Quantitative modeling of operational risk losses when combining internal and external data sources

Jens Perch Nielsen (Cass Business School, City University, United Kingdom)
Montserrat Guillén (Riskcenter, University of Barcelona, Spain)¹
Catalina Bolancé (Riskcenter, University of Barcelona, Spain)
Jim Gustafsson (Ernst and Young, Denmark)

Abstract

We present an overview of methods to estimate risk arising from operational losses. Our approach is based on the study of the statistical severity distribution of a single loss. We analyze the fundamental issues that arise in practice when modeling operational risk data. We address the statistical problem of estimating an operational risk distribution, both in standard abundant data situations and when our available data is challenged from the inclusion of external data or because of underreporting. Our presentation includes an application to show that failure to account for underreporting may lead to a substantial underestimation of operational risk measures. The use of external data information can easily be incorporated in our modeling approach.

1. The challenge of operational risk quantification

In the banking sector, as in many other fields, financial transactions are subject to operational errors. So, operational risk is part of banking supervision and it is also a key component of insurance supervision.

Regulators require to measure operational risk as part of the indicators for solvency. This research area has exploded in the last few years with the existence of the Basel agreements in the banking industry and Solvency II in the insurance sector and partly because the financial credit crunch has accelerated our appetite for better risk management. These two regulatory frameworks have set the path for international standards of market transparency of financial and insurance service operators (Panjer, 2006; Franzetti, 2010).

Measuring operational risk requires the knowledge of the quantitative tools and the comprehension of financial activities in a very broad sense, both technical and commercial. Our presentation offers a practical perspective that combines statistical analysis and management orientations.

In general, financial institutions do not seem to keep historical data on internal operational risk losses. So, models for operational risk quantification only have a scarce sample to estimate and validate on. Therefore, nobody doubts that complementing internal data with more abundant external data is desirable (Dahen and Dionne, 2007). A suitable framework to overcome the lack

¹ We thank the Spanish Ministry of Science / FEDER grant ECO2010-21787-C0301 and Generalitat de Catalunya SGR 1328.
Corresponding author: mguillen@ub.edu
of data is getting a compatible sample from a consortium. Individual firms come together to form a consortium, where the population from which the data is drawn is assumed to be similar for every member of the group. We know that consortium data pooling may be rather controversial because not all business are the same, and one can raise questions about whether consortium data can be considered to have been generated by the same process, or whether pooled data do reflect the size of each financial institution transaction volume. We will assume that scaling has already been corrected for or can be corrected during our modeling process, assuming that external data provide prior information to the examination of internal data. So, here we will not address pre-scaling much specifically when combining internal and external data sets.

Many different issues arise in practice when financial managers analyze their operational risk data. We will focus on four of them:

(i) A statistical distributional assumption for operational losses has to be fitted and it must be flexible enough.
(ii) Sample size of internal information is rather small.
(iii) External data can be used to complement internal information, but pooling the data requires a mixing model that combines internal and external information, and possibly a pre-scaling process.
(iv) Data used to assess operational risk are sometimes underreported, which means that losses do happen that are unknown or hidden. Data on operational losses are often not recorded below a given threshold due to existing habits or organizational reasons, which means that information on losses below a given level is not collected. One should not confuse underreporting with un-recording because the latter is a voluntary action not to collect all those losses below a given level.

The assumption that a normally distributed random variable is suitable to model operational losses is simply not valid. Operational losses are always positive and, many of them correspond to rather low values, while a few losses can be extremely high. For this reason, our approach to modeling operational losses completely departs from the Gaussian hypothesis and even from the log-normality assumption. Instead, we propose semi-parametric models that are much more flexible and that are much more suitable in this context than the normal distribution assumption.

