Uses and Misuses of Loss Data

In the field if operational risk management, it's hard to find good data. Internal loss data seem to be insufficient and external loss data appear to have many deficiencies. What are these data issues and what steps can a bank take to address them? **Ali Samad-Khan, Bertrand Moncelet** and **Thomas Pinch** discuss the operational risk data dilemma and offer a set of guidelines for banks that want to use loss data effectively.

In a world where business decisions are increasingly driven by hard facts and rigorous analysis, data and empirical models are commonly used to facilitate more educated decision making. When used properly, data and models can tell you something you don't know or can help you confirm or validate something you thought was true but weren't sure about.

Ali Samad-Khan

This is of course true when the data are of good quality and the models are robust. In such situations, data analysis and modelling are very straightforward processes and one can easily apply traditional methods to solve these simple problems. But what does one do when the data are not of good quality and the models (which can only be developed when one fully understands the data issues) are immature?

This describes the situation that exists today in the field of operational risk management (ORM). It is generally known that perfect data do not exist and that even the best available data are generally bad. However, in ORM, the data are horrible. Internal loss data appear insufficient and external loss data are affected by reporting biases and numerous idiosyncratic factors. This might suggest that it's just not possible to use these data in any sort of meaningful, scientific way.

Thomas Pinch

This article is the first of a two-part series that discuss the ways that data and models are commonly used in the ORM industry and explains the pros and cons of various alternatives.

The Fundamental Problem

Effective measurement of operational risk requires historical loss data. This requirement, however, poses two problems. First, most institutions don't have a lot of internal loss data. And second, many operational loss data sets have very "long tails," which means that to understand exposure to losses in the tail region requires very large amounts of data. To summarize: lots of data are required, but very little are available.

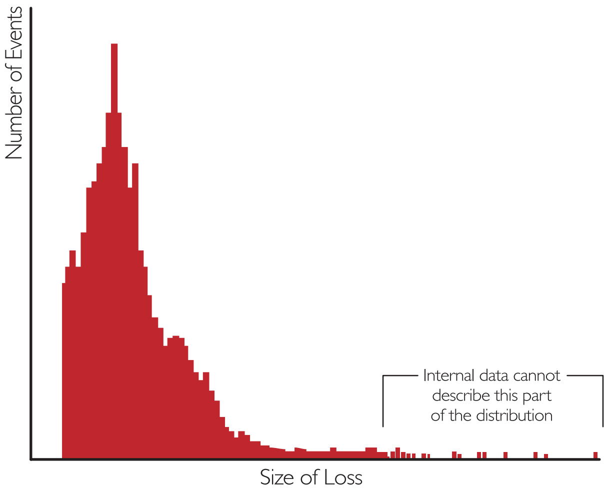
Bertrand Moncelet

To address this problem, many institutions have chosen to supplement their internal loss data with external loss data (that is, loss data drawn from other institutions). Bank regulators seem to agree with this approach. In fact, the use of "relevant" external loss data is mandatory for all banks intending to calculate capital under the Basel II advanced measurement approach (AMA) for operational risk. But this is problematic, because external loss data come from many different institutions, and these data are necessarily affected by idiosyncratic factors — including size, controls, culture, business processes, legal environment and geographic location — that are unique to the host institutions. It therefore seems impossible to apply these data to another institution without also transferring the attributes specific to the institutions from which these data were drawn. Furthermore, some of these data — particularly those drawn from public sources — are affected by reporting biases that seem to render them useless.

Based on these facts, it might appear that external data can't properly be used in connection with operational risk measurement. Then again, we note that the insurance industry has for decades been successfully using external data to calculate expected loss rates and the volatility (confidence intervals) around these estimates. This suggests that there may be scientific ways of addressing these data problems.

Before addressing these issues in detail, let's first describe what we mean by internal and external data, what types of data exist and what their respective deficiencies may be.

Figure 1: A Histogram of Internal Loss Data from a Typical Bank



Internal Data

Internal data are data drawn from the institution being modelled or, more specifically, from the unit within the institution being modelled. If you were to take the internal data from a bank with many years of loss experience and plot it as a histogram, it would probably resemble the graphical illustration in Figure 1(above).

