
Munich Personal RePEc Archive

Approaches and Techniques to Validate Internal Model Results

Dacorogna, Michel M

DEAR-Consulting

23 April 2017

Online at https://mpra.ub.uni-muenchen.de/79632/
MPRA Paper No. 79632, posted 09 Jun 2017 19:48 UTC

Approaches and Techniques to Validate Internal Model Results

Michel Dacorogna

DEAR-Consulting,
Scheuchzerstrasse 160, 8057 Zurich, Switzerland

April 23, 2017

Abstract

The development of risk model for managing portfolio of financial institutions and insurance companies require both from the regulatory and management points of view a strong validation of the quality of the results provided by internal risk models. In Solvency II for instance, regulators ask for independent validation reports from companies who apply for the approval of their internal models.

Unfortunately, the usual statistical techniques do not work for the validation of risk models as we lack enough data to significantly test the results of the models. We will certainly never have enough data to statistically estimate the significance of the VaR at a probability of 1 over 200 years, which is the risk measure required by Solvency II. Instead, we need to develop various strategies to test the reasonableness of the model.

In this paper, we review various ways, management and regulators can gain confidence in the quality of models. It all starts by ensuring a good calibration of the risk models and the dependencies between the various risk drivers. Then applying stress tests to the model and various empirical analysis, in particular the probability integral transform, we build a full and credible framework to validate risk models.

Keywords: Risk Models, validation, stress tests, statistical tests, solvency.

1 Introduction

With the advent of risk based solvency and quantitative risk management, the question of the accuracy of risk modelling has become central to the acceptance of the results of models both for management and regulators. Model validation is at the heart of gaining trust with the quantitative assessment of risks. From the legal point of view, the Solvency II legislation requires companies who are applying for approval of their internal risk model to provide an independent validation of both the models and their results. From a scientific point of view, it is not easy to ensure a good quality of models that are very complex and contain a fair amount of parameters. Moreover, a direct statistical assessment of the 99.5% quantile over one year is completely excluded. The capital requirements are computed using a probability of 1% or 0.5%, which represents a 1/100 or 1/200 years event. In most of the insured risks, such an event has never been observed or has been observed only once or twice. Even if, for financial return, we know better the tail of the distribution thanks to high frequency data [8], we do not have lots of relevant events for such a probability. This means that the tails of the distributions have to

be inferred from data coming from the last 10 to 30 years in the best cases. The 1/100 years Risk Adjusted Capital (RAC) is thus based on a theoretical estimate of the shock size. It is a compromise between pure betting and not doing anything because we cannot statistically estimate it. Therefore, testing the output of internal models is a must to gain confidence in their results and to understand their limitations.

The crucial question is: What is a "good" model? Clearly, the answer will depend on the purpose of the model and could vary from one purpose to the other. In the case of internal models, a good model would be a model that can predict well the future risk of the company. Since the internal models are designed to evaluate the risk over one year (see [5]), the prediction horizon for the risk is thus also one year. The threshold chosen for the risk measures 99 and 99.5% makes it impossible to directly test the predicted risk statistically, since there will never be enough relevant data at those thresholds. Thus we need to develop indirect strategies to ensure that the end result is a good assessment of the risk. These indirect methods comprise various steps that we are going to list and discuss in this article. First, we present in Section 2 the generic structure of an internal model in order to identify what is the job of model validation. The building of any model starts by a good calibration of its parameters. This is the subject of Section 3. In Section 4, we deal with component testing. Each model contains various components that can be tested independently before integrating them in the global model. We review the various possibilities of testing the components. In Section 5, we explain how to use stress tests to measure the quality of the tails of the distribution forecast. The use of reverse stress testing is explained in Section 6, while conclusions are drawn in Section 7. For completion, we quote the relevant articles of the European Directive in Appendix A and the Delegated Regulation in the appendices in B and C.

2 Structure of an Internal Model and Validation Procedure

There are various types and meanings of an internal model, but they all follow the same structure as it is also noticed in the European directive. First of all, it has to be clear that a model is always an approximation of reality, where we try to identify the main factors contributing to the outcome. This simplification is essential to be able to understand reality and, in our case, to manage the risks identified by the model. The process arriving at an internal model contains three main ingredients:

1. **Determining the relevant assumptions** on which the model should be based. For instance, deciding if the stochastic variable representing a particular risk presents fat tails (high probability of large claims) or can be modeled with Gaussian assumptions, or if the dependence between various risk is linear or non-linear? Should all the dependencies between risk be taken into account or can we neglect some? And so on ...

2. **Choosing the data** that describe best the risk and controlling the frequency of its updates. It is essential to ensure that the input data correspond really to the current extension of the risk. An example of the dilemma, when modelling pandemic: Are the data from the Black Pest of the 14^{th} century still relevant in today's health environment? Or can we use claims data dating back 50 years if available? It is very clear that the model results will depend crucially on the various choices of data that were made but also on the *quality* of the data.

3. **Selecting the appropriate methodology** for developing the model. The actuaries will usually decide if they want to use a frequency/severity model or rather a loss-ratio model for attritional losses, or a natural catastrophes model depending of the risk they want to model. Similarly, selecting the right methodology to generate consistent economic generators to value assets and liabilities, is crucial to obtain reliable results on the diversification between assets and liabilities.

People who have built internal models are familiar with this structure and have discussed endlessly

the various points mentioned above. Yet, often, the validation process could neglect one or the other of these points due to a lack of awareness of the various steps involved in building a model. That is why it is important to understand well the structure of model development.

Once the model is fully implemented, it must also be integrated in the business processes of the company. With the Solvency II requirements of updating the model on a quarterly basis, there is a necessary industrialization phase of the production process leading to internal model results. It does not suffice to have chosen the right assumptions, picked the right data and settled on a methodology, but processes must be built around the model for verifying the inputs, the outputs and producing reports that are well accepted within the organization. Moreover, keeping the model on excel spreadsheets that were probably used to develop it, would not satisfy the criteria set by regulators (see, for instance, Appendix C) but also the need of management to count on reproducible and reliable results.

