

# A practical approach to model banking risks using Loss Distribution Approach (LDA) in Basel II framework

Raquel Barreira, Tristan Pryer and Qi Tang

Department of Mathematics

Sussex University

Mantell Building

Falmer

Brighton BN1 9RF

Contact e-mail: [q.tang@sussex.ac.uk](mailto:q.tang@sussex.ac.uk), Tel: +44 1273 877457

**Abstract:** In Basel II Capital Accord, the Advanced Measurement Approaches (AMA) is stated as one of the pillar stone methods for calculating corporate risk reserves. One of the common yet cumbersome methods is the one known as loss distribution approach (cf. [3]). In this article, we present an easy to implement scheme through electronic means and discuss some of the mathematical problems we encountered in the process together with proposed solution methods and further sought on the issues.

## 1 Introduction

In Basel II Capital Accord, the use of top-down or bottom-up method to calculate risks provisions are recommended as the ways to model and compute corporate risk value (VaR -- values at risk). The top down approach relies on general business income/cash flow side of the business, for example, a percentage of the total net cash managed. The bottom up approach relies on gathering loss data from within the company and carry out statistical analysis of these data to arrive at a figure. Bottom-up methods are regarded as more refined. Once bottom-up methods are adopted, it is no longer permitted to retreat to top-down approach.

Within the bottom-up approach, there are process based models, actuarial models and proprietary models.

The process based model splits banking activities into simple business steps, the management evaluates the situation according to these steps to identify risks. This is mainly a time series type of model. Regression analysis tools are often used when there are multi-factors in the problem (cf. [1], [2], [5], [6] and [7]).

The actuarial models or statistics models are generally parametric statistical models. Various statistical fitting techniques are used (see extensive discussions in [3]). In this article, we present an efficient, direct way for this approach and we also discuss some of the technical difficulties that need to be solved. The implementation of this method is based on a carefully designed algorithm. In places where mathematics computations encounter difficulty, management judgement is requested in the form of inputting parameters in programmed interfaces. We also set default should management is unwilling to make judgments. These default settings are carefully set with discussion with management before implementation.

The advantage of our actuarial model is that once it is set, the model itself will give results very close to historic expected total loss. It is also possible to carry out extensive Monte-Carlo disturbance to the multi-parameter model on various levels of the organization and simulate a complicated business operation. It can also incorporate features such as management control impact on reduction of losses. We only discuss the general philosophy of algorithm design but not the details of how to implement various technical control issues. Our final program operates in the world-wide-web environment. A free test version can found by opening a trial account through [www.care-web.co.uk](http://www.care-web.co.uk). The background programming is in C.

The proprietary models in risk management are mainly developed by major finance service companies. The approach involves a variety of bottom-up, top-down and qualitative analysis schemes. It is mainly spreadsheets based. It was mentioned in [3] that the currently available proprietary software include Algo OpVantage by Algorithmics Inc, Six Sigma by Citigroup and GE Capital, and Horizon by JP Morgan Chase and Ernest & Young. Interested readers should go to the internet search engines to obtain more information.

Our model is different from the existing ones as we incorporated a much more flexible adjustment mechanism and the aims of the prediction are very concentrated. In the following, we will discuss step by step the various features of the program and the mathematical thinking behind. We are required by a management consultancy firm to implement these details. We believe that although it makes the running of the program more technical (there is a large preparation data file to prepare), but it does have more features and give more flexibility in modelling and predicting the VaR.

The content of this paper is as follows:

- 1) Description of the modelling approach and algorithm (see Section 2)
- 2) How to give different weight to different year's data (see Section 3).
- 3) How to deal with insufficient number observations (see Section 4).
- 4) How to deal with near misses (see Section 5).
- 5) Incorporating [excluding] high impact, lower probability events when they are not in [already in] the data table (see Section 6).
- 6) Management intervention and cost/improvement comparison (see Section 7)
- 7) Problems arising from designing Monte-Carlo simulations (see Section 8).

