



"turning data into dollars"

Tom's Ten Data Tips – September 2008

Missing Data

Missing data are a fact of life. No matter how hard we try, and how careful we assemble our data sets, there will always be missing data. In fact, sometimes data are supposed to be missing, for instance because a particular attribute does not apply to a person. In some cases, the pattern in *missing* data can be equally informative as the information present. In general, however, the effect of missing data is to limit the amount and quality of available information.

Missing data are usually replaced for one of several reasons:

- A NULL value isn't handled well by the RDBMS
- Statistical operations require non-NULL values
- Imputed values are expected to better represent the "true" state of affairs in downstream reports that are generated
- One is trying to maximize the value of all data *present* in the database

1. Missing Value Replacement Should Be Based On An *Explicit* Model

Regardless of which approach is used to impute (or drop) the missing values in your data, (or even if you choose *not* to replace any missings) "a" model is being used (see also tip# 10). Nowadays many tools come with some kind of "automatic" or "default" missing value replacement option. This may or may not be a good choice.

What is more important is that replacement doesn't 'happen to you', but should be a conscious choice, where the analyst has given consideration to the pro's and con's of any given approach. That's why he should be able to formulate (make explicit) the kind of model used for replacement. The underlying mathematics aren't nearly as important as the assumptions and consequences, so *what* the model does matters, rather than *how*.

2. Missing Data Are In Essence A Fault In The Data Model

Sometimes missing values in a table are rightfully missing, and sometimes they are not supposed to be missing. Some "missing values" are truly "*missing* values" (where no true value exists), but

many others are actually 'special' values or at present undefined values. Take the case of a survey. A particular question may not have been answered by a respondent, and this may cause a missing. Or, the question may have been answered with the answer "does not apply", "no opinion", or whatever options have been provided. Of course it is also possible that the respondent answered, but that the data entry person missed the answer.

What happens in this case, is that three different causes for a "non-response" all are lumped together, mapped onto the same value: NULL. It is better to assign a unique value for every "type" of missing (see also tip# 3). This should, ideally, lead to a situation where *no* missings remain. If the data are comprehensively modeled in this way, any new and hence unexpected NULLs are a sign some process is broken which may be cause for investigation.

3. Use Multiple Codings For Missing Data

Whenever the origins for why a field has become missing have different causes, each should preferably get a unique coding. So for instance in a survey with a 5-point Likert scale, the respondent might check "n/a", or the respondent might not have checked *any* option. A third possibility might be that the proper answer wasn't legible. And then of course the data entry person might have missed the item, or provided an impossible value.

When you code these options in a range from, say, 6-9 for these four scenarios it becomes easy to eliminate missings with a computer program. This still allows for special purpose reporting on the frequency of occurrence of different sources of missing. If many respondents skip certain questions, this can indicate sensitive topics.

In database transformation like ETL processes there can often be even more variations on the origins of missings, and again it is helpful to code them in a special purpose range of values.

4. Scrutinize Your ETL Process For MV Replacement

In many cases, ETL processes are designed to *cope* with missing values. This coping often consists of "filling in the blanks" where data feeds have unexpected empty values. Especially because data from several feeds are confronted and integrated, the ETL stage hosts an excellent opportunity to infer what appear to be missing fields from integrated feeds.

This kind of “inferred imputation” has two caveats:

1. The inference is based on business rules which are subsequently rarely ever scrutinized and not necessarily transparent (are they documented in the meta data??). Consequently, a whole stream of potentially erroneous imputations can be invoked that live on staying unnoticed (*unless* disaster strikes somewhere...).
2. The fact that a field was missing can be very valuable (predictive) information. By replacing the missing field with an inferred value, the information of the field previously being missing gets lost in the ETL process. The way to counteract this is to add an additional “previously missing” Boolean flag (see also tip# 6).

5. Never Replace Missings With A Constant

Although it sometimes appears intuitive, there are some nasty problems associated with replacing missings with a constant (often the mean or mode). Even if we keep track of *which* missings have been replaced (see also tip# 6), the biggest problems are the introduction of bias and underreporting of variance. Another problem is that for some records the global constant is a truly impossible value.