In the past decades, the importance of information technology in the financial industry has increased significantly, up to a point where it is inconceivable for an insurance company not to have extensive IT departments headed by a chief information officer reporting to the top management. Together with the growing data density, grew the need to develop appropriate techniques to extract information from the data: The systems must be interlinked and the IT landscape must be integrated to the business operations. The first industrialization process was initially with a strong design focus on accounting and administration. The complexity of handling data increased, especially in business areas which were not the main design focus for the IT system. As it was the case 20 years ago with accounting systems, companies need now to enter in an industrialization process of the production of internal model results. It can be summarized in three important steps to be taken care off:

1. First and for all, the company must choose a **conceptual framework** to develop the software. The basic architecture of the applications should be reduced to a few powerful components: the user interface, the object model including the communication layer, the mathematical kernel, and the information backbone. This is a quite standard architecture in financial institutions and is called the three tiers architecture, where the user interfaces can access any service of the object model and of the mathematical kernel, that can in turn access any data within the information backbone. Such a simple architecture ensure interoperability of the various IT systems and thus also their robustness.

2. The next step is the **implementation framework**: How this architecture is translated in an operative design. The software must follow 4 overarching design principles:

   i) **Extensibility:** Allowing for an adaption to new methods as the methodology progresses with time, easily adapting to changes in the data model when new or higher quality data become available. The data model, the modules, as well as the user interfaces evolve with time.

   ii) **Maintainability:** Low maintenance and the ability of keeping up with the changes for example in data formats. Flexibility, in terms of a swift implementation, with respect to a varying combination of data and methods.

   iii) **Testability:** The ability to test the IT components for errors and malfunctions at various hierarchy levels, using an exhaustive set of predefined test cases.

   iv) **Re-usability:** The ability to recombine programming code and system parts. Each code part should be implemented only once if reasonable. In this way, the consistency of code segments is increased significantly.

   Very often, companies will use commercial software like Igloo from Willis Tower Watson, or Remetrica from Aon-Benfield, or others. Nevertheless, their choice of software should be guided

by these principles.

3. The last step is to **design processes around the model**. Several processes must be put in place to ensure the reliable production of results, but also to develop a specific governance framework for the model changes due to either progress in the methodology or discoveries of the validation process (see for instance point 3 in Article 242 of Appendix B). The number of processes will depend on the implementation structure of the model, but they always include at least input data verification and results verification. Responsible persons must be designated for each of them and accountability must be clearly defined.

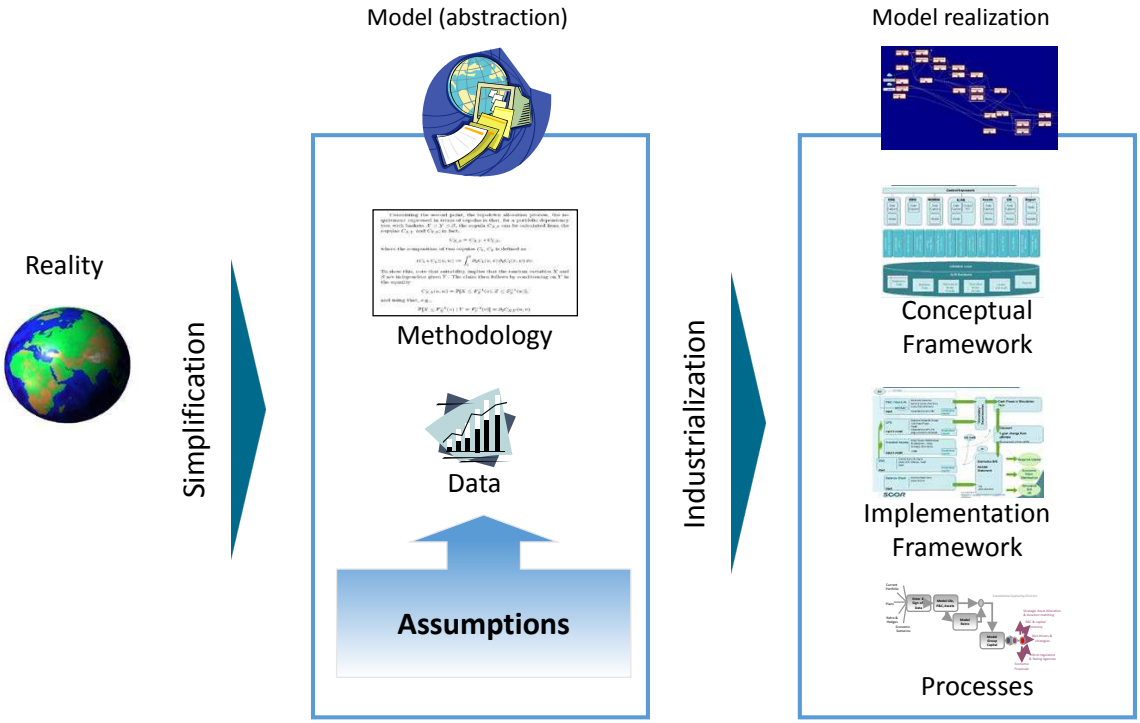


Figure 1: Schematic representation of the modelling framework

In Figure 1, we schematically illustrate the points we present in this section starting from the reality to be modelled, to the industrialization phase that is needed to ensure a smooth production of risk results. Reading the 3 appendices at the end of this document, we see that all those points are present but not structured as proposed here. Model validation will, of course, need to be articulated around the structure described in this section and around the various points mentioned above. The final validation report will be then much more understandable and could be reused for future validations, as regular validation is a pre-requisite of the regulators.

We already said that statistical testing of the capital requirements is impossible given the lack of data, nevertheless having a clear understanding of what needs to be done gives us a good framework for organizing the validation process around these points. In the rest of the text, we list some of the validation procedures that we propose.

3 Calibration

The first step of a good validation procedure is to make sure that the calibration of model parameters is done properly. Any model needs to determine few parameters. These parameters are set looking at data of the underlying process and fitting them to these data. Pricing and reserving actuaries often develop their models based on statistical tests on claims data. This is called "experience rating". Sometimes, they also use risk models based on exposure data, for instance in modelling natural catastrophes ("exposure rating"). There are many models for estimating the one-year variability of claims reserves (see for instance [13] or [11]). In general, internal models are usually composed of probabilistic models for the various risk drivers but also of specific models for the dependence between those risks. Both sets of models need to be calibrated.The most difficult part is to find the right dependence between risks because this requires lots of data. The data requirement is even more difficult to achieve when there is only dependence in the tails. As mentioned above, the probabilistic models are usually calibrated with claims data for the liabilities and with market data for the assets. In other cases, like for natural catastrophes, pandemic or credit risk, stochastic models are used to produce probability distributions based on Monte Carlo simulations.