## **2 The modelling approach**

In order to describe the loss profile of a corporate entity, it is important to collect the past data from the company. If the company has accumulated enough data over the past three to five years, then we can use the method defined in this section to model the company risk. The method falls into the general category of LDA but without the parts of estimating parameters and fitting to an existing distribution function, the model is based on direct modelling of the existing loss data.

The reasons for using direct modelling rather than fit onto an existing probability distribution are as follows:

- 1) Computational techniques make the handling of thousands of data automatic and instant, it can also highlight many critical issues automatically (such as few observations among some risk classes, very high loss values in a risk class etc) to alert the management in a mechanical manner.
- 2) The estimate of the loss value and frequency can now be done by Monte-Carlo simulation. When hundreds or thousands of data are involved, the Monte-Carlo simulation of the business operation becomes much more closer to real life situations.

In this article, we discuss LDA and use frequency as vertical axis against actual loss as horizontal axis. Time is only used as a background information (see Section 3). This is based on the observation that a detailed time series analysis of the risk events may not be very useful in this particular setting because businesses run in cycles. Accounting periods and many management factors could have intervened in the reporting system that result in inaccuracies in terms of report timing of the risk events. Many operational risks have a build up period and risk releasing period. This is one of the fundamental differences between business risk measurement and equity / derivative price movement measurement. By assuming that management is pro-active, we have also involved the feature that gives differentiation to the importance of past data depending on how far the event has happened from current time. For example, three-year-old data will be given less weight compared to one-year-old data.

Our discussions are based on real industrial consultancy experience, the modelling problems / challenges mentioned are real world situations. In each subsequent section, we concentrate on one particular issue, we give the background information on why these issues are arising, what is the expectation of the company management and what we can do to build a robust mathematical model.

Before going into further discussion, we first introduce the notations and basic concepts.

It is a common agreement that the risks faced by a company should be classified according to the hierarchy of



At the lowest level, it is common now to give a label to each individual risk event (just an example) in the way of

110	corresponding to	“Default on payment”
320	corresponding to	“Falsified identity”

etc. To simplify the discussion, these events are regarded as probabilistically independent events for the business.

Of course, they can also be treated as dependent events, then the aggregation of different probability distribution functions will be different from what we present now. In Section 8, we will highlight some of the possibilities. To keep things simple for management consultancy companies, we avoided using too complicated modelling techniques.

The business departments within a company will be directly responsible to a number of risk events. It is a common practice not to let different departments to share same risk events. So the structure looks like

Department 1	Department 2	••• •••
Risk 110	Risk 210	••• •••
Risk 120	Risk 220	••• •••
••• •••	••• •••	

It is common that risk are analysed per risk event and aggregated back to departmental and company level. The labelling of the risks provides a convenient way for the assessment and analysis using internet forms and programmed algorithm.

## 2.1 The data collection system

A web based reporting system installed across the business's offices, departments should provide a reporting form as

Risk No	Date reported	Loss value
110	02/02/2003	312,227.71
210	03/07/2003	0
300	07/10/2003	536.24

Table 1.1: Example of risk report form from a web reporting system.

The reported event with loss value “0” in the report form is called a “near-miss” event, where an event has happened against the interest of the company and has been observed, but there is no immediate observable money loss. However, it cannot be excluded that hidden loss has happened or will happen.

The meanings of the quantities involved in the table are as follows:

- Risk No – an identity number given to the risk event.
- Time reported – the time when the report of the loss is made OR the actual date when this event happened.
- Loss value – the actual recorded value of loss – this may be inconclusive as hidden losses may appear later.

It is assumed that each accident is reported in this format. We use “time reported” rather than “time happened” for convenience. Considering the annual business (accounting and management checking) period, this is a reasonable assumption.

Of course, real life risk management requires more information regarding the nature and quantification of risk events. In the actual simulation, additional information will need to be provided before the data have been fed. They are

- Total number of risks-risk numbers: they do not have to be pre-assigned, they can be any number string. But the total number of different risk numbers have to be known so that the data loading process can be carried out without errors.

