Model Validation for Insurance Enterprise Risk and Capital Models

Sponsored by
CAS, CIA, SOA Joint Risk Management Section

Prepared By
Markus Stricker
Shaun Wang
Stephen J. Strommen

April 2014

Acknowledgments

This research project was sponsored by the Casualty Actuarial Society, Canadian Institute of Actuaries, and Society of Actuaries' Joint Risk Management Section. Steven Siegel and Barbara Scott of the Society of Actuaries provided excellent guidance and coordination of the project. The authors thank the Project Oversight Group, and particularly Ben Neff, Daniel Hui, David Schraub, Glenn Meyers, Jonathan Glowacki, Madhu Windo, Richard de Haan, and Ron Harasym, for their valuable input and detailed comments.

Abstract/Executive Summary

Internal risk models are widely used by insurance companies and banks in measuring risk, calculating overall capital requirements, and allocating capital for business decisions. In addition, the valuation and pricing of complex financial and insurance products is regularly carried out using sophisticated mathematical models because of the unavailability of market prices for these products. The use of models for these purposes introduces risk through the potential for model error. This is especially true in situations where the financial framework for the operation of a business relies substantially on the results of a model.

The intent of model validation is to limit the risk that use of a model could mislead management into making poor decisions. That risk can be referred to as "model risk." The first part of this paper identifies and describes five distinct elements of model risk and then outlines a model validation process with specific steps to address each element.

Model validation often takes place in a dynamic environment where changes in reporting relationships and changes in models are taking place continuously. The second part of the paper discusses the challenges faced when performing model validation in a dynamic environment, with a focus on the risk management control cycle.

Models have played a significant role in some high-profile stories in the financial news. We briefly discuss the role of models in AIG's difficulty with credit default swaps and the role of models at JP Morgan Chase in the case of the London Whale trades. Finally, we provide a case study that examines the many ways models and model risk are playing a role in the developing story of a new life insurance product: Universal Life with secondary guarantees.

Model validation should be viewed as one part of the larger process of risk management, which includes managing the use of models within an enterprise. Management must decide when and where models will be used in decision making and should put a governance process in place to manage responsibilities and relationships among parties that design, implement, execute, and depend on the results of models. Model validation is a part of model governance, and because validation touches all of these parties, it can be confused with governance. However, although a validation determines whether the responsibilities and relationships between parties have been documented and are working correctly, it does not create those relationships or manage their evolution over time. A validation will determine whether a model is working correctly and providing reliable information, but does not determine whether a model should be built and relied upon in the first place. With all of this in mind, we have included an appendix that compares our five elements of model risk (and validation) to some other well-known frameworks for the use of models: the Solvency II criteria for the regulatory use of internal models, the Model Validation Principles published by the North American CRO Council, and the supervisory guidance on model risk management provided to banks by the Office of the Comptroller of the Currency.

Contents

Introduction

Definition of Model Risk

The primary purpose of model validation is to assess and communicate the level of model risk in light of the intended application. A side effect of the validation process is that it often leads to suggestions for improvement and, consequently, to a reduction of model risk.

Since validation is strongly connected with model risk, we need to define model risk. In broad terms, model risk can arise from various forms of errors or from inappropriate construction or use of the model. On this general level it is hard to derive validation principles. Thus, to make it more practical, it is important to break down the general definition into constituent elements that address specific kinds of errors and inadequacies. The following definition has been first stated [13]:

1. *Conceptual risk:* The risk that the modeling concepts are not suitable for the purpose of the application. This risk includes methodological risk. We think that a concept is slightly more general than a method, for example, whether or not a certain risk factor is taken into account in a model is conceptually important, even if we have not yet described an explicit method to handle this risk factor in the model. Decisions regarding which risk factors to simulate and the methods used to simulate them are both included in this category.
2. *Implementation risk:* There are two forms of implementation risk:
    - The risk that the wrong algorithms were chosen to implement the specified modeling concepts
    - The risk that appropriate algorithms were chosen, but they contain coding errors and bugs.
3. *Input risk:* The risk that the input parameters are inappropriate, incomplete, or inaccurate.
4. *Output risk:* The risk that the key figures and statistics that can be produced by the model do not support the business purpose or are too sensitive with respect to the provided input parameters. The latter cannot always be detected on the conceptual level. It is also not generally bad to have a sensitive model. Model risk is introduced if input data can only be coarsely estimated and output is very sensitive with respect to these parameters.
5. *Reporting risk:* The risk that the representation of the output for the business users is incomplete or misleading. This is related to what is called the use test under Solvency II. In a use test, the firm has to prove to the regulator that the model, that is, the reports, are used in a business decision process, for example, pricing and business planning. Some people consider this to be outside the scope of a model validation, but we disagree. It may well be that the detailed output of the model depicts the risk situation very well for a technical expert. But if the detailed output is condensed into a few statistical key figures that are reported to management without the details, these statistical key figures may lend themselves easily to misinterpretation.

The typical example is a measure of earnings volatility, that is, the standard deviation. If two portfolios have identical standard deviations, this does not mean that the risks in these portfolios are the same. Since the standard deviation is a symmetric measure around the mean, it could well be that one portfolio suffers from exactly the opposite effect of chances than the other. With almost any statistical key figure,

including Value at Risk (VaR) and Tail Value at Risk, we can create examples of portfolios with identical key figures and very different risk characteristics.

The separation between output risk and reporting risk is not so obvious, because both deal with the outputs of the model. The importance of separating them will become clearer in the validation process section when we discuss the embedding of the validation process in the risk management control cycle. For now, let us just point out that the difference between these two risks is that they are addressed to two different groups of people with different skills and responsibilities, and different levels of detail are provided to these two separate groups. The example above of a standard deviation or a VaR figure being unchanged should not provoke a technical expert to think that the risk situation is unchanged, but depending on the way such a result is presented, it might subtly suggest to management that nothing had changed. Hence, *reporting risk* has much to do with effective communication with the decision makers, whereas *output risk* is the technical issue of whether the outputs are correct and can be interpreted in the context of their intended application.

It is very important to have a clear and practical definition of the elements of model risk because this provides the basis for communication between the various stakeholders of a model validation. Without a clear separation of model risk sources, one can discuss very high-level validation principles on which everybody agrees, but be left puzzled as to what really needs to be done in a validation. One reference that does provide guidance on carrying out a validation is Lloyd's of London's guidance to managing agents on model validation in Solvency II [2]. However, that document lacks a definition of model risk and, as a consequence, misses one important model risk source, namely, implementation risk. We believe that the elements of model risk as defined here, combined with the specific actions we outline to address each element, provide a comprehensive and concrete approach to model validation that is unique in the literature at this time.