The new and difficult part of the calibration is the estimation of dependence between risks. This step is indispensable for the accurate aggregation of various risks. Dependencies can hardly be described by one number such as a linear correlation coefficient. Nevertheless, linear correlation is the most used dependence model in our industry. Most reinsurers, however, have long used copulas to model non-linear dependence. Yet, there is often not enough liability data to estimate the copulas, but copulas can be used to translate an expert opinion about conditional probabilities in the portfolio into a model of dependence. The first step is to select a copula with an appropriate shape, usually with increased dependencies in the tail. This feature is observable in certain insurance data, but is also known from stress scenarios. Then, one tries to estimate conditional probabilities by asking questions such as "What about risk Y if risk X turned very bad?". To answer such questions one needs to think about adverse scenarios in the portfolio or to look for causal relations between risks.

An internal model contains usually many risks. For instance, SCOR's model contains few thousand risks, which means a large amount of parameters for describing the dependence within the portfolio. The strategy for reducing the number of parameters must start from the knowledge of the underlying business. This allows to concentrate the efforts on the main risks and to neglect those that are by their nature less dependent. On way of doing this is to develop a hierarchical model for dependencies, where models are aggregated first and then aggregated on another level with a different dependence model. This would reduce the parameter space and concentrate the efforts in describing more accurately the main sources of dependent behavior. Such a structure allows to reduce the number of parameters to estimate from essentially n^2 to n, where n is the number of risks included in the model. If the upper level is modelled by a rv Z and the lowest level by a rv X, the condition for using a hierarchical tree is:

$$\mathbb{P}(X \leq x \mid Y = y, Z \leq z) = \mathbb{P}((X \leq x \mid Y = y)$$

In other words, given that the result of Y influences the information about the result in Z, the latter is not influenced by the distribution of X in Y. Business knowledge helps separating the various lines of business to build such a tree with its different nodes.

Once the structure of dependence for each node is determined, there are two possibilities:

1. If a causal dependence is known, it should be modelled explicitly.
2. Otherwise, non-symmetric copulas (ex. Clayton copula) should be systematically used in presence of tail dependence.

To calibrate the various nodes, we have again two possibilities:

1. If there is enough data, we calibrate statistically the parameters

2. In absence of data, we use stress scenarios and expert opinion to estimate conditional probabilities

For the purpose of eliciting expert opinion (on common risk drivers, conditional probabilities, bucketing to build the tree), we have developed a Bayesian method combining various sources of information in the estimation: PrObEx [1]. It is a new methodology developed to ensure the prudent calibration of dependencies within and between different insurance risks. It is based on a Bayesian model that allows to combine up to three sources of information:

1. **Pr**ior information (i.e. indications from regulators or previous studies)

2. **Ob**servations (i.e. the available data)

3. **Ex**perts opinions (i.e. the knowledge of the experts)

For the last source, experts are invited to a workshop where they are asked to assess dependencies within their Line of Business. The advantage of an approach using copulas is that they can be calibrated once a conditional probability is known. The latter are much easier to assess by experts than a correlation parameter. Once the elicitation process is completed the database of answers can also be assessed for biases. Lack of data cannot be an excuse to use the wrong model. It can be compensated by a rigorous process of integrating expert opinions in the calibration.

4 Component Testing

Every internal model contains important components that will condition the results. Here is a generic list of main components for a (re)insurer:

- An Economic Scenario Generator (ESG), to explore the various states of the World Economy
- A stochastic model to compute the uncertainty of P&C reserving triangles,
- A stochastic model for the natural catastrophes,
- A stochastic model for the pandemic (If there is a significant life book),
- A model for credit risk
- A model for operational risk
- and a model for risk aggregation

Each of these components can be tested independently, to check the validity of the methods employed. These tests vary from one component to the other. Each requires its own approach for testing. We briefly describe here some of the approaches that we use for testing some components.

4.1 Testing ESGs' with PIT

We start with the Economic Scenario Generator as it is a component that can be tested against market data and is central to the valuation of both assets and liabilities. The ESG produces many scenarios, i.e. many different forecast values. Thousands of scenarios together define forecast distributions. We use backtesting for checking how well did known variable values fit into their prior forecast distributions. Here, we need to test the validity of the forecast of a distribution, which is much harder and less straightforward than testing point forecasts. The Testing Method we choose is the Probability Integral

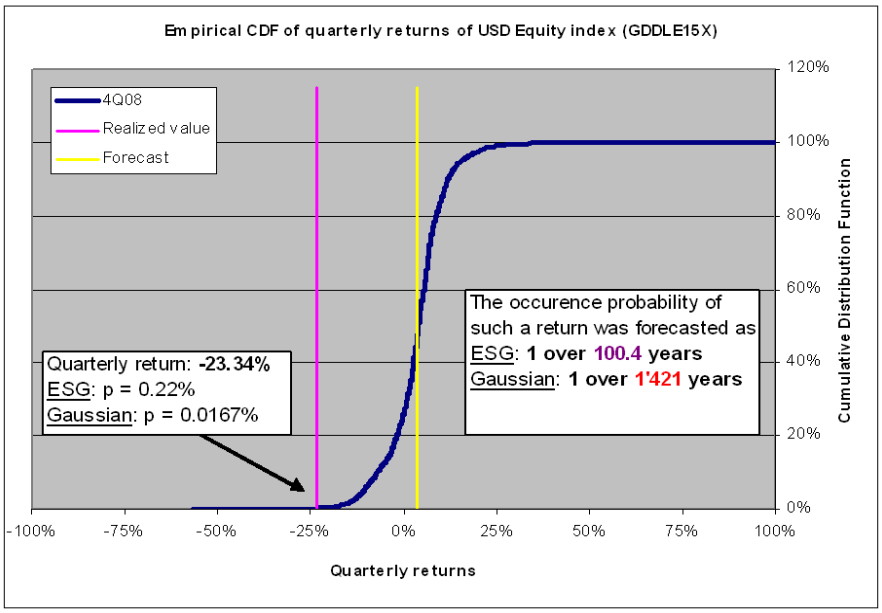


Figure 2: Cumulative distribution forecast of an US Equity Index made in June 2007 for 31.12.2008, the purple line is the actual realization, while the yellow line is the expectation of the distribution forecast.

Transform (PIT) advocated in [9] and [10]. The question is to Determine the cumulative probability of a real variable value, given its prior forecast distribution. The idea of the method is to test the probability of each realized value in the distribution forecast.

Here is a summary of the steps:

1. We define an in-sample period for building the ESG with its innovation vectors and parameter calibrations (e.g. for the GARCH model). The out-of-sample period starts at the end of the in-sample period. Starting at each regular time point out-of-sample, we run a large number of simulation scenarios and observe the scenario forecasts for each of the many variables of the model (see [2]).