Sparse data, underreporting, and extreme values are three fundamental reasons for operational risk being extremely difficult to quantify. Based on market experience, one cannot just do a standard statistical analysis of available data and hand results in to the supervisory authorities. It is necessary to add extra information. We call extra information “prior knowledge.” While this term originates from Bayesian statistics, we use it in a broader sense: prior knowledge is any extra information that helps in increasing the accuracy of estimating the statistical distribution of internal data. In this article, we consider three types of prior knowledge, and one is the prior knowledge coming from external data. First, one has data that originate from other companies or even other branches (for instance, insurance companies have so few observations that they need to borrow information from banks) that are related to the internal operational risk data of interest—but only related; they are not internal data and should not be treated as such. Second, one can get a good idea of the underreporting pattern by asking experts a set of relevant questions that in the end make it possible to quantify the underreporting pattern. Third, we take advantage of prior knowledge of relevant parametric models, and it is clear that if you have
some good reason to assume some particular parametric model of your operational risk statistical distribution, then this makes the estimation of this distribution a lot easier.

Prior knowledge on a parametric assumption can be built up from experience with fitting parametric models to many related data sets, and in the end one can get some confidence that one distribution seems more likely to fit than another distribution.

For our studies, we have so far come up with the prior knowledge that often the generalized Champernowne distribution is a good candidate for operational risk modeling. This distribution was initially used to model income, so it adapts well to situations where the majority of observations correspond to small values, whereas a few observations correspond to extremely high values (Buch-Larsen et al., 2005). We also consider the art of using this prior knowledge exactly to the extent it deserves. This is, when we have very little internal data at our disposal, we have to rely heavily on prior information.

Our approach based on kernel smoothing has the appealing property that small data sets imply that a lot of smoothing is necessary when estimating the statistical probability density function, and this again implies that we end up not differing too much from our prior knowledge. However, when data becomes more abundant, less smoothing is optimal, which implies that we have sufficient data to reveal the individual distributional properties of our internal data downplaying the importance of prior knowledge. Another key result based on real data is that our transformation approach also has a robustifying effect when prior knowledge is uncertain (Bolancé et al., 2012). For example, in a real data study where three very different parametric models were used, very different conclusions of the quantification of operational risk were given. However, after our transformation approach, similar results were obtained, independent or almost independent of the parametric assumption used as prior knowledge. Therefore, our recommendation is to “Find as much prior knowledge as possible and then adjust this prior knowledge with the real observations according to how many real observations are available”.

The structure of this article is as follows Section 2 presents briefly the semiparametric model approach for modeling operational risk severities. Sections 3 and 4 present the method to combine internal and external data and a procedure to take into account underreporting, i.e. the fact that some small losses may not have been recorded and so the density is somehow underestimated in the intervals corresponding to small losses. Section 5 presents an application and the results are discusses. We present this illustration in order to show the steps towards quantification of operational risk under the premises of data scarcity, presence of external data sources and underreporting. A final section concludes.

2 Semiparametric Models for Operational Risk Severities

Traditional methods for studying severities or loss distributions use parametric models. Two of the most popular shapes are based on the lognormal distribution and the Pareto distribution. Traditionally, a parametric fit can easily be obtained, but this is not the only possible approach. In fact, the main reason why the parametric approach can be controversial is that a certain distribution is imposed on the data. Too often this distribution is the Gaussian model or the log-
normal distribution, but those two models have very thin tails, which means that extreme events hardly ever happen. This assumption leads to a conservative risk measurement, which is sometimes unrealistic and dangerous.

Alternatively, nonparametric smoothing is suitable for modeling operational loss data because it allows for more flexible forms for the density of the random variable that generates the information. The versatility of a nonparametric approach is convenient in our context because the shape of the density can be very different for small and large losses. Our approach is based on a slightly adjusted version of the semiparametric transformed kernel density estimation method (Bolancé et al., 2003).

We show that the estimation is easy to implement (including the often complicated question of choosing the amount of smoothing) and that the new method provides a good and smooth estimate of the density in the upper tail. Parametric models have often been justified due to their simplicity and their ability to solve the problem of lack of smoothness of the empirical approach, so we provide the practitioner with a nonparametric alternative to traditional parametric estimation techniques.