Survey of Validation Approaches Used Elsewhere

There is a wealth of experience with validation processes in other fields from which we can learn. Model validation takes place in the context of engineering products, software development, and food and drug and environmental models.

The most directly applicable results stem from software development where significant effort goes into test procedures [4]. The actuarial models we wish to validate are implemented in software, and thus the development team should apply these procedures to reduce implementation risk. The validation team can use software metrics to detect whether test procedures have been applied diligently. A typical software metric is how the code is covered by automated test cases. Certification processes for software used in avionics gear prescribes the level of code coverage. Other measures deal with the complexity of the program, and yet others measure how many independent people have been involved in creating and testing sections of the code or its documentation. It is surprising that such measures yield insights into the correctness of the implementation. In particular, the latter one, which shows only how many people have been involved, runs contrary to most business people's intuition. Panko [7] has shown in his studies that spreadsheet users are overly confident: If they created a spreadsheet model themselves and nobody else

was involved, then they believe that the implementation is likely to be less error prone than if other people were involved—but the contrary appears to be true.

Although the connection between implementation risk and software testing is straightforward, we are surprised to see that professional test procedures have not been adopted by many risk management departments. Since the software market is not regulated, software validation procedures do not address some of the other model risk components, in particular, the conceptual risk and the reporting risk.

Although the market for engineering products is also not regulated in the same way as the insurance market, failure of some of the products (e.g., jet engines) would have a significant impact on people's lives, and thus, there are strict procedures (e.g., FAA regulations) to bring such products to the market and to operate them. As these products usually don't change and are operated in a stable environment, the validation procedures are more like a set of standardized tests.

The situation with food and drugs is different because the product remains stable and few models are involved. Yet, it is similar to the risk model environment for two reasons: First, the producer cannot completely test his or her product in a real-life setting. Lab mice play the role of the model, but mice are not humans. And clinical trials usually focus on the normal dosage, whereas in reality one cannot control how individuals apply the product—people may take too much or too little or a combination of medicines, and as well as they buy too little or too much or a strange combination of financial products. Second, the regulator is there to protect the individuals.

The closest relation to risk model validation can be found in environmental model validation. Here we are talking about climate models, or models of a specific ecosystem that are used to evaluate the environmental impact of some sort of action. There, the situations of management and of regulators are exactly like those in the financial services industry. Data from the past are scarce yet used to calibrate models. But who can guarantee that the data cover a reasonably broad range of what can happen in the future? Second, data are processed by models that only experts can understand and that can be built in different ways. The results of these models are used to set policies. And like in the financial industry, the only true test is the real-life application, which is dangerous because the policy effects can be observed only with a time delay, and unwinding potentially negative effects of the policy comes with a significant cost to society—tax money to clean up the environment or to bail out systemically important financial institutions. Thus, our definition of model risk above can be applied completely to environmental modeling. It is no surprise that this field has brought forth literature on model validation [5].[1]

Of course, the banking industry has to deal with model validation as well [6]. And there are parallels between some of the investment risks faced by banks and investment risks faced by insurers that focus on life insurance and pensions. But even where investment risks overlap, there are major differences in time frame—the insurance models usually deal with much longer time frames—and in liquidity considerations—insurance companies do normally not hold a trading book—that make banking models less applicable to insurers. And banks are not exposed to insurance claims risks such as disaster-related property damage, health care costs, or longevity. Further, insurer models must deal with issues such as

[1] The summary at the beginning of this book serves as an excellent introductory reading for many aspects on which we elaborate in this paper in the context of risk models for insurance companies. The list of key validation issues on p. 145 resembles our definition of model risk.

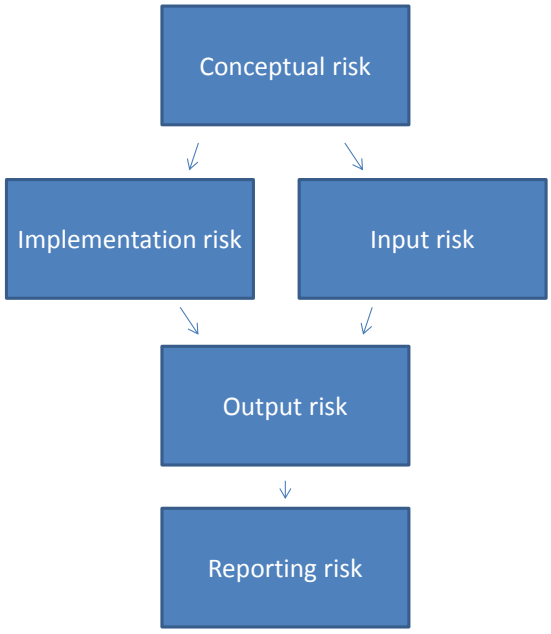
limited data availability and lack of standardization. All of these differences mean that insurers should not blindly adopt the risk models used by banks. The model validation process can help determine when and whether a model used elsewhere can be used appropriately for insurance risk management.

Process Guide

Commonly accepted standards for a validation process do not yet exist; most often they are just referred to as validation principles that essentially express that the range and rigor need to be in line with the potential risk. But this leaves far too much room for interpretation: What kind of documentation does a validator have to check, of what does such a check consist, and how is a validation process structured and documented?

The validation process should follow the major subcategories of model risk mentioned above. For an efficient process, it is important to notice that dependencies are found between the subcategories:

Order of model risk validation



The consequence of these dependences is that if the validation for a risk subcategory fails, then there is no need to validate the dependent subcategories. If the model is split into submodels, then this argument applies for each submodel. It is important to notice that if the model is broken up into submodels, the aggregation must be considered as a submodel that has to be validated according to the same standards. Even if all the submodels make perfect sense, it does not imply that their aggregation can be done consistently.

Many of the explicit steps in the validation of conceptual, input, implementation, and output risk have been mentioned previously [14].[2] The validation process for each of the subcategories of model risk is discussed in a separate section below. Within these sections, the paragraphs marked with check marks can be viewed as a checklist of action steps to perform during the validation.

Conceptual Risk

Conceptual risk cannot be evaluated without knowing the purpose of the model.