This histogram reveals that the loss data are collected above a certain threshold.[1] It also makes clear that there is a distinct "body" and "tail" to this distribution and that the tail region contains a number of "outliers." However, if you were to analyze this data by risk class, you would notice that there is an obvious reason for this: these data are not identically distributed. The body consists mainly of execution errors (primarily high-frequency/low-severity losses), and the tail consists mainly of losses from other (primarily low-frequency/high-severity) risk classes, which have different distributional characteristics. When many data points from the low-severity classes are mixed with a few data points from the high-severity classes in a small data sample, the latter data appear as outliers. However, if one were to examine data from the high-severity classes in a large external loss database, one would observe that the data in these data sets are continuously distributed. In other words, these so-called outliers actually do follow a distribution of their own.

To avoid potential confusion, let us clarify that this does not mean, for example, that the execution, delivery and process management (EDPM) data set — which contains execution errors — represents no exposure to large losses; rather, it means that because this data has a lighter tail, internal data (where one often finds hundreds or even thousands of data points) may be sufficient for statistical analysis. This is because for thin-tailed distributions one needs a much smaller data sample to determine accurately the shape of the tail.

The reason internal data appear to be disjointed is that they contain data from several different non-homogenous distributions. To overcome this problem, one would have to increase the sample size, because even non-homogenous data sets — when aggregated in sufficient proportions — can be seen to follow a single distribution (when sufficient data are available) and are therefore modelled together.[2]

However, if we were limited to using internal data alone, we would have to wait several thousand years (in a static risk environment). An obvious alternative is to use external data.

External Data

There are, broadly speaking, three types of external data — public data, insurance data and consortium data. The following is a brief description of each:

Public Data. These data are drawn from publicly available information: newspaper reports, regulatory filings, legal judgments, etc. Because the smaller losses are less likely to be reported than are the larger losses, these databases have a strong size-based reporting bias. Public data vendors also tend not to capture losses below a high threshold (often one million US dollars[3]).

Because of this reporting bias, one cannot extrapolate frequency or severity parameters directly from the data. Suppose, for example, that the actual number of losses that had taken place at the $1 million level was 100,000 and at the $1 billion level was 10. If loss reporting was 1% at the $1 million level but 100% at the $1 billion level, the ratio of losses in the database would be observed as 100 to 1, instead of 10,000 to 1. This would indicate that the probability of a $1 million loss occurring was only 100 times higher than that of a $1 billion loss, instead of 10,000 times higher.

Since the reporting bias completely distorts the loss frequency and severity distribution parameters, using this data directly would result in misleadingly high value-at-risk (VaR) estimates.

There are two types of public data and three vendors who offer such products (see “Public Data Vendors” box, below).

Insurance Data. Insurance data represent losses that have been submitted as claims to insurance companies. Aon is currently the only vendor in this space. These data are captured only in risk classes where the insurance company has offered insurance coverage. Aon does not reveal the identity of the firms that experienced the losses.

Consortium Data. These are pooled sets of internal data submitted by member organizations. Because they consist of non-public information, an effort is made to keep the data confidential. While there are many consortia in existence today, they generally adhere to similar principles. The initiative which appears to have the most formal structure is the Operational Riskdata eXchange (ORX). To maintain confidentiality, ORX data does not contain descriptive information. The British Bankers Association (BBA) consortium does provide some form of descriptive information, but this data is classified by “cause” rather than by the Basel II “event” category; this is problematic for risk analytics.

Public Data Vendors

Class I – Quantitative
Vendors: Aon, Fitch (OpVar) and SAS

The three products offered by these vendors are similar in nature and scope. They are all designed to support risk analytics including risk-scenario analysis. All contain loss information, descriptions and supplemental analytic data (e.g., size of firm) and all appear to classify losses according to the Basel II standards. However, the data quality and size varies significantly among the three products.

Class II – Qualitative
Vendors: Fitch (First)

This product is available online and is updated frequently. It contains long write-ups and a lot of useful information, often obtained from multiple sources. It provides a comprehensive analysis of the circumstances that brought about the losses, but it does not provide any supplemental analytic data.

Data is classified according to internal classification structure, with mapping to Basel II categories. This database currently has significantly fewer large events than the two larger Class I products. The focus appears not to be to capture every event that has taken place, but instead to capture the events that are of greater relevance and interest to the subscribers (this may create an additional reporting bias).

The advantage of consortium over public data is that consortium data are not subject to public (media) reporting biases. However, consortium data have their own issues. In some organizations, internal reporting is not yet comprehensive; for example, in some situations, it is not clear that the largest losses are reported.