2. The scenario forecasts of a variable x at time t_i sorted in ascending order, constitute an empirical cumulative distribution forecast. In the asymptotic limit of very many scenarios, this distribution converges to the marginal cumulative probability distribution $\Phi_i(x) = \mathbb{P}(x_i < x \mid \mathcal{F}_{i-m})$ that we want to test. It is conditioned to the information $\mathcal{F}_{i-m}$ available up to the time t_{i-m} of the simulation start. In the case of a one-step ahead forecast, $m = 1$. The empirical distribution $\hat{\Phi}_i(x)$ slightly deviates from this. The discrepancy, $\Phi_i(x) - \hat{\Phi}_i(x)$ can be quantified by using a formula given in [2]. Its absolute value is less than 0.019 with a confidence of 95% when choosing 5000 scenarios, for any value of x and any tested variable. This is accurate enough, given the limitations due to the rather low number of historical observations.

3. For a set of out-of-sample time points t_i, we now have a distribution forecast $\hat{\Phi}_i(x_i)$ as well as a historically observed value x_i. The cumulative distribution $\hat{\Phi}_i(x_i)$ is used for the following Probability Integral Transform (PIT): $Z_i = \hat{\Phi}_i(x_i)$). The probabilities Z_i, which are confined between 0 and 1 by definition, are used in the further course of the test. A proposition proved by

[9] states that the Z_i are i.i.d. with a uniform distribution $U(0,1)$ if the conditional distribution forecast $\Phi(x_i)$ coincides with the true process by which the historical data have been generated. The proof is extended to the multivariate case in [10]. If the series of Z_i significantly deviates from either the $U(0,1)$ distribution or the i.i.d. property, the model does not pass the out-of-sample test.

In Figure 2, we illustrate the PIT procedure with just one example. We display the forecast of the cumulative distribution of returns of a US stock index as produced in June 2007 for the 31.12.2008. We also draw the expected value (yellow vertical line) and the value actually reached at that time (-23.34%, purple vertical line) and look at its probability in our forecast. We note that our model had, in June 2007, a much too optimistic expectation for the fourth quarter of 2008. we remind the reader that the ESG is not thought to be a point forecast model. It is here to assess the risk of a particular financial asset. We see that it does this pretty well, it attributed in June 2007 a reasonable probability (1 over 100 years) to the occurrence of the third quarter 2008[1], while the Gaussian model would give an extremely low probability of less than 1 over 1400 years. This is an extreme case, but it shows how the PIT test can be applied to all the important outputs of the ESG to check its ability to predict a good distribution, and thus the risk, of various economic variables.

4.2 Testing the One Year Change of P&C Reserves

One of the biggest changes in the methodology that has been initiated by the new risk-based regulation, is the computation of the one-year risk of P& C reserves. It is an important component of any P&C insurance risk. Testing the quality of the model to compute the one-year change is thus also one of the important steps towards validating a model. There are many ways one can think of testing this. We present here a method developed recently (see [7]) that can be applied also for other validation procedures. It consists in designing simple stochastic models to reach the ultimate claim value that can then be used to simulate sample paths to test the various methods for computing the one-year change risk. Since claims data are scarce to do rigorous statistical tests on the methods, we generate with these models enough data to run the methods. The advantage of this approach is that, by choosing simple models, one is able to obtain analytic or semi-analytic solutions for the risk against which the statistical methods can be tested.

In this example, we present the results of the model testing using two methods for computing the one-year change risk:

1. The approach proposed by Merz and Wüthrich [13] as an extension of the Chain-Ladder following Mack's assumptions [12]. They obtain an estimation of the mean square error of the one-year change based on the development of the reserve triangles using the Chain-Ladder method.

2. An alternative way to model the one-year risk developed by Ferriero: the Capital Over Time (COT) method [11]. The latter assumes a modified jump-diffusion Lévy process to the ultimate and gives a formula, based on this process, for determining the one-year risk as a portion of the ultimate risk.

We present here results, obtained in [7], for two simple stochastic processes to reach the ultimate and for which we have derived *explicit formulae*:

1. a model where the stochastic errors propagate linearly (linear model)

2. and a model where the stochastic errors propagate multiplicatively (multiplicative model)

[1]The yearly return of 2008 was the second worst performance of the S&P500 measured over 200 years. Only the year 1933 presented a worst performance!

Table 1: Statistics for the first year capital on 500 simulated triangles with the *linear model*. Are displayed: the mean first year capital, the standard deviation of the capital around that mean and the mean absolute and relative deviations (MAD,MRAD) from the true value. The mean reserves estimated with chain-ladder are 101.87, which are consistent with the reserves calculated with our model, i.e. $n(1-1/2^I)p = 100000(1-2^{19})0.001 = 100.00$.

Method	Mean	Std. dev.	MAD	MRAD
Linear Model:				
Theoretical value	18.37	3.92	–	–
COT, without jumps	19.08	3.93	0.71	4.14%
COT, with jumps	18.81	3.86	0.43	2.47%
Merz-Wüthrich	252.89	149.6	234.5	1365.6%
Multiplicative Model:				
Theoretical value	29.36	21.97	–	–
COT, without jumps	26.75	19.84	2.54	8.19%
COT, with jumps	28.30	20.98	1.07	3.48%
Merz-Wüthrich	22.82	15.77	12.7	43.2%

For both processes, we compare the two methods mentioned above to see how they perform in assessing a risk that we explicitly know thanks to the analytic solutions of our models. The linear model does not follow the assumptions under which the Chain-Ladder model works, We this expect that the Merz-Wüthrich method will fair poorly.

In Table 1, we present capital results for the linear model. The mean is the one year capital, while the reserves are 101.87 with the chosen parameters. This is the typical capital intensity (capital over reserves) of the Standard Formula of Solvency II. We immediately see that the Merz-Wüthrich method gives results that are way off due to the fact that its assumptions are not fulfilled. This illustrates the fact that the choice of an appropriate method is crucial to obtain credible results. We also see that the COT method gives more reasonable numbers particularly the one "without jumps" as we would expect from the nature of the stochastic process that does not include any jump.

The multiplicative model is better suited for the Chain-Ladder assumptions as we cn aee this in Table 1 where we report similar results also for this model. As is to be expected, the capital intensity is higher than for the linear model (29%) as multiplicative fluctuations are stronger than linear ones. In this case, all the methods underestimate the capital but all of them fair about the same. The standard deviation is smallest for Merz-Wüthrich but the error the largest. One should also not here that the COT with jumps leads the best results as one would expect due to the nature of the stochastic process, which would involve large movements.