Purpose of the Model

Often we see that the purpose of the model is treated as an introduction to the model documentation: a brief statement of reminder why the model is being built. We consider this inadequate, because no validator can assess the suitability of the modeling concepts without knowing how users will apply the model in decision making. Thus, we believe that it is imperative that everybody on a validation team is well informed about the intended users and their use of the model's outputs.

Rather than describing the application setting in general terms, we recommend the documentation of the purpose of the model should do the following:

- ✓ Include reference to the relevant sections from underwriting, investment, or risk management guidelines or other appropriate documents that make reference to the use of the model output. It should be considered a warning signal if it is not clear from these texts which decisions the model outputs support and to which degree the users can supersede or augment the model outputs with qualitative information or judgment. Model outputs serve as quantitative decision support. Management needs to include other sources of information, apply judgment, and possibly challenge model results. By no means do we suggest automated decision making based on model results, but the role of the model and the role of judgment in the decision process have to be documented.

- ✓ Include a description of the users: At this stage of the validation, we need to verify only that the reports are addressed to a well-defined audience. It is not yet necessary to assess whether the information provided in the reports gives meaningful decision support to the users. The latter is assessed later under the reporting risk/use test section, where we explicitly assess the presentation of the content to decide whether it is appropriate for these users.

From a regulatory point of view, the purpose of the modeling exercise is also important. Regulators want to know whether the model results have a significant influence on the business. The argument is that the company invests more into modeling the risk situations appropriately if its business depends on it. CEIOPS, for example, states in their advice for level 2 implementing measures on Solvency II [8]: "Insurance and reinsurance undertakings shall demonstrate that the internal model is widely used in and

[2] The section on reporting risk in this report is significantly extended compared to this previous work.

plays an important role in their system of governance." Obviously, the regulators should not be relying upon the model if the management is not, so they apply a "use test."

This introduces a paradox that we have not seen addressed by regulators. The regulators want companies to use their risk models widely to ensure that they take care to build good models and thus hope to minimize the model risk. But if a model is used too widely, then even a modest amount of model risk could be devastating for the company because too many decisions depend on the model's output. Even worse, if the regulators prescribe a specific model (e.g., the standard model for Solvency II), then if all companies must use the same prescribed model, model risk can lead to systemic risk.

We believe that the validator should check whether the model is in use, that is, the two checks mentioned above, but should not seek to enhance the breadth of model application because any application outside the originally intended purpose constitutes model risk. Thus, one must be careful when interpreting CEIOPS's encouragement to use a model as widely as possible; it is more important to ensure that the model is used wisely and appropriately. Expansion of use may require changes or enhancements to a model. Companies need to recognize that there can be a tradeoff between *widely* and *wisely*.

Concepts and Their Limitations

Unless a standard model is being used like the one for Solvency II, two fundamental conceptual challenges are faced:

1. Which risks need to be modeled?
2. Which modeling methods should be applied to model them?

It is important that the question of which risks are material enough to necessitate being modeled is answered first and independently of the modeling method. Often we see the first question being answered implicitly by the choice of the modeling methods. We consider this to be incorrect because it is in reverse order; the starting point needs to be the identification of the risks that should be modeled. Choice of methodology comes later.

Choosing the risks that need to be modeled is absolutely central in terms of model risk. Some authors propose subjective risk ranking or materiality tests [2]. We consider such tests problematic because the application of risk statistics to linearly order the risk factors does not recognize that some risks influence the model more than others. This ordering of risks (e.g., by coefficient of variation or standard deviation, or by measuring how heavy a tail of the distribution is) assumes that all risk factors have the same influence on the model. If different statistics are used for different risk factors, then it is unclear how they can be compared; or if a different statistic is applied, then the order might change. The choice of which risk factors need to be modeled must be made by an expert. Consequently it is important that the rationale behind the choice of risks is well documented. Once the choice of which risks to include is determined, the model's output will not reflect any risk that was excluded. This may seem trivial, but it is an important statement. Failure to spend time to select the necessary risks at this stage could lead to surprises later that require more time and energy to fix.

- ✓ Check whether a process is in place to determine which risks need to be modeled. A warning signal has to be raised if the modeled risks are implicitly determined by the risk model methods or by a linear ordering of risk derived from test statistics, as this might indicate that technical people have started modeling without involving business people who are more familiar with the practical reasons why the model is being used in the first place.

It is important to note that once this step is completed, all the following steps of the validation refer only to the modeled risks. Thus, when the materiality question is raised again in the input and output risk sections, it is not the same as the one raised above. The materiality questions that follow are all conditional on the risk selection decision. Consequently, validation approaches that do not separate clearly between materiality in the context of risk selection and materiality in the context of model output sensitivity are bound to be imprecise in answering the materiality question.

Most new models are built largely by recomposing concepts in existing models. There is a value in using existing modeling concepts because their advantages and limitations are already understood. The focus of the concept validation should be to do the following:

- ✓ Check whether the concept documentation makes reference to external sources. If existing modeling concepts are used, then they should be documented by making reference to a publically available source. A warning signal should be raised if a company claims that all of their modeling concepts are proprietary.
- ✓ Check that the concept documentation describes how the modeling pieces are connected and why they can be used together. Sometimes this is falsely referred to as the aggregation. The aggregation model documentation must contain arguments of submodel consistency, but this is not the only place where this issue arises: Any subsequent processing units in a model must ensure that the output of one part of the model can be consistently used by the following one. For example: If a company decides to model claims on an annual aggregate level and for the same model describes a detailed nonproportional reinsurance model, then this is inconsistent.
- ✓ Check whether there is enough emphasis on describing the limitations of the concepts. It cannot be overemphasized how important it is to validate that the limitations are documented. First, it is important that the business users are informed about the limitations, which means they should be mentioned in the report. Second, it is important for the people who implement the model; they are usually less familiar in modeling concepts and thus will not know the limitations that have to be checked during the implementation. If the model limitations are not documented thoroughly, it is likely to lead to major implementation and reporting risk. For example, if the methods do not provide tail dependence, then such a model will not produce any tail dependence, and this must be directly stated. This may sound trivial, but if this restriction is not explicitly stated, a user could easily fall into the trap of interpreting the excellent diversification in the tail as a business effect rather than a consequence of a model limitation. Another example is as follows: If an economic scenario generator does not provide claims inflation, or if the model does not use the claims inflation that it provides, then the user must be made aware of such a limitation.
- ✓ Check vendor model concepts: If vendor models are used as submodels of the risk model, then the modeling team has to rely on the vendor's information about the modeling concepts. The documentation of limitations is particularly challenging when working with vendor models

because vendors are reluctant to disclose the limitations of their products. If a vendor does not provide documentation of the limitations, then the modeling team has to derive these themselves or collaborate with a service provider that has experience with this modeling approach. In any event, the validation needs to verify whether vendor modeling concepts and limitations are declared as vendor provided or self-derived. A warning signal should be raised if a substantial part is vendor provided. If it appears that the limitations of the vendor model have not been reviewed, this is yet a more serious red flag. In this case the validation team would need to verify that the modelers understand the models well enough to accomplish their objective. This can be a challenging task given that most validation teams are small in comparison with the modeling teams, and thus it is not easy for the validation team to match all the skills of the various modeling teams.