- Number of years in table: this will tell the program the prediction is based on the combination of how many years of past data.
- Number of groups, their labels and the risk numbers they contain: This will tell the program that the simulation is done on how many groups, each risk number is allowed to belong to only one group.
- Initial year of data: the program will understand how old the data are.
- Confidence level: the confidence level of predicted loss values, usually 95%.
- Loss impact and probability of occurrence for each risk number: these are management perceived quantities. Impact and probability are classified as V (very high), H (high), M (medium), L (low) and E (extremely low).
- Loss range: for each particular risk number  $j$ , a loss range is given to impact (not probability) factor V, H, M, L and E. These can be different for different risk numbers. For example, for risk 1289, its impact is M, its loss range could be (100000,300000), but for risk 22369, it is also M, but its loss range could be (1000,2000).
- Cash related indicator: some losses are not cash related, but could have implicit impacts. Numerical simulation can include them by assigning cash value or exclude them according to the indicator = 1 or 0.
- Management factor: This is a factor between 0 and  $\infty$ , but usually between 0 and 1. This is again a management judgement. If certain management measures are expected to reduce (or exacerbate) the loss, then this factor can be taken  $<1$  ( $>1$ ) accordingly. This is risk number dependent, default value = 1.
- Final information is annual weight: it is customary in management to regard older data as less relevant. We give the flexibility for management to assign different weight to data from different years. These weights are also risk number dependent.

When numerical simulation is finished, we have a number of information given in an exit report (issued after simulation) in parallel to the numerical results

- 1) If actual loss has happened and the real loss is out-of range for that risk number significantly (ie, risk is regarded as H, and range is (100000, 1000000), but there is a recorded loss event with value 2000000), then the management perception of loss range may need modification, our program will send a warning to the exit report.
- 2) Long tail risk: if V or H impact, L or E probability event has no recorded loss in data table, a mock loss event will automatically be added, with a message in the exit report on the nature and magnitude of the add event.
- 3) If a risk number contains only near misses, a warning will be given in exit report.
- 4) If a risk number contains high frequency loss event (a single loss value recorded many times repeatedly), a warning will be issued in exit report.

## 2.2 Construction of probability distribution frequency/loss value

Local probability distribution for a single observed cash loss value: Based on the reporting system described above, we can accumulate and sort all data for Risk No 10, for example, we list all loss data in increasing order (here time factor has been ignored, it could also be ordered according to the time when it happens), if the loss observation  $x_j$  has occurred  $p_j$  ( $\geq 1$ )

times, the associated frequency will be  $p_j$ . The loss observations will be in the following format

Observations for risk No k		
Date reported	Loss sum	Frequency
02/2007	$x_{k1}$	$p_{k1}$
...	...	...
01/2004	$x_{ks}$	$p_{ks}$

Remark: The index can be arranged in any convenient order.

We can then construct a local probability distribution around each loss value, say

$$g_s(\mu_s, \sigma_s) \text{ at } (x_s, p_s)$$

where  $g_s$  is a probability density function,  $\mu_s$  is the “assigned” expected loss and  $\sigma_s$  is the “assigned” spread of the loss (risk of the loss). The strategy for computing these parameters should be by using neighbouring loss data.

For example, if loss  $x_s$  is one of the actual loss observations, to construct the probability distribution  $g_s(\mu_s, \sigma_s)$ , we can use loss values

$$x_{s-p} < x_{s-p+1} < \dots < x_s < \dots < x_{s+q}$$

together with corresponding frequencies

$$p_{s-r}, p_{s-r+1}, \dots, p_s, \dots, p_{s+q}$$

to calculate the expected loss and risk (spread):

$$\mu_s = (x_{s-r} p_{s-r} + x_{s-r+1} p_{s-r+1} + \dots + x_{s+q} p_{s+q}) / (p_{s-r} + \dots + p_{s+q})$$

$$\sigma_s^2 = ((x_{s-r} - \mu_s)^2 p_{s-r} + (x_{s-r+1} - \mu_s)^2 p_{s-r+1} + \dots + (x_{s+q} - \mu_s)^2 p_{s+q}) / (p_{s-r} + \dots + p_{s+q} - 1)$$

here loss observations are from the same risk number.