Operational risk analysts are interested in having good estimates of all the values in the domain range: small losses because they are very frequent, medium losses causing a dramatic increase of expenses (demanding liquidity), and large losses that may imply bankruptcy. We study all severity sizes at the same time, and we do not break the domain into separate intervals. Actually, defining subintervals of the domain and fitting a distribution in every subinterval is an obvious approach when a unique parametric shape cannot be fitted across the whole domain range.

In finance and non-life-insurance, estimation of severity distributions is a fundamental part of the business. In most situations, losses are small, and extreme losses are rarely observed, but the size of extreme losses can have a substantial influence on the profit of the company. Standard statistical methodology, such as integrated error and likelihood, does not weigh small and big losses differently in the evaluation of an estimator. Therefore, these evaluation methods do not emphasize an important part of the error: the error in the tail.

As we have mentioned, practitioners often decide to analyze large and small losses separately because no single, classical parametric model fits all claim sizes. This approach leaves some important challenges: choosing the appropriate parametric model, identifying the best way of estimating the parameters, and determining the threshold level between large and small losses. In this section we present a systematic approach to the estimation of loss distributions that is suitable for heavy-tailed situations.

Our approach is based on the estimation of the density of the severity random variable using a kernel estimation method. In order to capture the heavy-tailed behavior of the distribution, we propose to transform the data first and then to go back to the original scale using the reverse transformation. This process is called transformation kernel density estimation. For the transformation, we use the generalized Champernowne distribution function, which has three parameters and adapts very well to the presence of small and large losses simultaneously. The combination of a nonparametric estimation with a previous parametric transformation of the data is the reason why we use the term semiparametric.
Our version of the semiparametric transformation approach to kernel smoothing is very straightforward and can easily be implemented in practice (see, Bolancé et al. 2012, for a guided example chapter with programmes and routines ready for implementation). Classical kernel estimation is substantially improved if data are transformed as a preliminary step. The method is called semiparametric because we use a parametric transformation function and then nonparametric kernel estimation. We introduce a transformation method with a parametric cumulative distribution function, but it could also be implemented with other types of functions, and we advocate the semiparametric transformation method because it behaves very well when it comes to estimating severity distributions.

Loss distributions, that is, densities associated to the severity of operational risk events, have typically one mode for the low loss values and then a long heavy tail. Existing results based on simulation studies have shown that the transformation kernel estimation method is able to estimate all three possible kinds of tails, namely, the Fréchet type, the Weibull type, and the Gumbel type (Embrechts et al. 2006). This makes the semiparametric transformation approach extremely useful for operational risk analysis. A specific feature of the semiparametric transformation principle is that it is also able to estimate the risk of a heavy-tailed distribution beyond the data. The reason is that extrapolation beyond the maximum observed value is straightforward.

Let us assume that a sample of n independent and identically distributed observations $X_1,...,X_n$ is available, they correspond to random variable $X$. We also assume that a transformation function $T(\cdot)$ is selected, and then the data can be transformed so that $Y_i = T(X_i)$ for $i = 1, ..., n$. We denote the transformed sample by $Y_1,...,Y_n$ and the corresponding random variable is $Y$. The first step consists of transforming the data set with a function $T(\cdot)$ and afterwards estimating the density of the transformed data set using the classical kernel density estimator. The density of transformed random variable at point $y$ is:

$$\hat{f}(y) = \frac{1}{bn} \sum_{i=1}^{n} K\left(\frac{X_i - Y_i}{b}\right),$$

where $K(\cdot)$ is the kernel function, $b$ is the bandwidth, and $Y_i$ for $i = 1, ..., n$ is the transformed data set. The estimator of the original density is obtained by a back-transformation to the original scale, using the inverse of $T(\cdot)$. See more details in Bolancé et al. (2005 and 2012)

The choice of bandwidth is an essential point when using nonparametric smoothing techniques. The bandwidth itself could be thought of as a scaling factor that determines the width of the kernel function, and thereby it controls how wide a probability mass is provided in the neighborhood of each data point.