Implementation Risk

The implementation of risk modeling is software based, and thus, a realistic assumption is that it may contain errors. Implementation risk is quite likely the most underestimated model risk. Most actuaries and risk managers want to implement their models in their departments, which are usually not staffed with IT professionals. Their argument is that the risk management domain knowledge is more important than the software engineering knowledge. We believe that they are both equally important. The notorious love that many risk professionals have for systems in which they can easily change the models, such as Excel and many of the broadly accepted actuarial modeling tools, stems from their lack of knowledge of modern software engineering tools and the lack of collaboration with people who are in command of these tools. It is well documented that the implementation risk of Excel-style systems is heavily underestimated by the business domain experts who implement them [7].

This is surprising in the light of substantial research and experience from the software industry [7]: End-user computing systems, such as Excel workbooks, are very hard to test and validate in a professional way. The major issue is the application and compliance with best practice software engineering methods.

In many situations, the models that have to be validated are available for detailed inspection. This includes both the documentation of the algorithms and the computer code used for implementation. Nevertheless, our model validation approach will not involve a direct review of computer code because in many situations an analysis of the software at the code level is unrealistic; it is simply too time consuming. In fact, code review would necessitate a benchmark implementation in the form of an independent test implementation or an extensive independent test suite.

The focus of the implementation validation should be to perform the following:

- ✓ Check that the risk modeling experts have been involved in the selection of the algorithms that implement the modeling concepts. This is not an issue if the risk modeling experts have implemented the model themselves. This is a development process issue: The validation should check that the modeling experts have signed off on the algorithms, not just the concepts.
- ✓ Check whether everything in the model development is versioned: This includes the model code, the reports, the test cases, and the test reports. Versioning only the productively released software

is insufficient because it may well be necessary to unwind a change that did not produce the desired result. A warning signal has to be raised if the versioning is missing or was performed by manually renaming files. The development process requires frequent changes, and a manual process is far too error prone and unreliable.

- ✓ Check whether the accountability for code changes, bug fixes, or improvements is clear: A change needs to be initiated by an authorized person, coded by an assigned programmer, and assigned to a tester. All three people or teams must be visible in the issue tracker or the code versioning system.
- ✓ Check whether there is an automated test procedure in place that is run in regular time intervals or, even better, after every new version is checked into the versioning system: A warning signal should be raised if a risk model is only manually tested, because such tests are too infrequent and usually focus only on the new features or new bug fixes. Checking that the most recent changes did not affect already correctly running parts of the model is often neglected.
- ✓ Check who has specified the test cases: The test cases must be specified by the business domain experts and not the implementation specialists. A warning signal has to be raised if the implementation experts have specified the test cases.
- ✓ Check the test coverage: Software is available to build test coverage reports. Ideally, the developers provide these reports to the validation team. In the absence of automated testing, the validation team's only way of assessing the level of testing is to check the existence of test protocols and to verify that enough time has been allocated to testing. It should be noted that this simple time check can count only as a superficial validation of the implementation risk.
- ✓ Check the test content: The validation team has to decide which test cases are checked in full detail. Ideally, it should be those that have a major impact on the output. But this is not always easy to figure out, because it may well be that a major error in a minor routine has an equally bad effect as a minor error in a major routine. For this reason, it is important to verify that the selected test cases are thoroughly implemented and that the other procedural items listed here are strictly followed.
- ✓ Check the limitations of the algorithms: Strictly speaking, this is covered by the previous point. One very frequent flaw of the test procedures is that only positive tests are formulated; thus we mention it here as a separate point. Positive test cases check that if the input parameters are correct, then the output is correct. Algorithms have limitations, and those should be in line with the conceptual limitations formulated under conceptual risk. Negative tests must be formulated to ensure that the software catches situations in which the parameters are outside an admissible range. In these situations, it has to be verified that the software warns the user or even blocks the calculation. This sort of test is often deemed unnecessary by expert users because they are convinced that they will not enter parameters that are not admissible. But given that these models have many parameters—in the case of the Solvency II standard model, several hundred—it is unlikely that an expert can keep track of and maintain a good overview of the complete set of parameters being used for a run.
- ✓ Check whether the process of automatically loading data from source systems is tested by integration tests. Again, strictly speaking, this is also covered by the general checking of the test content. But the nature of these tests is very different. During development, the developers usually work with mock-up data to be independent of the source systems. These tests have to be

carried out by IT professionals. The validation team simply verifies that they were carried out and that the failure rate was acceptably low, and if failures happen, then the data are flagged and reported as such.

- ✓ Check whether user acceptance testing has been performed: User acceptance testing is very different from the other forms of testing. It cannot be automated, and it is usually performed only on stable versions that are intended for release. Users may have tested real-world scenarios or theoretical worst cases or just verified that the average outcome of a model coincides with their expectation. It is important that it is well documented in a test protocol that records the tests the users performed. Often no user acceptance test protocols are available. In this case, the only thing that can be verified is that the users have been given access to the test environment and that they have been given enough time to test the new model version. A warning signal has to be raised if the users were not involved before the release of a new model version or if there is no documentation of user acceptance testing.
- ✓ Back-testing should be performed when possible. However, the approach to back-testing can depend on the nature of the business being modeled and the model itself.
    - Some insurance models deal with very low-frequency data, making back-testing of stochastic results less useful due to the wide range of variability in results, for example, the occurrence of natural catastrophes. What can be done is to switch the model from a probabilistic one to a deterministic one. If it is possible to run the model with old realized asset returns, loss ratios, and explicit large losses, then one can check whether the model produces profit and loss (P&L) figures in line with the realized ones. Two things should be considered with this approach: First, the model must produce P&L figures, and many risk models still do not provide them. Second, passing this test verifies that the business mechanics of the model are most likely correct, but it does not tell us anything about the probability of extreme losses and extreme events that could threaten the solvency of a company.
    - Some models, especially those for life insurance companies, deal with business where the range of results has been comparatively narrow and there is a clear underlying trend line. In the context of these models back-testing can be especially useful because the model should be able to reproduce the trends of multiple measures such as premiums, claims, amount of insurance in force, expenses, and baseline investment yields.

