



Working Paper

A Review of Correction Techniques for Inherent Biases in External Operational Risk Loss Data

Shane Wilson – November 2007



Copyright

The material in this publication is copyright.

You may download, display, print or reproduce material in this publication in unaltered form for your personal, non-commercial use or within your organisation, with proper attribution given to the Australian Prudential Regulation Authority (APRA). Other than for any use permitted under the *Copyright Act 1968*, all other rights are reserved.

Requests for other uses of the information in this publication should be directed to APRA Public Affairs Unit, GPO Box 9836, Sydney NSW 2001 or public.affairs@apra.gov.au

© Australian Prudential Regulation Authority (2007)

Disclaimer

While APRA endeavours to ensure the quality of this Publication, APRA does not accept any responsibility for the accuracy, completeness or currency of the material included in this Publication, and will not be liable for any loss or damage arising out of any use of, or reliance on, this Publication.

Inquiries

For more information on the contents of this publication contact:

Shane Wilson,
Supervisory Support Division

Australian Prudential Regulation Authority
GPO Box 9836
Sydney NSW 2001
Tel: 61 62 9210 3000
Email: shane.wilson@apra.gov.au

Acknowledgements

Author: Shane Wilson

I would like to thank Harvey Crapp, André Levy, George Efstathakis, Sarah He and Emily Watchorn for their helpful comments and suggestions.

Contents

Objective	4
Data Sources	5
Inherent Bias	6
Reporting Bias	6
Control Bias	7
Scale Bias	8
Incorporation into AMA	10
Conclusion	11
References	12

Objective

The Australian Prudential Regulation Authority (APRA) states in its prudential standard (APS115), that ADIs¹ (hereafter banks) wishing to implement the Advanced Measurement Approach (AMA) to calculate their Operational Risk Regulatory Capital (ORRC), must incorporate either implicitly or explicitly, internal and external loss data (ILD and ELD), scenario analysis (SA) and business environment and internal control factors (BEICFs) into their operational risk measurement system.

Through the collection of ILD, banks are able to ascertain information on commonly occurring low impact operational risk losses. However, to complete their loss profile, both ELD and SA are used to supplement the bank's internal loss experience with the "infrequent yet potentially severe operational risk loss events"² not usually experienced in a banks loss history.

The main limitation of utilising ELD for such a purpose is the inherent biases apparent in the data. This paper explores the biases inherent in ELD and the subsequent problems faced when incorporating external data into the AMA.

¹ Authorised Deposit-taking Institutions (ADIs) are corporations which are authorised under the *Banking Act 1959*. ADIs include banks, building societies and credit unions.

² APS115 (Attachment B, para 32) Australian Prudential Regulation Authority.

Data Sources

Because operational risk loss recording and measurement is a relatively new discipline, sources of ELD are still in their infancy. There are three types of external loss databases currently available to banks, each with its own pros and cons:

- **Publicly available data** – These databases are made up of operational risk losses reported in newspapers, magazines, press releases, etc. Losses are generally collected over a common threshold and compiled in a database with information regarding the size and location of the firm. These databases typically contain only very large losses from large financial institutions, which are likely to attract the attention of media and shareholders. For example, a large loss which results in litigation is difficult to keep out of the media and as such is more likely to appear in public data. Dahen and Dionne (2007, p7) comment that using only publicly available losses to supplement ILD will fill out the extreme losses missing in the tail, but certain types of risks, such as ‘Execution, Delivery and Process Management’, will not be properly represented, as losses of this nature are unlikely to be reported in the media. It should be noted that losses sourced from public data may be subject to rounding errors arising from the nature of their collection and the imperfect information available in the media.
- **Insurance data** – Provided by insurance brokers such as AON, Willis and Marsh, this data originates from insurance claims made by financial institutions. Insurance data is generally quite reliable as the figures are sourced directly from the institutions (Dahen and Dionne 2007). However, the coverage of risk types depends on the range of policies held by institutions, the deductibles taken, and is restricted to those risk types which are insurable.