In combination with the transformation, the bandwidth excerpts a very different effect in the original scale. For instance, a small bandwidth that is applied to a transformation that has compressed the original data quite considerably, would clearly cover a wide interval in the original situation (untransformed scale). Conversely, a large bandwidth in a transformation that has widened a subdomain in the original scale may be equivalent to a small bandwidth that is used in the original scale. So, a constant bandwidth applied after the transformation to obtain a kernel estimate has the same effect as implementing a variable bandwidth in the original scale with no transformation.
3 The Mixing Transformation Technique

For simplicity, we introduce superindex I to indicate an internal data source, and we use superindex E to denote an external source. The transformed data set will have a density equal to the true internal density divided by the estimated external density distribution transformed such that the support is on [0,1]. When the external data distribution is a close approximation to the internal data distribution, we have simplified the problem to estimating something close to a uniform distribution. When this estimation problem has been solved through kernel smoothing, we backtransform to the original axis and find our final mixing estimator.

Let \( X_{I1}, X_{I2}, \ldots, X_{In} \) be a sequence of \( n^I \) collected internal losses, and let \( X_{E1}, X_{E2}, \ldots, X_{En} \) be a sequence of \( n^E \) external reported losses. Then external reported losses represent prior knowledge that should enrich a limited internal data set. We assume that scales are comparable and a filtration has been made, so that the external information has already been prepared to be directly mixed with the internal data.

The internal data sample \( X_{I1}, X_{I2}, \ldots, X_{In} \) is then transformed with the estimated external cumulative distribution function. The parametric density estimator based on the external data set offers prior knowledge from the general financial industry. The correction of this prior knowledge is based on the transformation density estimation of the internal data points that is used as explained in section 2. This leads to a consistent and fully nonparametric estimator of the density of the internal data.

4 The Underreporting Function and Underreporting Correction

When modeling a publicly reported database, one needs to fit the operational risk exposure to a company’s internal situation. This is done by introducing a function into the model, which could be thought of as a scaling function explaining the likelihood of a loss being reported at different threshold levels.

A very simple interpretation of how the function affects the distribution is that, for small and medium-sized losses, the density up-weighted to include more probability mass in that domain, while larger losses are down-weighted. This can be called an underreporting function.

The underreporting function should, as well as being continuous, pass exactly through the predetermined reporting level values. The requirement is mathematically termed as interpolating the report levels, which is a prediction made through and between the report levels. A further requirement is to have a continuous and well-defined function after the last observed report level. Mathematically, this is defined as extrapolation and is a prediction made beyond the last observed report level. Of course, a prediction based upon extrapolation is potentially less reliable than a prediction between report levels since one must assume that a historical pattern
will continue outside the reported value domain. Figure 1 shows an example of underreporting function.

![Figure 1 Example of underreporting function for several operational risk categories (Guillén et al., 2007).](image)

Figure 1 presents the estimated probability of reporting for each risk category. It can be seen that the event risk category number seven (which corresponds to Execution, Delivery, and Process Management) has the lowest reporting likelihood, which means that losses for this kind of operational risk are likely to be underreported. On the other hand, smaller losses that belong to category number five (which corresponds to Clients, Products, and Systems Failure) have a much higher probability of being reported than small losses for other types of risks. It is also noteworthy that, for all event risk categories, expert opinion assesses that the probability of reporting a loss is smaller or equal than 99%, even for very large losses. In fact, this acts as a bound, and it implies that there is at least a 1% chance that a very large loss occurs and is not reported.