In addition, because consortium data are obtained from many organizations, categorization tends to be less consistent. Where consortium data do not contain ancillary descriptive information (e.g., ORX data), there is no way for a recipient of this data to remedy the categorization problem. Lastly, because consortium data represents only a subset of the loss data universe, it is not clear that consortium data contain sufficient information in the high risk classes (for example, unauthorized activities).

“Relevance” in the Context of External Data

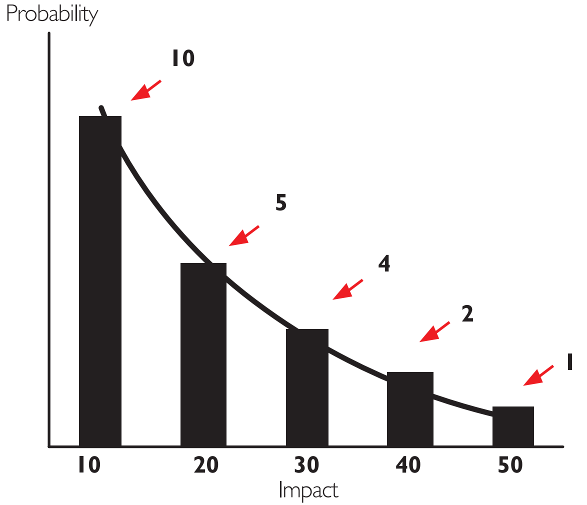
The Basel II regulations require that banks use “relevant” external data in their models. But what is relevance in the context of operational risk modelling? Or to put it another way: What actions should I carry out on external loss data so that they can be used meaningfully in connection with my bank’s internal loss data? Perhaps the answer is best expressed through the following set of guidelines:

- *Cautiously consider scaling individual loss data to the size of one’s institution*. There is strong empirical evidence to suggest that both loss frequency and severity are dependent on the size of a firm.[4] While scaling may be feasible, any scaling of individual losses should be based on empirically determined scale adjustment factors. The use of subjectively determined scale factors is inappropriate. Without knowing which scale variable to use (e.g., revenues or assets) or the nature (e.g., linear or log) and degree (e.g., 0.25 or 25) of the scale relationship, any subjectively based scaling is likely to degrade the data and do more harm than good.

- *Be wary of scaling individual losses to the quality of one’s internal control environment*. It seems reasonable to think that both loss frequency and severity are dependent on the quality of the internal control environment. However, no empirical studies have been conducted on this topic, and — given the difficulty in obtaining consistent, objective control scores for all firms in the loss data universe — it is clear that no such data are available today and its not likely that any such data will be available in the near future. Without such data, objective, controlled scaling is infeasible. And as explained above, since it will degrade the data, subjective scaling should not be attempted.

- *Don't try and select "relevant" data points from an external database based on the question, "Could this loss happen to me, given my internal control structure?"* Such an approach ignores the important point that a bank engaged in a certain business is exposed to the "inherent risks" of that business. So the relevant question is not, "Could this happen to me?" — after all, virtually anything could happen to you — but rather, "What is the relative probability of a loss of this size taking place in my business in relation to other losses of different sizes?" Since it is virtually impossible for anyone to subjectively determine this probability estimate, a better approach would be to use all external data and to let the data — in the context of a distribution — explain the relative probabilities associated with each loss level for an average bank in that line of business. See Figure 2 (below) for an example. If there were no systematic reporting biases in the external data, then good quality and bad quality banks would be equally represented in the database.

Figure 2: Relative Loss Proportions Before "Relevance" Adjustment



Therefore, the data distribution (above) would represent the average quality bank. While this may not reflect the exact control environment of my bank, it is an objective starting point. (The insurance industry uses this method for arriving at an initial risk estimate and then uses empirically-based factor analysis — for example, driving record, age or residence location — to adjust the results to match the unique characteristics of an individual client.) In fact, because one cannot know whether a certain loss is really more relevant than another, the arbitrary selection of "relevant" loss data points will often distort the data and could lead to the situation shown in Figure 3 (below).