- **Consortium Data** – Consortium databases contain non-public data sourced from participating financial institutions. Providers such as Operational Risk Exchange (ORX) and the British Bankers Association (BBA) compile data from consenting institutions and provide anonymous statistical analysis to them with “data based upon the business lines and/or locations and/or events for which they provided data.”(Baud et al 2002, p3)

Consortium based data span a wider array of loss amounts, and cover a broader range of loss types than other external data sources. However, to ensure anonymity, any descriptive information regarding the origin of the loss is removed, and only general statistics are provided to the consortium members, making meaningful scaling of the data more difficult. Each consortium database has its own collection threshold (typically €20,000 – €25,000), but it is unknown what the internal threshold is for each of the contributing banks, and whether it falls above or below the consortium threshold. The major disadvantage of this source of data is that it does not allow event-by-event access to the losses. Therefore, these data can not be used to construct a database combining internal and external data. (Dahen and Dionne 2007)

Utilising publicly reported operational losses for modelling purposes involves a number of considerations. As Gustafsson et al (2006) point out, the losses chosen for modelling should be representative of the organisation as far as is practicable, facilitating the use of both internal and external data in the modelling process.

Australian banks have used a combination of publicly available and consortium external databases as an input into operational risk measurement. However, due to the biases inherent in ELD, the use of the data in its raw state is somewhat limited.

³ For example Fitch First Database has a threshold of US\$1m .

Inherent Bias

Reporting Bias

Reporting bias affects all external data types. It occurs when loss data in an external database is not considered a random sample of the population of data, indicating that each loss is reported with unequal probability (de Fontnouvelle et al 2003).

Reporting bias in consortium data relates to the unknown loss collection thresholds of the contributing institutions. Generally, consortium databases will have their own loss collection threshold 'A', and each contributing institution will have a differing loss collection threshold 'B'. However, banks using the consortium database as a source of ELD cannot assume a complete loss profile, as the institution's collection threshold 'B' is unknown.

- If 'B' < 'A' then valuable data collected by the institution is lost when the consortium truncates the losses at their threshold 'A', giving a false impression of the quantiles in their loss distribution.
- If 'A' < 'B' then it appears that the contributing institution suffered no losses below 'B', when in reality the institution's reporting threshold could be higher than 'B' or they made a strategic decision to only submit losses above 'B' to the consortium.

In current consortium databases, it is not possible to observe whether an institution has reported losses above or below the consortium threshold, because losses from individual institutions are not identified. In response, Baud et al (2002) developed a method of removing the collection bias from the ELD, whereby each institution's collection threshold is treated as an unknown variable following some statistical distribution. Their method uses statistical techniques to find the optimal (expected) value of the threshold for each institution, allowing both ILD and

ELD to be pooled together. As noted by the authors, the fundamental limitation in their method is that they assume that ELD is drawn from the same loss distribution as ILD, but with external data truncated above a common threshold. Because this assumption is not statistically true for all data sets, results become spurious if the assumption is violated. Their method is also computationally intensive where there are many contributing institutions.

In insurance data, reporting bias occurs because the probability that an operational risk loss is claimed for, and thus included in the insurance data set, depends on the size of the deductible and the type of insurance policy. Because insurance datasets share similar characteristics to consortium datasets, similar methods for removing reporting bias can be used.

In publicly available data, larger firms and losses are more likely to be reported in the media due to factors such as the size and nature of the loss etc. If the probability of a loss being reported increases with the severity of the loss, then there will be an over-representation of large severity losses in the database. For example, a database that hypothetically reports every \$100m loss but only one in every ten losses of \$20m, would overstate the proportion of \$100m losses to \$20m losses by a factor of 10, giving a false impression of the population of losses. As such, using publicly available data without correction for reporting bias may lead to an overstatement of large losses, thus inflating the capital requirement (Baud et al 2002).