This example is presented here to illustrate the fact that one can, with such an approach, test the use of certain methods and gain confidence about their ability to deliver credible results for the risk[2]. In general, a technique to make up for the lack of data, is to design models that can generate data where the result is known and use these data to test the methods. It is what we also do in the next section.

4.3 Testing the Convergence of Monte Carlo Simulations

One of the most difficult and least tested quantity is the number of simulations used for obtaining aggregated distributions. Until recently, internal models would use 10'000 simulations. Nowadays, 100'000 simulations seems to become the benchmark without clear justifications other than the capacity of the computers and the quality of the software. Nevertheless, it would be important to know how well the model is converged. The convergence of the algorithm is definitely an important issue when one is aggregating few hundreds or thousands of risks with their dependence. One way to do this is to

[2]Note that the results in Table 1 are taken from Reference [7].

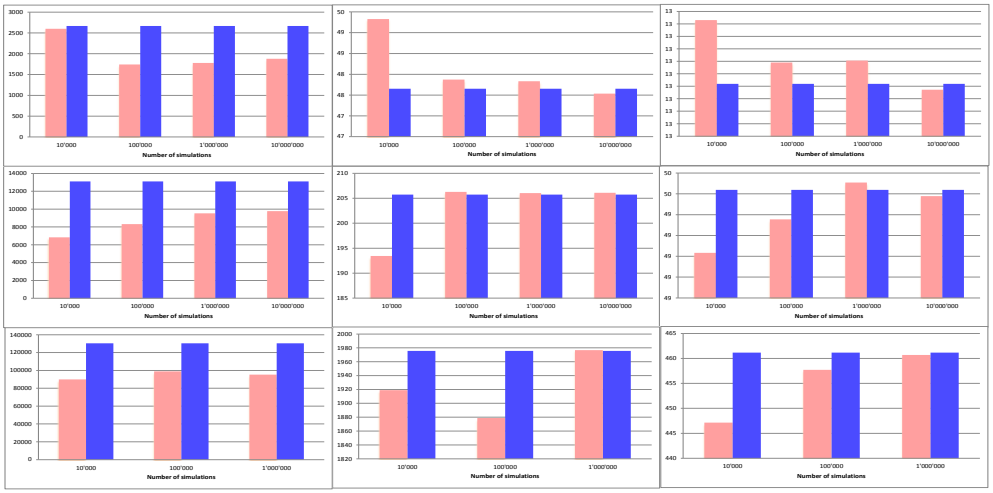


Figure 3: Convergence of the TVaR of S_n at 99.5% for $\alpha = 1.1,\ 2,\ 3$ from left to right, for an aggregation factor $n = 2,\ 10,\ 100$ from up to down. The dark plots are for the analytical values and the light ones are the average values obtained from the MC simulations. The y-scale gives the normalized TVaR ($TVaR_n/n$) and is the same for each column.

obtain analytical expressions for the aggregated distribution and then test the Monte Carlo simulation against this benchmark.

It is the path explored in [6] where we give explicit formulae for the aggregation of Pareto distributions coupled via a Clayton survival copulae and Weibull distributions coupled with Gumbel copulae. In Figure 3, we present results for the normalized TVaR (Expected Shortfall) ($TVaR/n$) for various tail indices $\alpha = 1.1, 23$ and different level of aggregation $n = 2, 10, 100$. On the figure, we see that:

- The normalized TVAR of S_n, $TVaR_n/n$, decreases as n increases
- The TVaR decreases as α increases
- The rate of convergence of $TVaR_n/n$ increases with n
- The heavier the tail, the slower the convergence
- In the case of very heavy tail and strong dependence ($\alpha = 1.1$ and $\theta = 0.91$), we do not see any satisfactory convergence, even with 10 million simulations, and for any n
- When $\alpha = 2, 3$, the convergence is good from 1 million, 100'000 simulations onwards, respectively.

The advantage of having explicit expressions for the aggregation becomes evident here: We can explore in details the convergence of the Monte Carlo simulations (MC).

We can go one step further by looking at other quantities of interest. For this, we define also the diversification benefit as in [3]. Recall that the diversification performance of a portfolio S_n is measured on the gain of capital when considering a portfolio instead of a sum of standalone risks. The capital is defined by the deviation to the expectation, and the diversification benefit (see [3]) at a threshold κ ($0 < \kappa < 1$), by

$$D_\kappa(S_n) = 1 - \frac{\rho_\kappa(S_n) - \mathbb{E}(S_n)}{\sum_{i=1}^n (\rho_\kappa(X_i) - \mathbb{E}(X_i))} = 1 - \frac{\rho_\kappa(S_n) - \mathbb{E}(S_n)}{\sum_{i=1}^n \rho_\kappa(X_i) - \mathbb{E}(S_n)} \tag{1}$$

where ρ_κ denotes a risk measure at threshold κ. This indicator helps determining the optimal portfolio of the company since diversification reduces the risk and thus enhances the performance. By making sure that the diversification benefit is maximal, the company obtains the best performance for the

lowest risk. However, it is important to note that $D_\kappa(S_n)$ is not a universal measure and depends on the number of risks undertaken and the chosen risk measure.

The convergence is even clearer seen in the following table:

Table 2: Relative errors (when comparing results obtained by MC and analytical ones) of the $TVaR_n$ and the diversification benefit D_n for S_n, at 99.5% and for various α, as a function of the aggregation factor n computed with 1 million simulations.

	n=2	n=10	n=100
α=3			
$TVaR_n$	0.30%	0.14%	-0.10%
D_n	-1.30%	-0.25%	0.15%
α=2			
$TVaR_n$	0.38%	0.14%	0.05%
D_n	-2.61%	-0.44%	-0.14%
α=1.1			
$TVaR_n$	-33.3%	-27.3%	-26.9%
D_n	1786%	742%	653%

where we see a decreasing estimation error by MC when increasing the aggregation factor, with small errors for $\alpha = 3$ and 2 and substantial errors for very fat tails and strong dependence. In the latter, we also see a systematic underestimation of the TVaR and an overestimation of the diversification benefit, whatever the aggregation factor. While with the thiner tails and lower dependence, MC has a tendency to overestimate the TVaR and underestimate the diversification benefit except for $n = 100$. Note that the error decrease is large between 2 and 10 but much smaller afterwards[3].