When a variable requires imputation of missings, one of the ground rules is that the association with other columns in the data set should remain unchanged. Unfortunately, this requirement is violated when missing is replaced by a constant. Another reason why missings should not be replaced by a constant (and a source of bias), is that a lower variance for a variable will make it appear a stronger predictor in a regression model, when all that has been done is replacement by a constant! In reality, a variable which contains (many) missings can not be a very good predictor.

6. Missing Data *Patterns* Need To Be Retained

For many (data mining) algorithms, it is necessary to replace missing fields by an appropriate value (regression, for instance). However, using standard statistical operations, there is no way to capture exactly which fields were missing after replacement.

To this end, it is good practice to append a table representing the pattern in the missings. This can be a single column with a separate value for each missing data pattern, or this may be a table with multiple columns, at most the number of variables in which missings

were replaced. In theory there are 2^n possible patterns for the missings, where n represents the number columns in which missings were replaced. In practice, the number of patterns is much, much smaller than 2^n , in particular in behavioral data sets.

7. MCAR, MAR, NMAR Patterns Decrease In Desirability, But Increase In Occurrence

The acronyms Missing Completely At Random (MCAR), Missing At Random (MAR), and Not Missing At Random (NMAR) are a bit of a misnomer. MCAR means there is no discernible pattern in the missings. The somewhat confusing term MAR means that although there is no significant association between *the target variable* and the pattern in missings, there *is* association within the dataset and missings. Your worst case is NMAR where 'missingness' is associated with other columns as well as the target variable.

So MCAR is the most 'desirable' pattern to have, yet it occurs the least in practice. NMAR is the most problematic (challenging) scenario to run into, yet it occurs the most in practice! MAR falls in between on both counts.

8. Hot Deck Substitution Is A Move In The Right Direction

Two terms that are frequently used are "Cold Deck" and "Hot Deck" substitution. "Cold Deck" substitution is a method whereby an 'external' constant is used to replace missings. So mean, mode or median substitution are special cases of a "Cold Deck" substitution (which is sometimes erroneously referred to as "mean substitution"). Incidentally, the mean has certain desirable statistical properties that the median is lacking but when the variable is (very) skewed, the median represents it much better.

"Hot Deck" substitution, sometimes referred to as "group mean substitution" means that groups of records are assigned the same value, and grouping is based on some other value(s) in the dataset. If for example weight is missing, it makes sense to impute a different value for men than for women since the latter weigh less on average. Of course the grouping can become much more elaborate, and then we gradually move towards sophisticated model based substitution.

9. SI Performs *Faster*, MI Works *Better*

MV replacement models can be separated in two clusters: single (SI) and multiple imputation (MI) models. Single imputation models only require a *single* pass through the data to provide a solution (hence their name). Multiple imputation models converge on a “best” solution after multiple passes through the data.

Examples of SI models are regression and Case Based Reasoning (CBR), also called Nearest Neighbor. Hot Deck substitution should be placed here as well. Regression models take the records where a field *is* present and build a regression model on the other columns to estimate the missing. CBR models look for records that are nearest in multi-dimensional space and infer the value of the missing from the N-closest records. Note this approach has conceptual similarities to K-means clustering.

MI models are the most advanced techniques available. The drawback is computational cost: (sometimes prohibitively) longer run times. They also require a bit more experience to parameterize. Examples are Structural Models and Expectation-Maximization (EM). These methods are usually not feasible in large databases.

10. Choose A MV Replacement Model In Line With Your Needs

The value of your data, or rather the misclassification costs of decisions that result from missings should drive your choice of imputation model. If you're after a fly, flypaper or a flytrap are better than a howitzer. One company that sought my advice was trying to run SAS Proc MI on a huge database with large numbers of missings. The value of imputed data was substantial but their tool broke down.

In some medical studies with very few cases and expensive treatments it makes sense to go “all out” to replace all missings as accurately as possible. In such a case one might want to test and develop different models *per variable*.

In other cases “good” is “good enough for government work” as they say (not to discredit the value of some of the data government holds!). Awareness of strengths and weaknesses of all MV Replacement models should guide you to a reasonable approach in a wide range of circumstances.