It may be surprising that only a fraction of the suggested checks deal with content, the large part deals with process. Applying sound software engineering techniques and extensive automated tests cannot guarantee error-free software. But it can substantially reduce the implementation risk.

Input Risk

The principles for the validation of the inputs can be expressed very simply: internal and external data must be demonstrably appropriate, accurate, and complete [13]. This is easy to understand, yet it is very hard to derive explicit guidance for the validation from this principle. Another problem with the definition of this principle is that the three terms characterize partially overlapping issues that sometimes move in

opposite directions: more appropriate does not necessarily imply more accurate, more complete does not necessarily imply more appropriate—it can be the opposite. In addition, accuracy is very hard to quantify in this context. Thus, the formulation of this principle is not indisputable, and we would have preferred a principle that mentions consistency instead of accuracy or in addition to it.

The input data can be segmented into two classes:

1. *Raw data:* Data from a source system used as input for the model without processing the data.
2. *Calibrated data:* Parameters or input for the risk model that have been derived from source data by means of clustering to reduce the amount of data, by a statistical procedure such as distribution fitting, etc.

- ✓ The validation needs to check that the model input data and parameters are clearly assigned to one of the above two classes. A warning signal should be raised if raw data are directly loaded into the model and can be edited by model users without leaving a trace because this would mean that calibrated data are declared as raw data. For the calibrated data, the validation needs to verify that the data source is well defined, the calibration procedure is documented, and the persons performing the calibration have the required skills.
- ✓ Check that raw data are being interpreted correctly. The definition and encoding of data elements from a source system should be checked against the way each element is being interpreted in the model. This is especially important in cases where company-specific codes are used to denote important information such as contract type or options elected. Sometimes data encoding in a source system can change without notice to modelers.
- ✓ For raw data it has to be verified that the tool does not allow a user to edit such data. If it does, it creates implementation risk. We mention it here because the segmentation of input data into two classes was not mentioned in the section about implementation risk.

Generally it is true that more data lead to better calibration. But in the context of risk models that often conceptually deal differently with different phenomena, it has to be verified that the data used for calibrating parameters are consistent with the modeling concept. In particular, in some insurance models, attritional losses and large or catastrophic losses are modeled differently and both of them differ from scenario testing. Although the first two are often modeled probabilistically, the scenarios are normally just a list of predetermined and fixed parameter values.

- ✓ The validation needs to verify that the calibration uses the data consistently: Data used to calibrate attritional losses must not be used again for calibrating large losses and vice versa. Hence, there must be a clearly defined threshold, and it needs to be verified that this threshold is meaningful. Depending on the integration of the scenarios with the probabilistic model output, the same issue appears between large losses and scenarios. If scenario losses are integrated with the probabilistic model results, then large losses that stem from scenarios have to be excluded from the large loss calibration.

Sometimes the use of conservative assumptions is seen as a good or acceptable method to reduce input risk. We disagree with this. The use of raw data does not allow the introduction of conservatism, by

definition. Conservatism can be introduced sometimes when data have been calibrated data, but this contributes to the opacity. We do not believe that there is any benefit to introducing conservatism at the input parameter level. A conservative approach to risk management may well be worthwhile to consider, but it is a management decision that leads to separate scenarios that are clearly set apart from the base calibration of a model derived from actual data.

Conservatively estimated parameters also render a comparison with a company's historic parameters difficult. Effectively, this practice constitutes a change in actuarial method. Comparison with industry benchmarks is also rendered useless. But both comparisons should be valuable instruments in a validation.

- ✓ Check whether any substantial deviation of input parameters from their values used in the previous reporting cycle has been explained. Parameters that stay the same with respect to the previous reporting cycle do not necessarily have to be correct because it could be that changes in the underlying portfolio require changes in parameters. Such changes are more difficult to detect and assess in a validation process. We recommend that the validation team ask for a summary of major changes in the source data used in the calibration.
- ✓ Benchmark the major input parameters against the industry's distribution of parameters or a peer group's selection: here we do not elaborate on how to determine which parameters are considered major. This will be dealt with in the section output risk. Benchmarking is a highly valuable aspect of the validation. We propose that if the parameters chosen are in the interquartile range (i.e., the range from 25 to 75 percent) of the benchmark distribution for this parameter, then the validation should check only whether the calibration procedure is well documented. If the parameter value lies outside the interquartile range of the benchmark distribution, then the validation should be more rigorous: In addition to verifying the existence of the calibration documentation, we consider it necessary either to check for a more detailed explanation of the parameter values or to verify the actual calibration procedure.
- ✓ Many input parameters of risk models cannot be derived solely by applying statistical methods. In such cases the validation needs to verify that an effective peer review process is in place and was appropriately conducted by qualified reviewers.

The above guideline puts substantial emphasis on benchmarking. Model builders must often complete their work with little or no benchmark data. Although benchmarking is valuable for validation, significant research on benchmarking still needs to be carried out: Should overall industry benchmarks be considered, or should they be further segmented? If so, what is a useful level of segmentation of the overall industry? Which time frame should be considered to construct the benchmark distributions?

Output Risk

We would like to remind the reader that before the output risk can be assessed, first, the implementation and input risks must be assessed and deemed acceptable. Validation of output risk should be an additional and subsequent process. The assessment of output risk needs to check whether skilled people can interpret

the model outputs in the context of their intended application. Presentation and communication of these outputs to business decision makers are dealt with in the reporting risk section.

Before any interpretation takes place, we need to make sure that everyone involved is clear about the outputs and exactly what they refer to.