Gustafsson et al (2006) propose a correction method whereby Subject Matter Experts (SMEs) are asked to estimate the extent to which reporting bias appears in publicly available data per Basel risk type at different loss sizes. These estimates are then used to derive an under-reporting function which, when combined with publicly available data, creates an estimate of a bank's true loss profile. However, using SMEs to estimate the under-reporting bias can lead to additional biases instilled in the estimates, arising from the inherent subjectivity of the probability elicitation process.

De Fontnouvelle et al (2003) have used the methods of Baud et al (2002) mentioned above and applied them to public datasets. They state that “an operational loss is publicly reported only if it exceeds some unobserved truncation⁴ point. Because the truncation point is unobserved, it is a random variable” (de Fontnouvelle et al 2003, p10), which can be modelled using a random truncation model.

Few Australian banks try to correct for reporting bias in external databases. Some have stated that by not correcting for reporting bias they are adding conservatism to their capital figure. For example, in public databases the over-representation of large losses may lead to an increase in capital. However, by not correcting for reporting bias, any parameter estimates drawn from the data will be similarly biased and produce spurious results, which are not representative of the banks loss profile.

Control Bias

Control bias appears in all external databases. It refers to the relevance of losses that come from institutions with different control mechanisms. All losses arise as a consequence of a specific set of circumstances due to a lack of, or failure in controls. As such, not all losses will be relevant to all banks. In such cases, institutions use simple rules to obtain a subset of the external database that is most relevant to their own operations and control structure. For example, a retail bank that operates predominately in residential and commercial mortgages with no trading operations would not experience any losses due to rogue trading. Thus banks must recognise the contrast between losses that are relatively unlikely (such as terrorist attacks), which must be incorporated into the data collection process, and those that should be excluded on the grounds of irrelevance.

Khan, Moncelet and Pinch (2006, p22) warn against selecting relevant data points on the basis of similar quality control standards unless an objective way of doing so is formulated. If a bank was to filter external data to eliminate control bias then it would need to identify which businesses, in which organisations have similar control structures to their own, and eliminate those which do not meet certain criteria. However, due to the disclosure constraints of external data, it is difficult to know what control structures are in place for each business line, from each institution.

When filtering ELD, Khan et al (2006, p26) suggest that a bank’s subject matter experts should ask themselves not if the loss could occur, but “what is the relative probability of a loss of this size taking place in my business in relation to other losses of different sizes?” Any answer given would be subjective, given that virtually any event has some probability of occurring. Khan et al (2006) suggests that banks should not ‘cherry pick’ losses from relevant subsets of data (i.e. relevant to the user of the data in terms of their business operations), because one cannot know whether a certain loss is really more relevant than another. Instead, a better approach would use “all relevant external data, so that the data – in the context of a distribution – can explain the relative probabilities associated with each loss level for an average bank in that line of business” (Khan et al p21, 2006).

⁴ The term “truncation point” refers to the unobserved, observation-specific random variable that determines whether a loss event is publicly reported and included in the external databases.