Note that there are some cases where left data or right data are missing. For example,  $x_1$ , the smallest observed loss value has no left loss values. In this case, minor adjustment is carried out using any reasonable method, this will usually not affect the overall risk profile.

For risk numbers containing single observation, we will have to use the impact bounds to help us to decide expected loss and risk.

In a similar way, we may construct higher order moments. Once we have these moments, we can construct a local probability distribution for the loss value  $x_s$ .

**Remark:**

- 1) The local distribution function can be normal, gamma or any other appropriate function. For example, if normal distribution is used, truncation at 0 (loss values are all regarded as positive) and re-weighting is necessary.
- 2) The assigned expected loss and assigned spread of the loss will depend on the neighbouring loss values as shown above.
- 3) This requires that the parameters are calculated “locally”, eliminating the need for estimating the parameters of the overall probability distribution.
- 4) The “localization method” can give large rare loss a large associated spread (risk), therefore smoothing the local probability distribution at that point. It will also

concentrate sharply where large number of losses are observed within a relatively compact range.

Example: In the case of normal distributions, the local distribution at  $(x_s, p_s)$  is

$$f_s = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{(x-\mu_s)^2}{2\sigma_s^2}\right) / \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{(x-\mu_s)^2}{2\sigma_s^2}\right) dx & x > 0 \\ 0 & x < 0 \end{cases}$$

Probability distribution for a risk number: After constructing the local probability distribution for losses in each risk number, the global probability loss distribution for a given risk number can then be constructed using weighted sum, the weight is the frequency of  $s^{th}$  loss value (the number of times it appeared) over the total frequency in that risk number. The aggregation formula is

$$f_{risk\ i} = \frac{\sum_s p_s f_s}{\sum_s p_s}$$

where  $s$  is summed over all observations in risk number  $i$ . To simplify notation when no confusion we simply re-label  $f_{risk\ i}$  as  $f_i$ .

**Remark:** The fact that the actual loss value is a random quantity is compensated by the recognition that its frequency is a random variable. It is easy to check that if the loss distribution thus constructed for risk number  $i$  is  $f_i$ , we have

$$\text{Expectance of } [f_i] \approx \text{Expectance of frequency} \times \text{Expectance of Loss}$$

Here we do not have equality because of truncation and re-weighting errors. These errors are usually very small. This conforms to the insurance risk modelling principle.

The probability distribution for the entire company or a group of risk numbers: events in different risk numbers are treated as independent, we simply aggregate the distribution functions for risk numbers

$$f = \frac{\sum_i (\text{total frequency for risk } i) f_i}{\sum_i (\text{total frequency for risk } i)}$$

The index  $i$  is taken for all risk events in a designated simulation group. The resulting probability distribution gives the basis of modelling of company risk.

**Remark:** In fact, these “total frequency” may need to be changed from true figures. We will gradually discuss cases where frequencies need to be re-assigned (for near-misses, for older data etc).

Finally, random perturbations can be given to each loss value and its frequency. The perturbed model will exhibit rather complicated behaviour and resembles to a real business operation. The perturbation pattern can be decided by the company need. After many simulations, an estimated loss interval can be extracted using certain confidence level.

### 3. Give different weights to different years of data

**Background:** Assuming that the management system is reasonably efficient, gradual improvements will be in place for controlling high / medium impact losses. Considering the delay in time in the management process, it is anticipated that j-year-old data will have less impact on current operation than (j-1)-year-old data for any j.

**Considerations:** The weight for different years should apply to frequency only. The reason is that the loss observations are used in calculating expected loss and risk (spread) of the distribution functions, therefore the actual loss observations will affect the distribution pattern of the losses and should not be altered lightly. The application on frequency will give a desirable rebalance of importance and is viewed as acceptable within our discussions.

