield different limits in the long term (50 years). If, for example, $p_U > \phi_D$ then every percentage improvement in a_U will have a bigger effect on the overall accuracy then the same improvement in a_I.

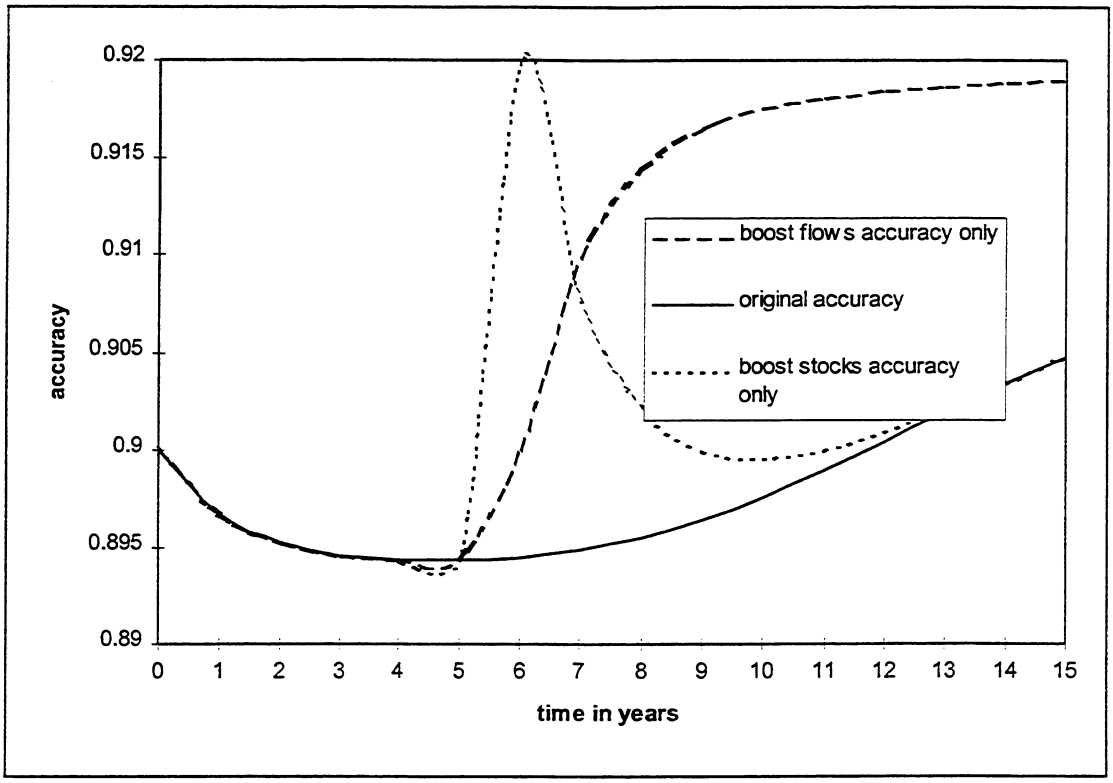

Figure 6: Responses to accuracy improvement.