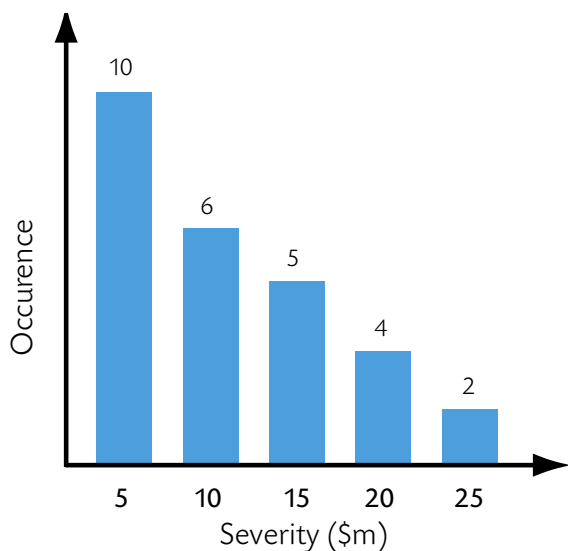
To illustrate, Figure 1 depicts the loss frequencies and severities for a relevant business line taken from an external loss database. It shows that the relative probability of a \$5m loss to a \$15m loss is 2:1. In Figure 2, notice that if individual losses are ‘cherry picked’ from the distribution because they are deemed ‘more relevant’, the relative probability of a \$5m loss to a \$15m loss is now overstated at 5:1. Looking at figures 1 and 2 it can be seen that picking individual data points from a relevant subset of losses completely distorts the loss profile for that business line. Khan et al (2006) points out that loss data essentially contains two pieces of information, the magnitude of loss and its relative frequency. By cherry picking losses a bank is effectively rendering the external data worthless as it removes the relative probability of loss and distorts the underlying risk profile.

Scale Bias

Scale bias is apparent in all external loss databases, and exists because the losses recorded within each database come from institutions of different magnitudes (in terms of their operations, assets, number of employees, revenue, etc.). As a result, many banks use a severity scaling mechanism applied to losses within the external database to scale loss values up or down.

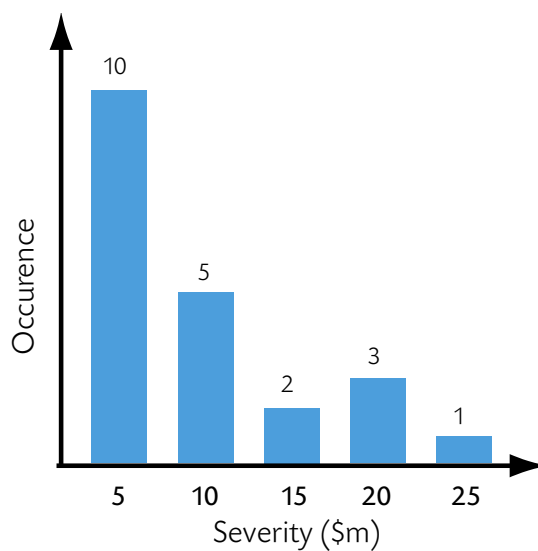
Shih, Khan and Medapa (2000) have shown that there is some correlation between both the frequency and severity of a loss and the size of a firm. However, the problem lies in determining a proxy for size and a scaling equation that adequately describes the relationship between the two. They found that approximately 9% of the variation in loss size can be explained by the size of the firm (using revenue as a proxy for size) using a non-linear weighted least squares model. Because so little of the variation in loss size can be explained by a firm’s size, the merits of scaling by only such a factor are questionable.

FIGURE 1



Loss proportions before Individual Losses Removed

FIGURE 2



Loss proportions after Individual Losses Removed

Adapted from Khan et al (2006, p21)

Na (2004) suggests that an operational risk loss amount can be broken into a component common to all banks and an idiosyncratic component. The common component explains the variation in the macroeconomic, geopolitical and cultural environment, whereas the idiosyncratic component captures the risk and control environment faced by individual banks.

Dahen and Dionne (2007) suggest that no single factor can explain a significant proportion of the variation in loss size, but instead a combination of factors, including firm size, location, business line and risk type, is more suitable.

Furthering the work carried out by Na (2004), Dahen and Dionne (2007) incorporated additional scaling factors into the idiosyncratic component of the loss equation in an effort to explain a greater proportion of the variation in loss severity. As a result, a normalisation formula was developed which scales a loss of bank X to its equivalent value in bank Y, through the quotient of some function of their idiosyncratic parts.

$$Loss_Y = \frac{g(Component_{idio})_Y}{g(Component_{idio})_X} \times Loss_X$$

To use such a scaling method in practice requires external databases to provide more institution specific information for each loss. However, due to the confidentiality requirements of consortium databases, this type of information is currently only available in public databases.

