Title: **Cleaning Financial Data**

Summary: Increasingly, sophisticated methods are available for analyzing financial data and helping decision makers. In practice, the data that will be used can be full of errors. It is often the more sophisticated methods that seem to be particularly sensitive to the presence of bad values in the data. Therefore, it makes sense to deal with the bad data before the modeling takes place – improve the quality of the data and you are very likely to improve the quality of the results.

Increasingly, sophisticated methods are available for analyzing financial data and helping decision makers. Many of these methods have been described in articles that have appeared in Financial Engineering News. But in practice the data that will be used by these methods can be full of errors; it is dirty data. And it is often the more sophisticated methods that are most affected by dirty data, time series and variance models, such as GARCH, seem to be particularly sensitive to the presence of bad values in the data. While it is sometimes possible to use robust techniques that are less sensitive to bad observations, for example, using a median instead of a mean, it makes sense to deal with the bad data before the modeling takes place. Improve the quality of the data and you are very likely to improve the quality of the results. So before using the data it should be cleaned, that is, as many of the errors in the data as possible are corrected.

Here we will look specifically at cleaning numerical information, but the need for data cleaning is perhaps more important in text information where problems with misspelling and use of different abbreviations etc. complicate the process further.

What can be wrong with the data? There is a hierarchy of problems that are encountered:

1. No values have been input
2. Impossible values have been input
3. Inconsistent values have been input
4. Unlikely values have been input

While the first seems straightforward, there needs to be a distinction made between structural missing and observational missing. Structural missing will relate to values you would not expect to be there, for instance, share price changes will not be available when stock markets are closed at weekends or holidays. The models used need to be able to cope with such values, inventing values to fill such gaps is not a good way to proceed. On the other hand, observational missing are just values that have gone astray. Clearly where possible they should be looked for, but this may either be impossible or just too expensive.

Impossible values should be checked for by the data handling software, ideally at the point of input so that they can be re-entered. These errors are generally straightforward like negative prices when positive ones are expected. If correct values cannot be entered, the observation needs to be moved up the hierarchy to the missing value category.

Inconsistent values represent a more sophisticated error. This is when several values together break a rule. For example, if component values do not add up to an input total value. The problem is which one is wrong? Considered independently each may be valid and if some of the components are from past inputs it may not be possible to check the values. One possible approach is to hope that the methods considered below will shed light on the situation by indicating which components are least likely to be correct.

Unlikely values are those that are theoretically possible but cause some surprise. Consider the values in ($'000):

2, 3, 5, 7, 10, 2000

The value 2000 could be correct, but this is unlikely. It is more likely that instead of entering the value in $'000 they have entered the value itself. In this case methods should be able to say with reasonable certainty that 2000 is definitely an odd value. So it can then be treated as a missing value. But what about the values

2, 3, 5, 7, 10, 200

Methods are likely to pick out the 200 as being odd, but should it be rejected out of hand? Or should it just be investigated further? Maybe it is correct, but is it so unusual that you do not want it in your analysis?

The problem with large data sets is that the effort involved in chasing-up missing and checking suspect values can be too expensive and time consuming. Automatic methods are therefore needed that can do a reasonable job in cleaning the data. First we will look at approaches to handling missing values.

Given data with holes in it, either from original missing values or by designating doubtful values as missing, we need to be able to get suitable values. This is an imputation problem. There are two fundamental approaches.

The first is to find the observation that is most similar to the one with the missing value; this is called the donor. If there are more than one possible donor, then the donor used can be selected in different ways, either by choosing one at random or by selecting the first in the list. In these approaches the skill is in selecting the best variables to match the recipient and the donor.

The second approach is to use models to predict the missing value. The fundamental approach is

1. Fit a model.
2. Use to model to predict the missing values.

As with any modeling exercise the prime requirement is to select a suitable model. The better the model, the better the prediction used to replace the missing values. The model should try to include all the relevant information. So if the data is from a time series it makes sense to use a time series model. However, there is also a need to keep the model relatively simple, as it needs to be robust to changing circumstances. Local model may be more effective and avoid the problem of inaccuracy in modeling long-term trends.