Overall, we see that, if $\alpha \leq 2$, the convergence is good with 100'000 simulations. Problems start when $\alpha < 2$. Luckily, the first case is the most common case for (re)insurance liabilities, except for earthquakes, windstorms and pandemic. This is reassuring even though it is not clear what would happen with small α's and very strong dependence. Some more work along those lines is still needed to fully understand the convergence of MC given various parameters for the tails and the dependence.

5 Stress Test to Validate the Distribution

Stress testing the model means that one looks at the way the model reacts to a change of its inputs. There are at least three ways of stress testing a model:

1. Testing the sensitivity of the results to certain parameters (sensitivity analysis)
2. Test the predictions against real outcomes (historical test, via P&L attribution for lines of business (LoB) and assets)
3. Test the model outcomes against predefined scenarios

The sensitivity analysis is important. It is not possible to base management decisions on results that could drastically change if some unimportant parameters are modified in the input. Unfortunately, note that this statement contains the adjective "unimportant", which is hard to define. Clearly, the

[3]Figure 3 and Table 2 are taken from Reference [6]

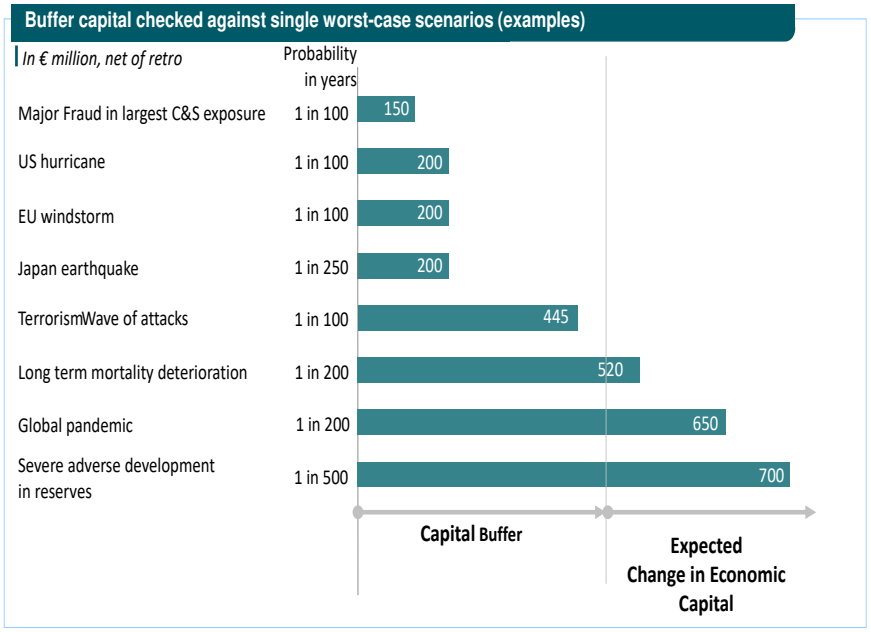


Figure 4: We display the results of scenarios that could affect the balance sheet of a reinsurance company with its estimated probability of occurence. We also compare the values to the size of the capital buffer and the expected next year profit of the company.

question is delicate because one has to determine, in advance, what are the parameters that justify a strong sensitivity of the results, like, for instance, the heaviness of the tails or the strength and the form of the dependence, from those that should not affect too strongly the results. We studied one of these important parameters in the previous section talking about the convergence of the MC. An increase of the number of simulations should not affect too much the results. In any case, sensitivity analysis must be conducted on all parameters and the results should be discuss according to the expected effects these parameters should have. In certain cases, big variations of the capital is justified particularly when we change the assumptions that affect directly the risk.

The second point, is closely related to the PIT method described in Section 4.1, except that here we do not have enough data to test if the probabilities are really i.i.d. The only thing we can do is ensure that the probabilities obtained are reasonable both at a disaggregated level (lines of business or types of assets) as well as an aggregated level (company's results for the whole business or for a large portfolio). This type of backtest must be performed each year, and with experience accumulating, we should be able to draw conclusions on the overall quality of the forecast. In a way, we are testing here the belly of the distribution rather than the tails but nevertheless this is also important as the day to day decisions have often to do with those types of probability in mind rather than the extremes.

Scenarios can be seen as thought experiments about possible future states of the world. Scenarios are different from sensitivity analysis where the impact of a (small) change to a single variable is evaluated. Scenario results can be compared to simulation results in order to assess the probability of the scenarios in question. By comparing the probability of the scenario given by the internal model to the expected frequency of such a scenario, we can assess whether the internal model is realistic and has actually taken into account enough dependencies between risks. Recently, scenarios have caught the interest of regulators because they allow to represent a situation that can be understood by both management and regulators. On one hand, analyzing how the company would fare in case of a big natural catastrophe or in face of a serious financial crisis is a good way to gain confidence

on the value of the risk assessment made by the quantitative models. On the other hand, using only scenarios to estimate the capital needed for the company is a guarantee that they will miss the next crisis, which is bound to come from an unseen combination of events. That is why a combination of probabilistic approaches and scenarios is a good way of validating model results. In Figure 4, we present an example published by SCOR showing how some scenarios that could hit the balance sheet of the company measure against the capital and the buffer the company holds for covering the risks.

6 Reverse Stress Test to Validate Dependence Assumptions

Internal models based on stochastic Monte Carlo simulations produce many scenarios at each run (typically few thousands). Usually little of these data is used: some averages for computing capital and some expectations. Yet, these outputs can be put at use to understand the way the model works. One example could be to select the worst cases and look at what are the scenarios that make the company bankrupted. Two questions to ask on these scenarios:

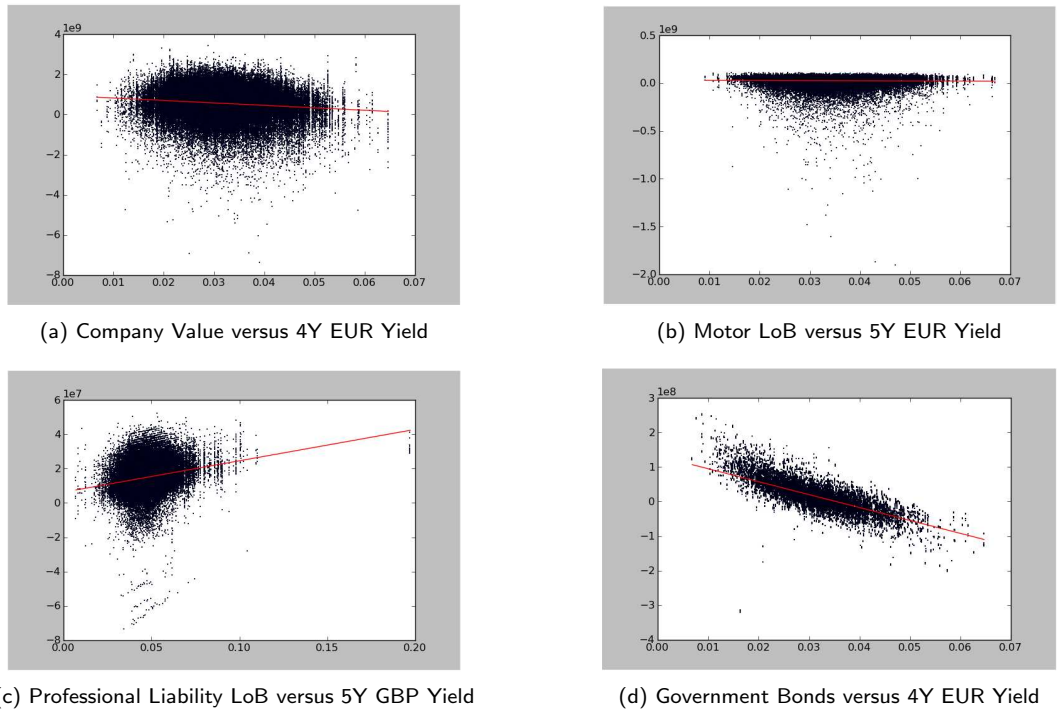
1. Are these scenarios credible, given the company portfolio? Would such scenarios really affect the company?

2. Are there other possible scenarios that we know of and do not appear in the worst Monte Carlo simulations?

If the answers to the first question is positive and negative to the second, we gain confidence in the way the model apprehend the extreme risks and is describing our business. Inversely, one could look at the question how often the model would give negative results after one year. If this probability is very low, we would know that our model is too optimistic and would probably underestimate the extreme risk. If the answer is the reverse, the conclusion would be that our model is too conservative and neglects some of our business realities. In the case of reinsurance, looking at the published balance sheets, a typical frequency of negative results would be once every ten years for a healthy reinsurance company. This is the kind of reverse back testing that can be done on simulations to explore the quality of the results.

Other tests can be envisaged and are also interesting like looking at conditional statistics. A typical question would for instance be: how is the capital going to behave if interest rises? Exploring the dependence of results on certain important variables is a very good way to test the reasonableness of the dependence model. As we already explained, validation of internal models does not mean statistical validation because there will never be enough data for reaching a good conclusion on high enough significance levels. In this context reasonableness, given our knowledge of the business and past experience, is the most we can hope to achieve. In the next few figures, we present regression plots where we show the dependency between interest rates and changes in economic value (of the company or certain certain typical risks). The plots are based on the full 100000 scenarios of the Group Internal Model (GIM). By analyzing, the GIM Results on this level, we can follow up on a lot of effects and test if they make sense.

We start this example with the company economic value after one year that is displayed on Figure 5(a). We choose to do a regression against the 4Y EUR government yield because the liability portfolio of this company has a duration of roughly 4Y and the balance sheet is denominated in EUR. In all the graphs, the chosen interest rate is the one corresponding to the currency denomination of the portfolio and its duration. We see that as interest rate grows the Value of the company slightly decreases. This decrease is due to an increase in inflation, which is linked to IR increase in our Economic Scenario Generator (ESG). In Figure 5(b), we regress Motor LoB versus the 5Y EUR yield. The value of motor business depends only very weakly on interest rate as it is relatively short tail. In Figure 5(c), we show the regression between professional liability and the 5Y GBP yield. The value of professional liability



(a) Company Value versus 4Y EUR Yield

(b) Motor LoB versus 5Y EUR Yield

(c) Professional Liability LoB versus 5Y GBP Yield

(d) Government Bonds versus 4Y EUR Yield

Figure 5: We display here typical regression analysis on the simulation results of the GIM. This is what is called reverse tests.

business depends heavily on interest rate as it takes a long time to develop to ultimate and the reserve can earn interest for a longer time. The last graph displayed in Figure 5(d) is related to the regression of the government bond asset portfolio and the 4Y EUR yield. Here the relation is obvious and also well followed by the simulations: Bond value depends mechanically on interest rate. When interest rate increases the value decreases.

Looking at all these graphs helps to convince us that the behavior of the various risks captured in the portfolio with respect to interest rate is well described by the model and that dependence on this very important risk driver for insurance business is well modeled. It is another form of gaining confidence in the accuracy of the model results and an important one as it makes use of the full simulation results and not only some sort of average or one particular point on the probability distribution (like VaR, for instance). On these graphs, we can also inspect the dispersion around the regression line. It represents the uncertainty around the main behavior. For instance, we notice that, as expected, in Figures 5(b) and (d), there is little dispersion, while in Figures 5(a) and (c), we have a higher dispersion as the interest rate is, by far, not the only risk driver of those portfolios. This is only an example of the many dimensions that can be validated through reverse testing. It is definitely an important piece of our tool box for validating the results of our models.

7 Conclusion

The development of risk models is an important step for improving risk awareness in the company and anchoring risk management and governance deeper in industry practices. With risk models, quantitative analysts provide management with valuable risk assessments, especially in relative terms, as well as guidance in business decisions. Quantitative assessments of risk help putting the discussion on a sensible level rather than oposing unfounded arguments. It is thus essential to ensure that

the results of the model delivers a good description of Reality. Model validation is the way to gain confidence in the model and ensure its acceptance by all stakeholders. However, this is a difficult task because there is no straightforward way of testing the many outputs of a model. It is only by combining various approaches that we can come to a conclusion regarding the suitability of the risk assessment.

Among the strategies to validate model let us recall those that we presented or mentioned in this paper:

- Ensure a good calibration of the model through various statistical techniques.
- Use data to test statistically certain parts of the model (like the computation of the risk measure, or some particular model like ESG or Reserving Risk).
- Test the P&L attribution to LoBs against real outcome.
- Test the sensitivity of the model to crucial parameters.
- Compare the model output to stress scenarios.
- Compare the real outcome to its predicted probability by the model.
- Examine the simulation output to check the quality of the bankruptcy scenarios (reverse back-test).

Beyond pure statistical techniques, this list provides a useful set of methods to be used and also researched to obtain a better understanding of the model behavior and for convincing management and regulators that the techniques used to quantify the risks are adequate and the results represent really the risks facing the company. No doubt that with the experience we are now gaining in this field. We will also make progress in the near future by doing research for defining a good strategy to test our models. As long as we keep in mind that we need to be rigorous in our approach and keep the scientific method for assessing the results, we will be able to improve both our models and their validation.