The estimation of the underreporting function is essential in our approach, but unfortunately, very little is known about the magnitude of underreporting and even less about the validity of experts’ estimation. We think that further research needs to be done on the stability of these estimates, and we believe that here we provide a starting point for discussion.

| Table 1. Event type definition for operational losses |
|-----------------|--------------------------|
| 1 Internal Fraud | 2 External Fraud          |
| 3 Employment Practices and Workplace Safety | 4 Business Disruption |
| 5 Clients, Products, and System Failures | 6 Damage to Physical Assets |
| 7 Execution, Delivery, and Process Management |
Underreporting is corrected using the definition of the probability density of the statistical random variable and re-scaling it in order to take into account that sample is not fully representative of the true losses. Underreporting correction is equivalent to weighting the data in order that the sample correctly represents the original distribution. This technique is well-known and it is widely used in statistical sampling theory. Each observation is weighted with the inverse of the probability of selection. In this case, the probability of selection is given by the underreporting function.

5 An operational risk application combining mixing and underreporting

We must emphasize that distributional assumptions can be modified in practice, and the ones assumed here are not necessarily the best choices for every applied situation.

A company should combine internal and external data for internal loss distribution modeling, when the sample of internal data seems insufficient to represent all the domain accurately. Here we take into account both minimum collection threshold as well as underreporting. The mixing model should utilize the external data as prior knowledge. Then, a nonparametric smoothing technique adjustment should be applied on the prior knowledge according to internal data availability.

In our example, the mixing model is applied as a first step. In the second step, the underreporting correction is introduced. We assume that the same underreporting applied both for the internal and the external data.

We use two data sets described in Table 2. The two data sets correspond to operational risk data from internal and external sources. These data have been generated for illustrative purposes from a particular company and type of operational risk. We will call these data the internal data set and the external data set, respectively.

<table>
<thead>
<tr>
<th>Name</th>
<th>Content of operational risk loss data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal data set</td>
<td>75 observed loss amounts</td>
</tr>
<tr>
<td>External data set</td>
<td>700 observed loss amounts</td>
</tr>
</tbody>
</table>

In Table 3 we show the basic descriptive statistics of our two data sets. Results in Table 3 show that the size of losses is very different in the internal sample and the external sample. The reason for the different scale is that data from the external sample have been obtained from companies that may be larger than the company from which internal data were obtained. Therefore with mixing transformation techniques we also provide a rescaling procedure that would solve the problem of combining both data sets.
We estimate four models for the density function of the losses associated with operational risk, named M1, M2, M3 and M4. Model M1 denotes the semiparametric model described in section 2 and will be used for the internal data set. Models M2 and M3 are mixing models. In M2 we use transformation parameters estimated with external data to obtain transformed kernel density estimation with internal data. The rescaling procedure is introduced in model M3. This model is similar to M2, but in M3 we use shape transformation parameters estimated with external data and the scale transformation parameter estimated with internal data. Finally, model M4 is the benchmark model with a Weibull assumption on the severity estimated with internal data.

From the severity density function, we estimate the distribution of total operational risk. We assume that the number of losses over a fixed time period follows a Poisson process. We know that the annual number of internal losses has mean equal 30, then we can assume that the annual number of internal losses follows a Poisson process with maximum likelihood estimated parameter $\hat{\lambda} = 30$. We denote the annual simulated frequencies by $\hat{\lambda}_r$ with $r = 1, \ldots, R$ and number of simulations $R = 10,000$. For each $\hat{\lambda}_r$ we draw random uniformly distributed samples and combine these with loss sizes taken from the inverse estimated cumulative distribution function of the severity distribution. The outcome is the annual total loss distribution denoted by

$$S_r = \sum_{k=1}^{\hat{\lambda}_r} \hat{F}^*(u_{rk}, \theta), r = 1, \ldots, R$$

with $u_{rk} \in U(0,1)$ for $k = 1, \ldots, \hat{\lambda}_r$, and $\hat{F}^*(u_{rk}, \theta)$ is the inverse of the cumulative distribution function obtained from the density that has been estimated from one of the models, i.e. M1, M2, M3 and M4.