Figure 3: Relative Loss Proportions After "Relevance" Adjustment

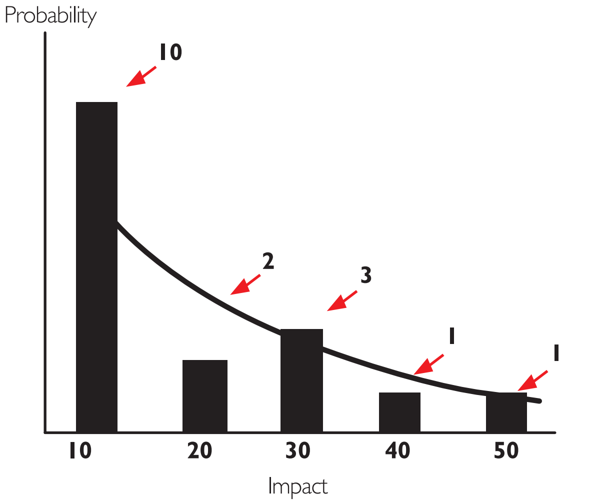


Figure 2, the original data set, depicts the inherent level of exposure for an average bank in this line of business. It shows the relative probability of a $10 million loss to a $20 million loss is 10 to 5 or 2:1. However, after subjective "relevance adjustment," the data in Figure 3 show this relationship has been changed to 10 to 2 or 5:1. Comparing Figures 2 and 3, one can see that the selection of individual data based on any sort of "relevance adjustment" completely distorts the data and renders it useless. Loss data (for severity) contains two integrally related pieces of information, the loss magnitude and the relative probability of the loss in the context of the distribution for which it was drawn. When loss data points are arbitrarily removed from a data set, the data set loses all informational value. Since there is no theoretically valid method of including or excluding individually "relevant" data points, the only reason for engaging in this type of data manipulation would be to force the results.

- *Think carefully before selecting "relevant" data points from an external database based on the question, "Is this organization similar to my organization in terms of control quality?"* This method may be feasible if objective means are used to determine what constitutes "similar control quality." This is done for whole data sets (groups of firms), not individual data points. For example, in circumstances where someone with detailed knowledge about the control quality across organizations (perhaps an external auditor)were to determine which business within which organizations had roughly the same control quality. However, including only firms of a similar control quality results in a smaller data pool.

  In addition, there is always the possibility of inadvertently introducing biases into this process. So it's important to recognize that the full data set naturally includes an equal number of good and bad institutions (assuming no systematic biases), and that any attempt to select "relevant" data may result in an inadvertent over-selection of either "good" banks or "bad" banks. Because it will be impossible to know which way the resulting data set is skewed, it is very hard to know whether the final results require an upward or down ward adjustment, if any. This issue is discussed in the next guideline.

> "When developing an operational risk model, the starting point should be an honest and objective assessment of the data; more specifically, it should be a detailed analysis of the nature of any biases present in the data. The next step is to determine how to address these biases in a manner that does not in any way degrade or distort the information contained in the data."

- *Think carefully before selecting "relevant" data points from external data based on the question, "Did this loss take place in a geographical region or type of business that is identical to my business?"* The same arguments expressed above apply here as well. In addition, one must consider that the issue is not whether this loss took place in a region or business that is identical to my business, but whether from an operational risk standpoint there is a material difference in the "inherent risk profiles" of the two geographical regions or businesses. In certain rare situations, where there is a legitimate, material difference in risk profiles, an argument could be made to objectively exclude an entire set of data points.

A Unique Challenge

Modelling operational risk is very different from modelling other types of risk because operational risk modelling is fundamentally about addressing the data issues.

When developing an operational risk model, the starting point should be an honest and objective assessment of the data; more specifically, it should be a detailed analysis of the nature of any biases present in the data. The next step is to determine how to address these biases in a manner that does not in any way degrade or distort the information contained in the data. Where there are systematic biases in the data, logical and objective means must be used to scientifically address these deficiencies. These methods will be discussed in part two of this series in an upcoming issue of *GRR*. ■

FOOTNOTES:

1. A solution to the problem of distribution-fitting loss data collected above a threshold can be found in most actuarial science textbooks.
2. This is a logical extension of the Central Limit Theorem, which the authors have validated through empirical analysis.
3. Because most losses under the threshold don't make the press, some vendors believe that capturing data in this region is generally not worth the effort.
4. Jimmy Shih, Ali Samad-Khan and Pat Medapa explained in *Operational Risk* magazine (February 2000) that there was a statistically significant relationship between size of firm and size of loss.

✎ **ALI SAMAD-KHAN** is the president of OpRisk Advisory, an operational risk management consulting firm.
BERTRAND MONCELET and **THOMAS PINCH** are consultants at OpRisk Advisory. They can be reached at opriskadvisory.com.