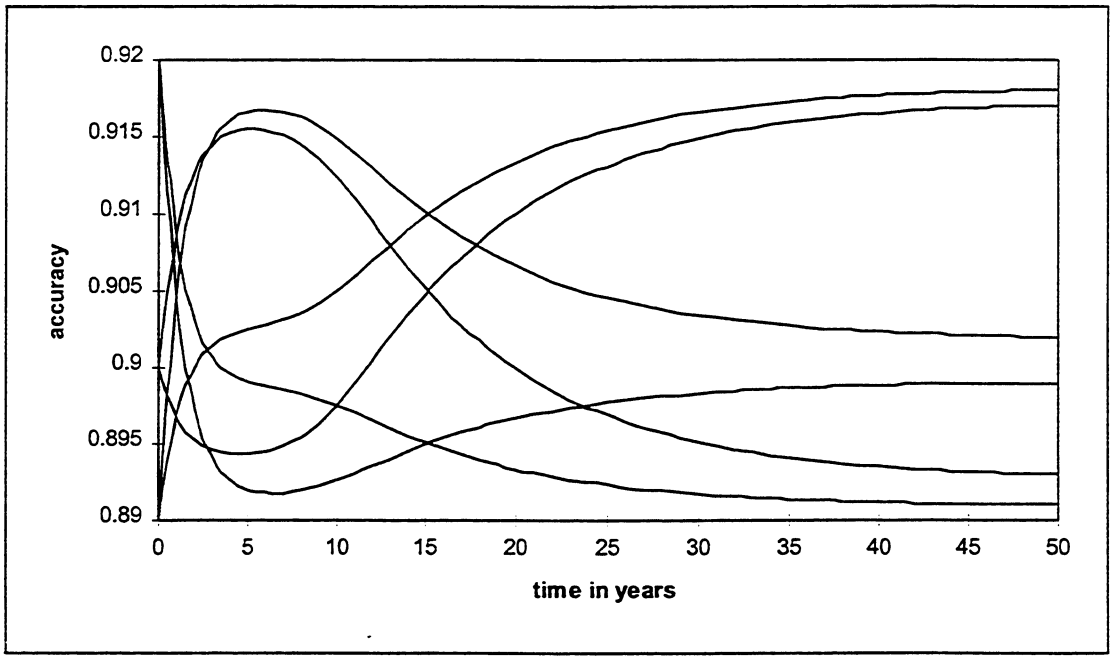

Figure 7: Patterns of accuracy in the long term

The model can be employed in a variety of other ways to evaluate the impact improvement initiatives. For example it could be used to examine the effect on accuracy of a bulk removal (archive or deletion) of old data from the database.

7. Fitting the Model

In order to fit the model in equation (3.) to a real case it is necessary to verify the assumptions set out in section 4.2 (including logistic growth) and to estimate values for the parameters. Estimates can be made using market information and measuring the database during a chosen year t as follows:

Size of the market	M
Ultimate market share	S
Number of instances of the data view at the start of the year.	n
Number of instances inserted during the year.	I
Number of correct instances inserted during the year.	c_I
Number of instances updated during the year.	U
Number of correct updated inserted during the year.	c_U
Number of instances deleted during the year	D

Then $N = M \times S$, $\phi_D = D/n$, $a_U = c_U/U$, $a_I = c_I/I$, $p_U = U/n$. Values for n_0 and a_0 can be taken from the database history, if it is available. ϕ_I can be estimated by inverting equation (2.), $\phi_I = A/n - Be^{-At}$, where t is the chosen year. The number of correct instances of various types can be estimated by sampling.

Ideally these parameters should be measured over several time periods to check that they are relatively constant and hence that the model assumptions are true. However most organisations do not measure the accuracy of their data at all (either stocks or flows) let alone keep a history of them. It is usually difficult to obtain a history of database size or flow rates. If there is insufficient information to calculate the parameters then educated guesses can at least provide some assistance in planning for data quality improvement.

8. Conclusion

The dynamic nature of databases makes data quality improvement difficult to manage. Since database accuracy may not be constant even under stable input conditions it is hard to determine whether a deterioration or an improvement in accuracy is intrinsic or due to external factors.

As with any form of quality control, measurement is crucial. In the data quality arena the measurement of a limited set of parameters associated with the data can yield a picture of what is happening and inform how best to improve the situation. In particular improving data stocks accuracy alone is likely to be ineffective. Ideally stocks and flows improvement should be tackled simultaneously but if there are limited resources then it is generally better to concentrate on flows improvement first.

9. Appendix: Sketch of the Model Derivation

9.1 Size

The standard logistic is derived from $n' = \phi_I n(N-n)$. We have modified this to allow for deletions by subtracting the term $\phi_D n$:

$$n' = \phi_I n(N-n) - \phi_D n \qquad (4.)$$

This can be solved along the lines set out in (Montroll, 1974) leading to equation (2.) in section 4.1, Database Size.