Table 4 presents summary statistics for the simulated $S_r$ total loss distribution for each model. Among the usual summary statistics we report VaR and TVaR for risk tolerance $\alpha = \{0.95, 0.99, 0.995\}$. Results for models M1, M2, M3 and M4 reveals that when we include external information (models M2 and M3) the estimate for the operational risk measure is higher.

In Table 5 we show the same results including underreporting. If we compare results in Table 4 to those in Table 5, we can see that results without underreporting subestimate risk.
Table 4. Statistics of simulated total loss distribution associated with operational risk

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>Median</th>
<th>Standard dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>4.3309</td>
<td>4.22</td>
<td>1.5416</td>
</tr>
<tr>
<td>M2</td>
<td>4.5685</td>
<td>4.42</td>
<td>2.0063</td>
</tr>
<tr>
<td>M3</td>
<td>4.8237</td>
<td>4.65</td>
<td>2.3677</td>
</tr>
<tr>
<td>M4</td>
<td>5.2042</td>
<td>5.08</td>
<td>2.0720</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Var_{95%}</th>
<th>Var_{99%}</th>
<th>Var_{99.5%}</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>6.53</td>
<td>7.65</td>
<td>8.84</td>
</tr>
<tr>
<td>M2</td>
<td>7.11</td>
<td>8.38</td>
<td>9.89</td>
</tr>
<tr>
<td>M3</td>
<td>7.60</td>
<td>8.93</td>
<td>10.68</td>
</tr>
<tr>
<td>M4</td>
<td>7.76</td>
<td>8.99</td>
<td>10.37</td>
</tr>
</tbody>
</table>

Source: Own and Bolancé et al. (2012)

Table 5. Statistics of simulated total loss distribution associated with operational risk including underreporting

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>Median</th>
<th>Standard dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>11.0148</td>
<td>10.46</td>
<td>11.6227</td>
</tr>
<tr>
<td>M2</td>
<td>13.3456</td>
<td>11.73</td>
<td>51.5824</td>
</tr>
<tr>
<td>M3</td>
<td>14.5299</td>
<td>12.75</td>
<td>62.2843</td>
</tr>
<tr>
<td>M4</td>
<td>12.2733</td>
<td>12.17</td>
<td>4.0734</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Var_{95%}</th>
<th>Var_{99%}</th>
<th>Var_{99.5%}</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>15.74</td>
<td>22.36</td>
<td>43.41</td>
</tr>
<tr>
<td>M2</td>
<td>23.04</td>
<td>45.08</td>
<td>90.82</td>
</tr>
<tr>
<td>M3</td>
<td>25.65</td>
<td>50.34</td>
<td>94.11</td>
</tr>
<tr>
<td>M4</td>
<td>15.74</td>
<td>17.45</td>
<td>19.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>TVar_{95%}</th>
<th>TVar_{99%}</th>
<th>TVar_{99.5%}</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>20.49</td>
<td>31.74</td>
<td>69.37</td>
</tr>
<tr>
<td>M2</td>
<td>36.63</td>
<td>64.32</td>
<td>109.69</td>
</tr>
<tr>
<td>M3</td>
<td>40.38</td>
<td>70.01</td>
<td>113.18</td>
</tr>
<tr>
<td>M4</td>
<td>16.84</td>
<td>18.55</td>
<td>24.04</td>
</tr>
</tbody>
</table>

Source: Own and Bolancé et al. (2012)

6 Discussion

Our recommendation is to pay attention to underreporting whereas model selection for combining internal and external data is especially relevant only when the tolerance level is high, which is when the quantile for assessing risk if high.

Our application shows that failure to account for underreporting may lead to a substantial underestimation of operational risk measures. The use of external data information can easily be incorporated in our modeling approach. So, even if the underlying theoretical foundation is technically difficult, the intuition and the practical implementation of operational risk models is very straightforward.
Our method addresses the statistical problem of estimating a density function with extreme values and then how to combine data sources, while accounting for operational risk underreporting. Underreporting, which is a feature that practitioners admit is the most dangerous problem of operational risk data quality, has generally been much ignored in the specialized literature.

References