#### Solutions:

1) Management controlled re-weighting: The re-weighting factor are imported from the data table, we use five-year as an example in the following:

Year from current time	1	2	3	4	5
Re-weighting factor	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$

Table 2.1: Re-weighting factor.

where  $A_j \geq 0$  for  $j=1,2,3,4,5$ , except for exceptional circumstances, we should have  $A_1 \geq A_2 \geq A_3 \geq A_4 \geq A_5$ .

**Remark:** It is OK for some of the  $A_j$ s to be bigger than 1, for example, the choice  $(A_1, A_2, A_3, A_4, A_5) = (1.4, 1.0, 0.7, 0.4, 0.2)$  is perfectly OK. Let  $a = 1.4 + 1 + 0.7 + 0.4 + 0.2 = 3.7$ , the real weight thus assigned are:  $(1.4/3.7, 1.0/3.7, 0.7/3.7, 0.4/3.7, 0.2/3.7)$

In the data table (see beginning of Section 2.2)

Date reported	Loss sum	Frequency
02/2007	$x_{k1}$	$p_{k1}$
...	...	...
01/2004	$x_{ks}$	$p_{ks}$

the program can apply suitable factors in the frequency column. Say we are modelling year 2008, hence 2007 is the closest year, after applying annual weight, the data table is reformulated to

Date reported	Loss sum	Frequency
02/2007	$x_{k1}$	$A_1 p_{k1}$
...	...	...
01/2004	$x_{ks}$	$A_4 p_{ks}$

#### 2) Regression re-weighting approach (when there is considerable history in reliable data recording)

We can use the best fitting AR(5) model (autoregressive model)

$$y_t = A_1 y_{t-1} + \dots + A_5 y_{t-5} + \varepsilon_t$$



to decide the coefficients  $A_j$ . The “total number of losses” (taking into account of near-misses and non-cash losses to a certain degree) for each risk number is used as one observation, regression is carried out over all risk numbers.

Note that  $t$  is the current year where the total loss is supposed to be known. The final probability distribution, using previous notations is

$$f = \frac{\sum_{j=1}^5 p_j A_j f_j}{\sum_{j=1}^5 p_j A_j}.$$

Since the management controlled re-weighting is more straightforward to understand, we use it as our default setting. Whether to choose management controlled re-weighting or regression re-weighting should be a simple matter of preference. However, in implementing regression re-weighting, a safeguard has to be implemented to invalidate the model when  $\sum_{j=1}^5 p_j A_j = 0$  (unlikely but possible) and switch back to approaches 1).

#### 4 Insufficient number of observations

**Background:** because of subdivision of risk into categories such as business lines and risk types, it is often that in the construction of the probability loss function, we end up having just a few observations or no observations at all. Recall that to construct loss distribution function, we need neighbouring data to define local data spread (risk) and expected loss, few neighbouring data means that the reliability of analysis is reduced.

For example, if 4 neighbouring (different) observed loss values are need to calculate expected losses and data spread (risk), and there are only three different loss values observed in that category, then we have to compromise the way we compute these quantities.

**Considerations:** Any reasonable way of calculating expected loss, data spread (risk of a loss, the possibility a new loss appear near this old loss and/or how neighbouring losses are related) will require a number of different observed loss values for the calculation/estimate to be carried out.

**Solutions:** Again here, management participation is necessary. It is understood that in the <<Impact>> column of the report form (see Table 1.1 and the bulleted explanation below that table), the indicator V, H, M, L and E have their corresponding “loss range”, this is the management judgement of possible range of loss for events in this particular risk. If the recorded loss falls inside this range, then the end data of this range, some interpolation of them and the actual loss(es) can be used as “additional data” to form a computational strategy. We do not construct local probability distribution for the end data and their interpolation but we use them for computing expected returns and risks.

It has to be pointed out that these “range” boundary values are usually very wide and may not be as precise as desirable, it is often that the management will require some fine tuning of their values based on the actual loss values. It is also possible that the observed loss values

falls significantly outside this range, if this is the case, we still carry on as usual but the program will issue a warning in the exit report that the range has been “broken” for the risk number concerned. Management need to review the situation and make further decisions.