- ✓ Check that the correct input data set and model version are referenced by the outputs. During an operational run there will be multiple input datasets in use and potentially even multiple versions of a model. It is thus imperative that proper data management is in place: Output data must reference the input data and the model used. The input data must be locked as long as the output data are available to users. A warning signal has to be raised if this is all done manually.
- ✓ Check whether the outputs can be reproduced. For deterministic models, this means verifying that nobody can edit or delete the input data sets if they have been used to produce outputs. Although this also applies for Monte Carlo models, it is significantly more complex in this situation. Obviously, the output data set should reference the random number generator seed(s). If the application runs in a distributed environment, then this computing environment might have to be referenced in the output set as well. Most implementations of distributed Monte Carlo simulations do not guarantee reproducibility across different compute environments. Some risk models are more compact and run in only one environment. For newer, bigger, more detailed models, it is likely that at least three computing environments will be used:
    - The developers use their laptops
    - The testers use a server and
    - The operational run might again use a different, more powerful server environment.

  Note that it is possible to construct distributed Monte Carlo simulation methods that do not depend on the computing environment. But it is beyond the scope of this paper to explain these approaches.

- ✓ Check whether breaches of input parameter limits are indicated in the output. Good reasons may be given to run a model with an input data set that contains some breaches of parameter limits. These include checking whether the input parameter limits are highly sensitive, or demonstrating that the lack of such limits could lead to outputs that lend themselves to misinterpretations. Such output sets need to be clearly marked.

Once the above operational issues have been checked, then the validation of the dynamic behavior of the model must be verified. This is the most demanding job in the validation process, because it requires technical expertise and business understanding. The overall goal is to check that the outputs are meaningful. It is generally accepted that a measurement of the sensitivities of the output key figures with respect to the input parameters yields good insight into the dynamics of a model. This is only partially true. Even if all derivatives of a function like the economic capital are zero or very small, there can be significant sensitivities with respect to joint movements of input parameters. No mathematical procedure exists to determine which sensitivities should be calculated. Usually many parameters are found in a model. This makes the computational time of measuring sensitivities with respect to all input parameters

quite costly. The computational time increases exponentially when we compute the sensitivities with respect to simultaneous movements of multiple variables. Business knowledge has to be applied to propose which sensitivities are interesting to study. This introduces output risk: Even if the model were a perfect fit of reality and the input data were absolutely accurate and complete, it could still be that the outputs of the model are misleading its users. This could occur if users were unaware of the dynamics created by changing some parameters. How the dynamic behavior of a model can be explored and verified is still a matter of ongoing research. In the reporting risk section of this paper, we discuss one data exploration technique that can be used for this purpose.

The validation should include the following checks:

- ✓ Check whether documentation exists concerning the selection of input parameters against which the sensitivities of the outputs are measured. People with business knowledge should be involved in this part. All involved parties must understand the critical importance of selecting the input parameters.
- ✓ Check whether the sensitivities are documented. If the sensitivities are given as estimated values of first derivatives, then the meaning of such derivative estimates has to be explained thoroughly. We recommend the graphic visualization of the sensitivities.
- ✓ Check materiality of input parameters based on the sensitivities. The sensitivities provide an important feedback loop between the outputs and the inputs. The input parameters, which are highly sensitive, need to be estimated more carefully, documented with more detail, and their limits observed more strictly.
- ✓ Check whether the ranges of the output key figures are made available. Given that most input parameters cannot be estimated exactly, the validation needs to verify that the uncertainty of the output key figures is explicitly communicated. This can be done using confidence intervals around the point estimates.

Apart from measuring and interpreting sensitivities, benchmarking can yield some insight into the validity of the output. Here we are not talking about the benchmarking of input parameters, but rather benchmarking in the sense of comparing the outputs with outputs of other, usually simpler models. After all, it is very likely that regulators are doing this as well. Candidates for benchmarking models are standard models provided by the regulator, models used by the rating agencies, or simpler models created by the validation team. The expectation is not that these simpler models yield the same output; otherwise it would not have made sense to build the more complicated, next generation model. But experts should be able to explain the differences between the outputs of different models, especially the new model and its predecessors.

- ✓ Check whether benchmarking models were used to validate the output. Experts should document why their model yields different results than a benchmarking model. Note that this is almost like presenting a value proposition for the internal model. If the model is not an initial version, then output from older model and input data versions might be available for comparison. Understanding the new model in relation to an older one may also help. This is usually referred to as an analysis of change. There are two typical forms of analysis of change:

1. Running a new model version with an old data set and comparing the results with the results obtained from the old model version. For a better understanding of the effects of the model changes, the various model changes are applied one after the other.
2. The second form of analysis of change is the same procedure, except that it is applied to changes in the data while the model is kept the same. Consequently the various changes from one input data set to another are applied one after another and their effects are measured.

The problem with analysis of change is that it is not independent of the order in which the changes are applied. Hence, if it is used for validation purposes, then the documentation must include arguments concerning a meaningful order of changes.

- ✓ Check that the analysis of change starts from a validated model and input data set. The order in which changes are applied must be documented and must include arguments why a particular order of changes has been chosen

Reporting Risk

This is the final step in a validation and the most crucial. Yet some people consider this step to be outside the scope of model validation because it is closely related to the use test. If this step is not included in the validation process, then the model's use in the intended real-world application cannot be assessed. We consider this to be the most essential point.

In this section, we assume we have already resolved the issue of which key figures must be presented in the report. Still, we have to validate that the way in which they are reported—the manner and order of presentation—is not misleading. The report provides quantitative decision support: the intended use and the addressed audience both matter. Reports are all about communication. The way in which the information is framed and presented influences the decisions that are based on it. This is well known and has been researched by well-known authors such as Kahneman and Paulson. In particular, if probabilities are involved, such as VaR, or scenarios have to be assessed, then it is shocking how many people draw the wrong conclusions due to the *framing* involved in the presentation. Risk communication is quite possibly the most difficult job of risk managers, and thus it should come as no surprise that assessing the reporting risk is the most challenging part of a validation. The validation team members may be affected by misinterpreting the report, while having to assess how others might misinterpret it. It is important to notice that this is not an assessment whether the addressed audience is fit for their job. A regulator might have to assess that, but the model validation team considers the management as a given and assesses only whether the decision support they are getting is suitable.