9.2 Accuracy

An equation for the accuracy can be developed by considering the number of correct instances, $c(t)$ (see Figure 2):

$c(t+dt)-c(t)=$			
	+	number of correct inserts	$a_I\phi_I n(N-n)dt$
	-	number of correct instances in the set of updated instances, before it was updated	$ap_U ndt$
	+	number of correct instances in the set of updates, after it was updated	$a_U p_U ndt$
	-	number of correct instances in the set deleted	$a\phi_D ndt$

$$= a_I\phi_I n(N-n)dt - c(t)p_U dt + a_U p_U ndt - c(t)\phi_D dt$$

Thus

$$c' = -(p_U + \phi_D)c + (a_I\phi_I(N-n) + a_U p_U)n \qquad (5.)$$

The accuracy is defined to be $a = \frac{c}{n}$, so that $a' = \frac{c'}{n} - \frac{cn'}{n^2} = \frac{c'}{n} - \frac{an'}{n}$

Substituting in this from equations (4.) and (5.) yields a differential equation in a:

$$a' + \left(p_U + \phi_I(N-n)\right)a = a_I\phi_I(N-n) + a_U p_U \qquad (6.)$$

This is a first order Differential Equation which can be solved by multiplying through by an integrating factor (Chaundry, 1969):

$$f = \exp\int_0^t \left(p_U + \phi_I(N-n)\right)dt = \frac{(\phi_I + B)ne^{(p_U+\phi_D)t}}{A}$$

transforming (6.) into:

$$ane^{(p_U+\phi_D)t} - a_0 n_0 = \int_0^t ne^{(p_U+\phi_D)t}\left(a_I\phi_I(N-n) + a_U p_U\right)dt$$

$$= \int_0^t e^{(p_U+\phi_D)t}\left(a_I\phi_I(n' - \phi_D n) + a_U p_U n\right)dt$$

$\int_0^t n' e^{(p_U+\phi_D)t} dt$ can be expressed in terms of $\int_0^t n e^{(p_U+\phi_D)t} dt$ using integration by parts. This allows the last line above to be transformed into equation (3.)

9.3 Power Series

The integral $\int_0^t n e^{(p_U+\phi_D)t} dt$ can be expanded to a power series as:

$$\int_0^w n e^{(p_U+\phi_D)t} dt = \frac{A}{B} \sum_{s=0}^{\infty} \left(-\frac{\phi_I}{B}\right)^s \frac{1}{(p_U+\phi_D+A(s+1))} \left[e^{(p_U+\phi_D+A(s+1))w} - 1\right]$$

$$\text{for } w < \frac{1}{A} \ln\left(\frac{B}{\phi_I}\right), \text{ and}$$

$$\int_w^t n e^{(p_U+\phi_D)t} dt = \frac{A}{\phi_I} \sum_{s=0}^{\infty} \left(-\frac{B}{\phi_I}\right)^s \frac{1}{(p_U+\phi_D-As)} \left[e^{(p_U+\phi_D-As)t} - e^{(p_U+\phi_D-As)w}\right]$$

$$\text{for } w > \frac{1}{A} \ln\left(\frac{B}{\phi_I}\right)$$

9.4 Generalisation

As indicated at the end of section 4.1, the logistic may not always be an appropriate model for database growth. In general the database size will be of the form:

$$n' = g(t) - \phi_D n$$

for some function $g(t)$. In this case the corresponding version of equation (6.) will be:

$$a' + \left(p_U + \frac{g}{n}\right) a = a_I \frac{g}{n} + a_U p_U \qquad (7.)$$

A more general formulation can be made by allowing for a mix of multiple flows with differing accuracies. The mixing of flows, although not the net effect on stocks, is addressed in (Ballou, 1985).

10. References

Ballou, D, P, Pazer, H, L, *Modelling Data and Process Quality in Multi-Input, Multi-Output Information Systems*, Management Science, vol 31, no 2, February 1985

Chaundy, T, W, *Elementary Differential Equations*, Oxford 1969

Montroll, E,W, Badger, W, W, *Introduction to the Quantitative Aspects of Social Phenomena*, Gordon Breach 1974

Redmond, T, C, *Data Quality Management and Technology*, Bantam, October 1992