In the reporting table, there is a column containing a statement that the probability for this risk number is regarded by the management as V, H, M, L or E. This should be ignored in our construction of probability distribution. The fact that there are few observations speaks for itself. They are used for other purposes described later.

## 5 Near-misses

**Background:** Near-misses are observed events that have led to no quantifiable immediate losses. Near-misses are actually quite frequently reported in real life. In the institutions we worked with, as many as about 8% of total reports are about near-misses. The near-miss data also spread over many different risk numbers.

**Considerations:** When an actual cash loss is reported, its loss value is very likely to be different from other reported loss values. In the case of near-misses, not only they do not have a real loss value attached to, the frequency is actually large. So we have a double problem to solve:

- 1) How to deal with events that have loss value = 0.
- 2) How to deal with events with large frequencies (say more than 20% of total frequency for that risk number)

### Solutions:

#### 5.1 Loss value = 0

To design a probability loss distribution, we can use a fixed percentage of the expected loss in that risk number category. The actual number should be decided by the management.

Correspondingly, a reduction in the total number of observations must also be applied.

After computation, an exit report is issued stating the likely contributory effect of these near-misses.

#### 5.2 Single loss value with large number of observations

From risk assessment point of view, we should view this with certain suspicion. The interpretation of this phenomenon is that either the fraudsters find this particular value attractive, or in the reporting procedure, the reporter simply added the various losses and took an average. So the loss values should be given more spread than other loss values.

In our approach, if such an event happens, the program will automatically detect it, and apply an adjusting factor  $f(m)$  to the risk (spread) factor, this adjusting factor  $f(m)$  is a function of  $m$ , the total number of observations for this particular loss value in consideration.

That is, if  $\sigma$  is the standard risk (spread) for the probability distribution for this loss, the adjusted risk (spread) is  $f(m)\sigma$ . We used  $f(m) = m^\phi$  for some  $\phi \in (0, 1)$ , a constant to be decided and tested by the management. Of course, in the actual process of simulation, the program will simply go on using our default setting which is to let  $\phi=1/8$ . An explanation has been added in the exit report explaining that such a high concentration event has happened. If the user is not happy with the current simulation result, he/she can ask the programmer to change the value of  $\phi$ .

## 6 Incorporating very high (V) and high (H) impact, low (L) and extremely low (E) probability event

**Background:** It is often true that if a risk number is labelled as having V or H impact (in terms of value of loss) and L or E probability (in terms of appearance frequency), we may find that in the period of data collection, there is no actual observation of such event (notice here that this no-observation must have happened for the entire time period of observations where computation is based upon, say over the whole 3-5 years).

**Considerations:** If this happens, the management may want to add an artificial loss event. With this imagined event in the probability distribution, the subsequent Monte Carlo simulation process will take it into account to give predictions for the company reserve level including a perturbation of this “artificial” data.

**Solutions:** First, a detection mechanism should be in place to warn that such an event has got no observations in the time interval concerned. In our program, if such a no-observation event happened, the program will automatically insert a imagined event with frequency 1 and a magnitude depending on the loss range of the risk number. In the exit file, a report will be issued clarifying the insertion and its likely impact.

## 7 Management intervention and cost/improvement comparison

One of the interesting factors we have introduced in our prediction of risk capital reserve is a management control factor which applied a factor to the existing loss data.

Usually this is a reduction factor in terms of both loss sums and frequency of losses.

If the management is interested in testing a management improvement measure with a cost X. They have a perceived reduction in number of losses.