Literature to date has focused on the severity side of scale bias, but as Dahen and Dionne (2007) show, frequencies can also be scaled to more closely reflect a bank's risk profile. They scale frequencies using a Poisson and Negative Binomial probability distribution truncated at zero (i.e. at least one loss has occurred). The parameters for each distribution are estimated using regression analysis involving a series of scaling factors, including assets, location, business line and risk type. It was shown that because of the equi-dispersion properties of the Poisson distribution, the Negative Binomial distribution proved superior for modelling operational risk loss frequencies.

Theoretically, if a bank scales external losses so that they are more reflective of the risk profile of the bank, then the final capital charge should be more representative than if unscaled ELD losses were included. As such, banks should ensure that their chosen scaling methodology can be supported by empirical analysis to ensure the appropriateness of the scaling factor to the frequency and severity of operational risk losses (Khan et al 2006).

Australian banks using ELD explicitly in their measurement methodology currently employ simple linear methods of scaling with respect to a size proxy. Although some have explored more sophisticated methods of scaling, no significant relationship has been determined. External loss data frequencies are not widely used explicitly in the measurement model because, in the banks' view they tend to be overstated, even after scaling. However, some banks use the unscaled frequencies as a benchmark for scenario likelihoods. As data in external loss databases becomes richer and more information becomes available, scaling methodologies will be able to take into account a greater variety of variables and thus attempt to explain the institution specific behaviour of operational risk losses.

Incorporation into AMA

Australian banks have generally used one of three main methods for including ELD into their risk measurement models:

- **Integration** – Integrating ELD with other data sources and using them indiscriminately in a consolidated dataset assumes that all data originates from the same underlying probability distribution. External databases are made up of a sample selection of financial institutions, some of them with better control structures than others. To assume that losses originating from different financial institutions, in different locations and regulatory environments will belong to the same distribution as losses originating from an Australian bank seems questionable.

Statistical tests exist that are able to provide quantitative evidence regarding the equality of the underlying probability distributions of two data samples. In the event of distributional inequality, a bank may try to select individual losses from a relevant subset of external data to supplement their own ILD, but as discussed above this distorts the data and hence has limitations in its usefulness. Integrating ELD into the measurement model in this fashion generates a large amount of uncertainty, which may be difficult to measure and mitigate with commensurate conservatism. Banks tend to be moving away from integrated models, as more efficient and robust methods of loss measurement are developed.

- **Model Independently** – ELD can be modelled independently from other data sources to obtain a separate ELD ‘component’ of regulatory capital. This component is then combined⁵ with the capital derived from the other data components⁶, to obtain a final capital charge.

Currently, scaling is undertaken to ensure that the capital charge is not overstated due to excessively⁷ large losses occurring in ELD. Methods of scaling are currently quite simple due to the lack of available information in external databases, but as discussed above more advanced scaling methods are emerging as data improves.

Because ELD tends to be biased towards high severity losses and certain risk types, banks need to ensure that all loss sizes and risk types are adequately represented in the final capital requirement by means of the other data sources.

- **Implicitly Incorporate** – Banks not wishing to use ELD explicitly in their operational risk measurement model, perhaps due to the uncertainty created by the inherent biases, have instead used ELD as a reference source in the formation of dependence structures and in scenario analysis workshops. SMEs are given pre-assessment information containing ELD to aid in the elicitation of scenarios as well as their frequency and severity estimates. This is especially useful for extreme losses for which the bank typically has little data. By using ELD implicitly in the measurement model, banks are not required to transform or scale the data, as it is used only as an indicator of what the frequency and severity might look like.

⁵ Using weighted or average or variance minimisation techniques etc.

⁶ Scenario Analysis, ILD and BEICFs.

⁷ Deemed excessive in relation to the bank’s risk profile.