This communication challenge is amplified by the fact that the people who build models and review them have a background that is almost exclusively quantitative (actuaries, financial engineers, et al.), and the people on the receiving end of the report often have a more limited quantitative background. The problem

of innumeracy is immense: the end users need to understand the report and communicate it to their fellow managers who may be less involved in risk issues.[3]

The validation team can and should ask the business users whether they consider the reports to be useful and meaningful decision support. But the validation team cannot rely on the users' opinion to the exclusion of all else. User feedback is asymmetrical for more than one reason: A user who is affected only to a limited extent by the model's outputs is inclined to provide weak feedback. But this does not imply that another user might not be affected significantly by the model's output. In fact, users that are significantly affected by the model's output are also inclined to provide more feedback: These users often mix the report's usefulness with the effect the report's numbers have on the user's business. If the model results are favorable, then these users are inclined to give a more positive review than if the model outputs are putting the manager's unit under pressure. This situation may be more obvious when the compensation of managers is tied more closely to the results produced by the report.

The validation has to check a few operational issues before the more complicated risk assessment is done:

- ✓ Check that the reports clearly state which model and data version were used.
- ✓ Check that business users are made aware of situations in which some of the parameters are outside a comfort range or even outside the agreed limits.
- ✓ Check whether the frequency and timing of the reports is in line with the decisions they support. It must be stressed that this has nothing to do with an automated decision process.
- ✓ Check whether the results are communicated using institutionally accepted metrics that are readily understood by all end users. Metrics that may capture or describe the risks well, but are not commonly known or used in the company, introduce reporting risk.
- ✓ Check whether the report uses any means to convey how robust the key figures are. Simply providing point estimates of the key figures does not give enough information to decision makers. They have to be made aware of the fact that estimation errors for the parameters and different modeling assumptions yield ranges of outcomes.

Other ranges that have to be communicated are the ranges of normal business volatility. Even if the intended application is to decide about the required economic capital, that is, the extreme outcomes, it is important that the business users know the normal business volatility. Extreme situations cannot be well understood without knowing what is considered normal. Interquartile ranges can be considered a good measure of normal fluctuation.

Risk Management Control Cycle

The development of a risk model may have taken many employee-years, and so it is clear that a proper validation could take a significant amount of time and involve a number of professionals. For larger models or organizations, it is likely that the validation process stretches with various levels of intensity across several reporting intervals and includes temporary approval from regulators or management. Thus, it interweaves with the risk management control cycle. This is a substantial challenge for management.

[3] A nice summary of the issues involved in innumeracy and communication can be found in chapter 4.3.6 of Franzetti's book [12].

The portfolios and the parameters change while the model is being validated. In this section, we discuss the embedding of validation in the risk management control cycle and propose how transparency can be increased to reduce the problems of validating a moving target. For all these reasons it is important that the validation process is clearly structured and documented to ensure that a common standard is being kept and that there is no gap in the process.

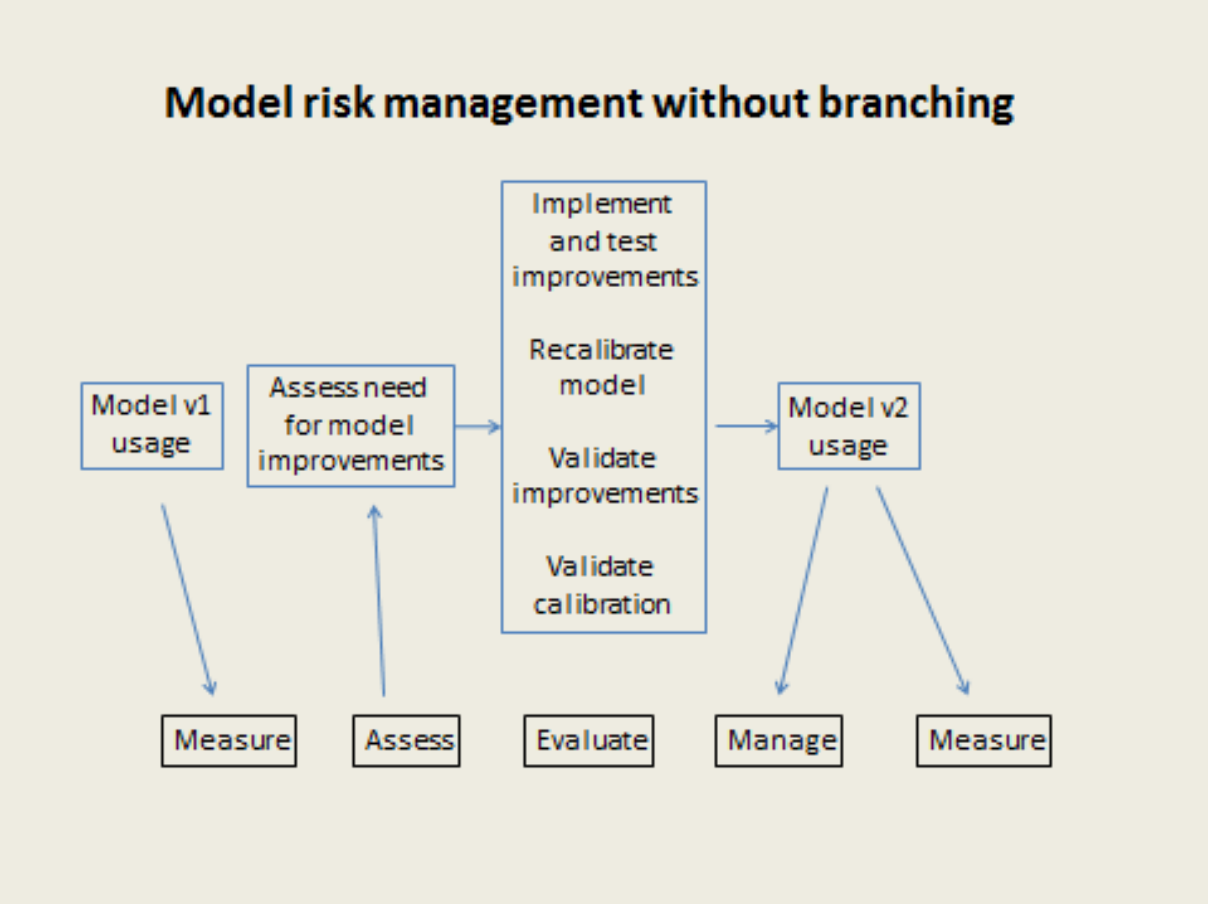
Finally, the validation process is an ongoing effort. Even if the risk model is kept stable, there will certainly be new input data, and thus at least the input risk has to be regularly reassessed. A standard process should be in place to check the validity of input data, such as reconciliation to key control totals or analysis of changes. At some less frequent interval, a validation team should check whether the regular input validation process is still effective.