We have a system of introducing control factor (loss reduction factor) based on risk numbers. That is, risks in the same risk number will be reduced by the same amount in terms of loss amount, frequency and loss spread (risk), but different risk numbers can have different control factors (loss reduction factors). Say for risk number 320, we apply reduction factors  $(\text{amount}_{320}, \text{frequency}_{320}, \text{spread}_{320}) = (a_{320}, f_{320}, s_{320})$ , then take a typical function

$$f_s = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{(x-\mu_s)^2}{2\sigma_s^2}\right) / \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_s} \exp\left(-\frac{(x-\mu_s)^2}{2\sigma_s^2}\right) dx & x > 0 \\ 0 & x < 0 \end{cases}$$

we have

$$f_s = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_s s_{320}} \exp\left(-\frac{(x - \mu_s a_{320})^2}{2\sigma_s^2 s_{320}^2}\right) & x > 0 \\ \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_s s_{320}} \exp\left(-\frac{(x - \mu_s a_{320})^2}{2\sigma_s^2 s_{320}^2}\right) dx & x < 0 \end{cases}.$$

The factor  $f_{320}$  will be applied in the aggregation process in a similar pattern.

This gives the management the maximum flexibility in estimating perceived improvements in implementing perceived management measures. It will also be straightforward to look at the financial impact of improvements in comparison to cost.

To simplify the process, in our program, we unify the three parameters into a single management factor by assuming that  $a_{320} = f_{320} = s_{320}$ .

## 8 Issues arising from Monte-Carlo simulations

In the Monte-Carlo simulations, we perturb the observed loss values and the associated frequencies. We repeat the simulation many times and pick the confidence interval. This approach agrees with the BASEL II requirement that the simulations must be based on models using true company data.

In our program, we have designed a concept called risk groups, this is an intermediate level between risk numbers and the entire company. In the input file, the total number of groups  $N$ , the risk numbers each group contains and the label for each group should be given. The program will automatically model and simulate the  $N$  groups and give out  $N$  simulations results and  $N$  exit files. In management terms, these groups correspond to the certain section of the business (such as departments) that the management is interested in looking at the performance. The only restriction is that each risk number should belong to no more than one group (when a risk number is in no group, it does not participate in modelling and simulation overall).

The aggregation of the model from risk number case up to groups (departments or business lines) or whole company can also be done in different ways, thus affecting the simulation results.

- 1) Simply combine the distribution functions for each risk numbers using weighted sum. The underlying assumption is that different risks are statistically independent events.
- 2) Combine the observation data from different, but relevant risk numbers to form a new data pool, and reconstruct the distribution function. This is assuming that the underlying events are statistically relevant.
- 3) A combination of 1) and 2). Regroup the data into different really independent sets, each set may contain one or more risk numbers, and treat data within each group as dependent, treat data on inter-group level as independent.

We point out that the approach 3) makes the programming much more human intervention dependable, but could have the advantage when coming to simulate management improvement measures. It is often true that a prescribed management measure will affect a number of risks rather than just one risk. This points to interdependence of the risks under that particular context.

To simplify the structure of the input file, our program uses approach 1).

**Remark:** There are many ways of predicting the future capital reserve of the business simulated. We relied on Monte-Carlo simulation because we think our model has incorporated enough details for a meaningful simulation. If we use other statistical prediction methods such as regression, we will not be able to incorporate very fragmented information we have been requested to implement such as

1) Giving different weight to different year's data but different risk numbers may have different weight: Say for Risk No 10, weight are given as follows:

Year 1	Year 2	Year 3	Year 4	Year 5
1	0.9	0.8	0.7	0.6

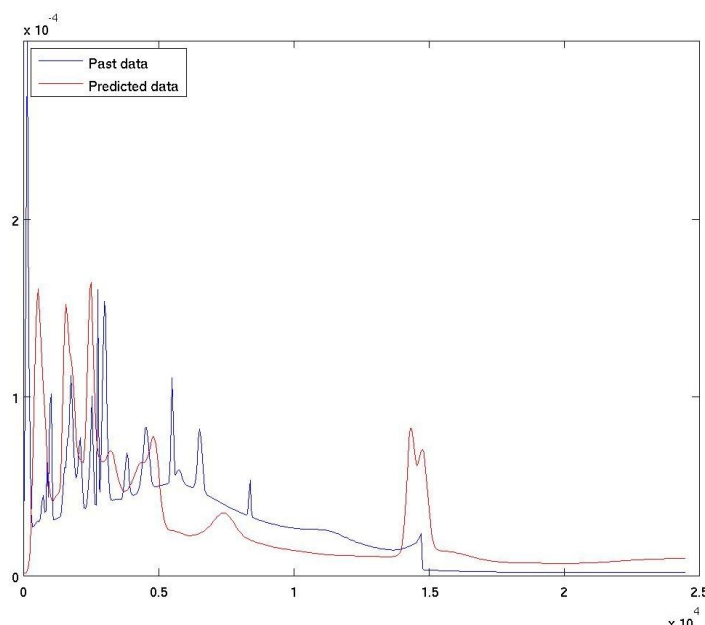
But for Risk No 110, we are required to implement