To fit the model you need complete data. If there are only a few holes in the data then this is not a problem because a substantial amount of good data will be available. If there are lots then an iterative approach can be used. Starting with guessed values for the missing values, the model is fitted to this data and then used to produce better estimates for the missing values. This procedure is then iterated until there is convergence. If the fitting is by maximum likelihood and the prediction by taking the expected value then this is the EM algorithm.

A key point is that any relevant model can be used. The choice of model can come from statistics or machine learning. So techniques such as regression, neural networks, support vector machine and decision trees can all be used. The simplest model is just to assume the data follow a normal distribution with no other structure. This leads to the missing values being replaced by the mean. For multivariate

data the methods used in data mining for prediction may be used such as those available in the NAG Data Mining Components. For time series a simple exponential moving average model may be adequate but more complex ARIMA models may be used. It is worth noting that if a Kalman filter approach is used there is a natural way of skipping over and/or predicting missing values within the series.

To detect unusual observations the common approach is a 'how far off' approach, that is, to predict what the value is expected to be and see how close the observed and predicted values are. This is like treating an observation as missing, imputing the value and then comparing the imputed value to the observed value. So the same methods can be used. How do you decide on which observations are too far off? In some cases it may possible to use the statistical properties of methods to compute theoretical bounds but often the more practical approach is to use simulation to generated suitable cut-off points.

There is a complication, to use a predictive mode you need to know which are the good observation, something you are trying to find out! One tool to help is the use of robust methods. A common robust approach is to use M-estimators; these down weight any observation that is far from where it is expected to be. These should not be affected if there are a small number of bad values. If higher numbers of bad values are expected then methods based on the median approach can be used. These methods can also be combined with approaches that grow the size of the good data. Here the robust methods are applied to a small initial set of data that has the greatest likelihood of being good. Other observations are added to this set if they pass the criterion of goodness.

An alternative approach to finding possible bad values is the 'usual suspects' approach. This approach needs data that has identified bad cases in it. The good/bad category can then be viewed as a classification problem and standard classification methods such as logistic regression, neural networks and decision trees can be used to classify future observations as to whether or not they should be considered to be bad.

While the focus above has been on analytical methods, the use of visualization can often be a powerful tool. It is particularly good at picking out bad values that are occurring in a regular pattern. For example, simple surface plots will reveal holes or spikes. However, care is needed in distinguishing between the natural variability and the presence of bad values. Data is often more dispersed that we think.

A word of caution is needed at this point. First, while automatic methods can detect unusual values that cannot distinguish between values that are unlikely but true and those that are just plain wrong. It may be that you are happy removing all unlikely values because they are difficult to model but in doing so useful, if awkward, information may be missed. Second, any form of automatic data cleaning will have an effect on the results of any subsequent modeling. In general it is hoped that the cleaning will enhance the results, but it is possible that the cleaning may occasionally distort the results. The effects of data cleaning on the whole process needs to be examined and should not be treated in isolation.

Given a wide range of possible methods for both error detection and imputation, how can you compare them? One approach is to start with a data set you are happy with, and then perturb the data adding odd values to replace any missing values, and then apply the different methods that you are considering. The results can be evaluated using suitable criterion such as those suggested in Chambers[1].

The problems of automatic data cleaning are increasing being studied and new ideas are being tried out. In time, practical experience and research will separate the useful from the merely interesting, and suitable software will become readily available. This is one of the aims of the Euredit Project (see [1]).

References:

[1] Ray Chambers, Evaluation Criteria for Statistical Editing and Imputation, National Statistics Methodological Series No. 28, HMSO, 2001. Available at http://www.cs.york.ac.uk/euredit/

by Geoff Morgan
Geoff Morgan is the Data Analysis and Visualization Group Leader at Numerical Algorithms Group, Ltd.

Originally published by Financial Engineering News, June/July 2002 (www.fenews.com).

Numerical Algorithms Group
www.nag.com
infodesk@nag.com