References

[1] P. Arbenz, D. Canestraro, 2012, Estimating Copulas for Insurance from Scarce Observations, Expert Opinion and Prior Information: A Bayesian Approach, *Astin Bulletin*, vol. **42**(1), pages 271-290.

[2] P. Blum, 2004, On some mathematical aspects of dynamic financial analysis, *ETH PhD Thesis*, pages 1183.

[3] R. Bürgi, M. M. Dacorogna and R. Iles, 2008, *Risk Aggregation, dependence structure and diversification benefit*. chap. 12, pages 265-306, in "Stress testing for financial institutions", edited by Daniel Rösch and Harald Scheule, Riskbooks, Incisive Media, London.

[4] M. Busse, U. A. Müller, M. M. Dacorogna, 2010, Robust estimation of reserve risk, *Astin Bulletin*, vol. **40**(2), pages 453-489.

[5] M. M. Dacorogna, 2015, A change of paradigm for the insurance industry, *SCOR Papers*, , available under `http://www.scor.com/en/sgrc/scor-publications/scor-papers.html`.

[6] M. M. Dacorogna, L. El Bahtouri, M. Kratz, 2016, Explicit diversification benefit for dependent risks, *SCOR Paper* no. 38, available under `http://www.scor.com/en/sgrc/scor-publications/scor-papers.html`.

[7] M. M. Dacorogna, A. Ferriero, D. Krief, 2015, Taking the one-year change from another angle, *submitted for publication*, available under `http://www.SSRN/XXXX`.

[8] M. M. Dacorogna, U. A. Müller, O. V. Pictet, C. G. De Vries, 2001, Extremal forex returns in extremely large data sets, *Extremes*, vol. **4**(2) pages 105-127.

[9] F. X. Diebold, T. Gunther, A. Tay, 1998,Evaluating density forecasts with applications to financial risk management, *International Economic Review*, **39**(4), pages 863-883.

[10] F. X. Diebold, J. Han, A. Tay, 1999,Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange, *Review of Economics and Statistics*, vol. **81**, page 661-673.

[11] A. Ferriero, 2016, Solvency capital estimation, reserving cycle and ultimate risk, *Insurance: Mathematics and Economics*, vol. **68**, pages 162-168.

[12] T. Mack, 1993, Distribution-free calculation of the standard error of chain ladder reserve estimates *Astin Bulletin*, vol. **23**(2), pages 213-255.

[13] M. Merz, M. V. Wüthrich, 2008, Stochastic Claims Reserving Methods in Insurance, *Wiley Finance*, John Wiley & Sons, Ltd, Chichester.

A Article 124 on "Validation Standards" of the European Directive

Article 124
Validation Standards

Insurance and reinsurance undertakings shall have a regular cycle of model validation which includes monitoring the performance of the internal model, reviewing the ongoing appropriateness of its specification, and testing its results against experience.

The model validation process shall include an effective statistical process for validating the internal model which enables the insurance and reinsurance undertakings to demonstrate to their supervisory authorities that the resulting capital requirements are appropriate.

The statistical methods applied shall test the appropriateness of the probability distribution forecast compared not only to loss experience but also to all material new data and information relating thereto.

The model validation process shall include an analysis of the stability of the internal model and in particular the testing of the sensitivity of the results of the internal model to changes in key underlying assumptions. It shall also include an assessment of the accuracy, completeness and appropriateness of the data used by the internal model.

B Article 241 on "'Model Validation Process"' of the Delegated Regulation of the 17^{th} of January 2015

Article 241
Model Validation Process

1. The model validation process shall apply to all parts of the internal model and shall cover all requirements set out in Articles 101, Article 112(5), Articles 120 to 123 and Article 125 of Directive 2009/138/EC. In the case of a partial internal model the validation process shall in addition cover the requirements set out in Article 113 of that Directive.

2. In order to ensure independence of the model validation process from the development and operation of the internal model, the persons or organisational unit shall, when carrying out the model validation process, be free from influence from those responsible for the development and operation of the internal model. This assessment shall be in accordance with paragraph 4.

3. For the purpose of the model validation process insurance and reinsurance undertakings shall specify all of the following:

   (a) the processes and methods used to validate the internal model and their purposes;

   (b) for each part of the internal model, the frequency of regular validations and the circumstances which trigger additional validation;

   (c) the persons who are responsible for each validation task;

   (d) the procedure to be followed in the event that the model validation process identifies problems with the reliability of the internal model and the decision-making process to address those problems.

C Article 242 on "'Validation Tools" of the Delegated Regulation of the 17^{th} of January 2015

Article 242
Model Validation Tools

1. Insurance and reinsurance undertakings shall test the results and the key assumptions of the internal model at least annually against experience and other appropriate data to the extent that data are reasonably available. These tests shall be applied at the level of single outputs as well as at the level of aggregated results. Insurance and reinsurance undertakings shall identify the reason for any significant divergence between assumptions and data and between results and data.

2. As part of the testing of the internal model results against experience insurance and reinsurance undertakings shall compare the results of the profit and loss attribution referred to in Article 123 of Directive 2009/138/EC with the risks modelled in the internal model.

3. The statistical process for validating the internal model, referred to in the second paragraph of Article 124 of Directive 2009/138/EC, shall be based on all of the following: (a) current information, taking into account, where it is relevant and appropriate, developments in actuarial techniques and the generally accepted market practice; (b) a detailed understanding of the economic and actuarial theory and the assumptions underlying the methods to calculate the probability distribution forecast of the internal model. 4. Where insurance or reinsurance undertakings observe in accordance with the fourth paragraph of Article 124 of Directive 2009/138/EC that changes in a key underlying assumption have a significant impact on the Solvency Capital Requirement, they shall be able to explain the reasons for this sensitivity and how the sensitivity is taken into account in their decision-making process. For the purposes of the fourth subparagraph of Article 124 of Directive 2009/138/EC the key assumptions shall include assumptions on future management actions.

4. The model validation process shall include an analysis of the stability of the outputs of the internal model for different calculations of the internal model using the same input data.

5. As part of the demonstration that the capital requirements resulting from the internal model are appropriate, insurance and reinsurance undertakings shall compare the coverage and the scope of the internal model. For this purpose, the statistical process for validating the internal model shall include a reverse stress test, identifying the most probable stresses that would threaten the viability of the insurance or reinsurance undertaking.