Year 1	Year 2	Year 3	Year 4	Year 5
1	0.8	0.8	0.75	0.66

2) Applying management factors: if a management measure reduces the impact of Risk No 10 by 20%, it may only reduce the impact of Risk No 110 by 5%.

## 9 Example, exit report and conclusion

The following is the probability distribution of “frequency against loss” constructed using part of the data file received from a company (no name disclosed for confidential reasons) and its 1000 Monte-Carlo simulations averaged. It is not the whole picture as the range goes from zero to positive infinity, there are some other concentration areas further beyond the range we plotted. But the effects are smaller.



In summary, based on a consistent web reporting system, our algorithm has the following advantages:

- 1) It is numerically very efficient to construct a probability distribution and calculating various statistical summaries.

- 2) The Monte-Carlo simulation can highlight distribution anomalies (say the jump around 15K loss value in the demonstrated example), and smooth out large number of small losses (say the peak appeared near 0).
- 3) The form of the probability distribution is fully adapted to the past situation of the company.
- 4) It is easy to decide what impact people give to distant past data, what to do with near-misses and the Monte-Carlo simulation is highly similar to real life situations.
- 5) Due to the algorithm used, although the final probability distribution combined thousands or more probability distributions at each individual loss observation, the computation speed can be almost instant (to complete one cycle, the time consumed is  $< 10^{-2}$  seconds on an ordinary PC). It is therefore possible to carry out tens of thousands of Monte-Carlo simulations to identify confidence interval.
- 6) The program can simulate over the whole interval, can simulate over some pre-indicated loss value interval, it can also simulate on any arbitrary group of risk numbers, can incorporate imagined large loss events or reduce the scale of an unexpected large loss observation which is unlikely to happen again.

In addition to the standard loss probability distribution function and predicted loss values, we have an exit file reporting the following events:

- 1) Very high and high impact, low and extremely low probability events that have no observations.
- 2) Large concentration of single observations (a single loss value that appeared many times).
- 3) Some risk numbers that have actual events with loss values much larger than the management perceived upper bound. This implies that the impact of this risk number needs to be re-assessed.

## References:

- [1] C Alexander and J Pezier, Taking control of operational risk, Futures and Options World 366, 2001.
- [2] L Allen, J Boudoukh and A Saunders, Understanding Market, Credit, and Operational Risk: The Value-at-Risk Approach, Blackwell Publishing, Oxford, 2004.
- [3] A S Chernobai, S T Rachev and F J Fabozzi, Operational Risk, a Guide to Basel II Capital Requirements, Models and Analysis, Wiley Finance, 2007.
- [4] P Cizek, W Hardle and R Weron eds, Statistical Tools for Finance and Insurance, Springer, Heidelberg, 2005.
- [5] P Giudici, Integration of qualitative and quantitative operational risk data: A Bayesian approach, in M G Cruz ed., Operational Risk Modelling and Analysis, Theory and Practice, RISK Books, London, 2004, pp 131-138.
- [6] C L Marshall, Measuring and Managing Operational Risk in Financial Institutions: Tools, Techniques, and Other Resources, John Wiley & Sons, Chichester, 2001.
- [7] M Neil and E Tranham, Using Bayesian Network to Predict Op Risk, Operational Risk, August 2002, pp 8-9.
- [8] J V Rosenberg and T Schuermann, A general approach to integrated risk management with skewed, fat-tailed risks, Technical Report, Federal Reserve Bank of New York, 2004.
- [9] B W Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, London.