Conclusion

Banks applying to APRA for the use of the AMA for operational risk must incorporate ELD either implicitly or explicitly into their risk measurement model.

Because of biases inherent in the data, there has been no convergence on the methods for integrating ELD into the risk measurement process, and removing the biases is often computationally intensive and difficult to implement on large data sets. However, as the industry evolves and improved methods for removing biases are determined, a greater reliance on ELD is expected to drive the modelling of severe operational risk losses and hence risk capital.

Most banks have limited the direct influence of ELD in determining their capital outcome because of the unmeasurable uncertainty it creates. Instead they rely on ELD as a reference point and benchmarking tool for scenario analysis. Those who use ELD implicitly in the capital calculation have taken due care to ensure that the uncertainty created by its use is directly correlated with its overall influence on the final capital number.

The research cited in this paper shows that there are methods to mitigate the biases inherent in ELD. However, applied research by banks needs to be conducted to see if academic research can be applied in practice. Due to the current limitations of external data it seems unlikely that any bank will be able to completely eliminate biases from their dataset until more efficient techniques are developed. Until then, banks are able to improve the quality of external data to some extent using the techniques mentioned, allowing for more solid conclusions to be drawn.

References

- Australian Prudential Regulation Authority (APRA) (2007) 'Draft Prudential Standard APS 115 Capital Adequacy: Advanced Measurement Approaches to Operational Risk', June 2007.
- Baud, N., Frachot, A and Roncalli, T. (2002) 'Internal data, external data and consortium data for operational risk measurement: How to pool data properly?' Groupe de Recherche Opérationnelle, Crédit Lyonnais, France, June 2002.
- Cagen, P. (2005) 'External Data: Reaching for the Truth' Algorithmics Incorporated, December 2005.
- Dahen, H. and Dionne, G. (2007) 'Scaling Methods for Severity and Frequency of External Loss Data', Working Paper 07-01- Canada Research Chair in Finance, January 2007.
- De Fontnouvelle, P., Jesus-Rueff, V., Jordan, J., Rosengren E. (2003) 'Using Loss Data to Quantify Operational Risk', Federal Reserve of Boston, April 2003.
- Frachot, A. and Roncalli, T. (2002) 'Mixing internal and external data for managing operational risk', Groupe de Recherche Opérationnelle, Crédit Lyonnais, France, January 2002.
- Gustafsson, J., Nielsen, J.P., Pritchard, P. and Roberts, D. (2006) 'Quantifying operational risk guided by kernel smoothing and continuous credibility: A practitioner's view', *The Journal of Operational Risk* 1 (1): 43-57.
- Gustafsson, J., Guillen, M., Perch Nielsen, J. and Pritchard, P (2007) 'Using External Data in Operational Risk', Available at SSRN: <http://ssrn.com/abstract=871181>, January 2007.
- Khan, A., Moncelet, B. and Pinch, T. (2006) 'Uses and Misuses of Loss Data' Global Association of Risk Professionals, May/June 2006.
- Na H., Miranda, L., van den Berg, J. and Leipoldt, M. (2005) 'Data Scaling for Operational Risk Modelling', ERS-2005-092-LIS - Erasmus Research Institute of Management- Report in Research Management.
- Pezier, J. 2002, 'A Constructive Review of Basel's Proposals in Operational Risk', ISMA Discussion Paper in Finance 2002-20, University of Rebankng, September 2002.
- Shih, J, A. Samad-Khan, and P. Medapa (2000), 'Is the Size of an Operational Loss Related to Firm Size?' *Operational Risk Magazine* 2, 1.
- Yap, J. (2006) 'External Data and Operational Risk: Practical Issues and Solutions', Actuarial Research Essay – University of Melbourne, October 2006.



Telephone
1300 13 10 60

Email
contactapra@apra.gov.au

Website
www.apra.gov.au

Mail
GPO Box 9836
in all capital cities
(except Hobart and Darwin)