More likely, because of resource constraints, an initial validation will not cover all model risks equally well. All this makes model validation for regulators and management a moving target, and it is highly important for risk management to know the status and level of the validation.

Normally the risk management control cycle is depicted with the following phases: Assess, evaluate, manage, measure, and then restart. These phases can be mapped to the process of model development, validation, and use as follows:

- *Assess:* Determine the need for a model (or model changes and improvements)
- *Evaluate:* Design, implement, and validate the model (or changes and improvements)
- *Manage:* Use the model results to help make decisions
- *Measure:* Report financial results and compare with any model expectations.

This mapping is illustrated in the figure below.



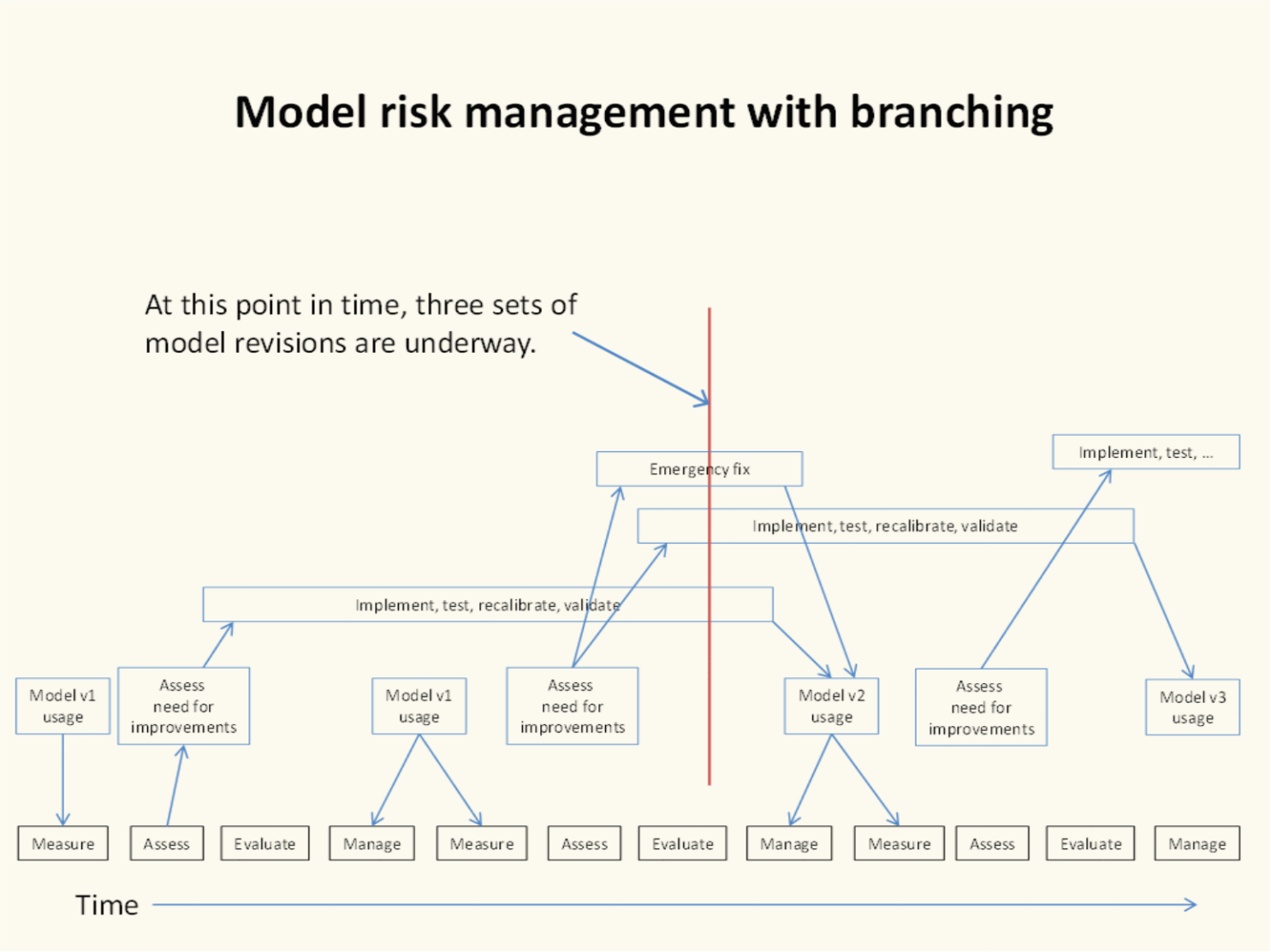
This figure makes it clear that model development and validation both fall within the evaluation phase of the risk management control cycle. In a real-world setting, the time required to implement, test, calibrate, and validate changes to a model may be longer than one cycle allows. This is especially true if the "Measure" phase of the cycle is equated with external financial reporting. The time required to implement, test, calibrate, and validate model changes can easily exceed one financial reporting cycle. Therefore model changes that are initiated in one cycle may not be usable until several cycles later.

This can become even more complex when the "Assess" phase of different cycles leads to different model changes. One can experience a situation where several changes are in the process of implementation, testing, calibration, and validation at the same time and are in different stages of that process. This is especially likely to happen when models are used in a regulatory setting, because the need for changes can arise both from regulators and from management.

This complexity is well known in software development and is usually called branching. To maintain quality, the organizational structure should not be too complex. Insufficient quality will result in rejection by the validation team, additional development cycles, and, thus, increased costs. Unfortunately the organizational structure for the development of software that involves branching necessitates the concurrent development of more than one version of the same product. This requires either multiple teams

or an extreme organizational discipline and significantly more management oversight. It has been stated in the literature [9] and empirically verified [10] that the failure rate increases significantly under these circumstances. The consequence for the validation process is the same as for the development: In a setting with branching, it becomes more complicated and time consuming and, thus, more expensive.

The following figure displays a relatively simple branching situation. At the point in time indicated by the vertical line, three sets of model revisions are underway. One is nearly finished with validation, another is just beginning implementation. A third represents an emergency fix that may relate to discovery of an error or a regulatory mandate. Organizationally, one needs to ensure that the people implementing and validating each set of revisions have knowledge of the others, even though the others are not yet validated or in use. Also, management needs to understand the status of various model enhancements when several are in process at the same time.



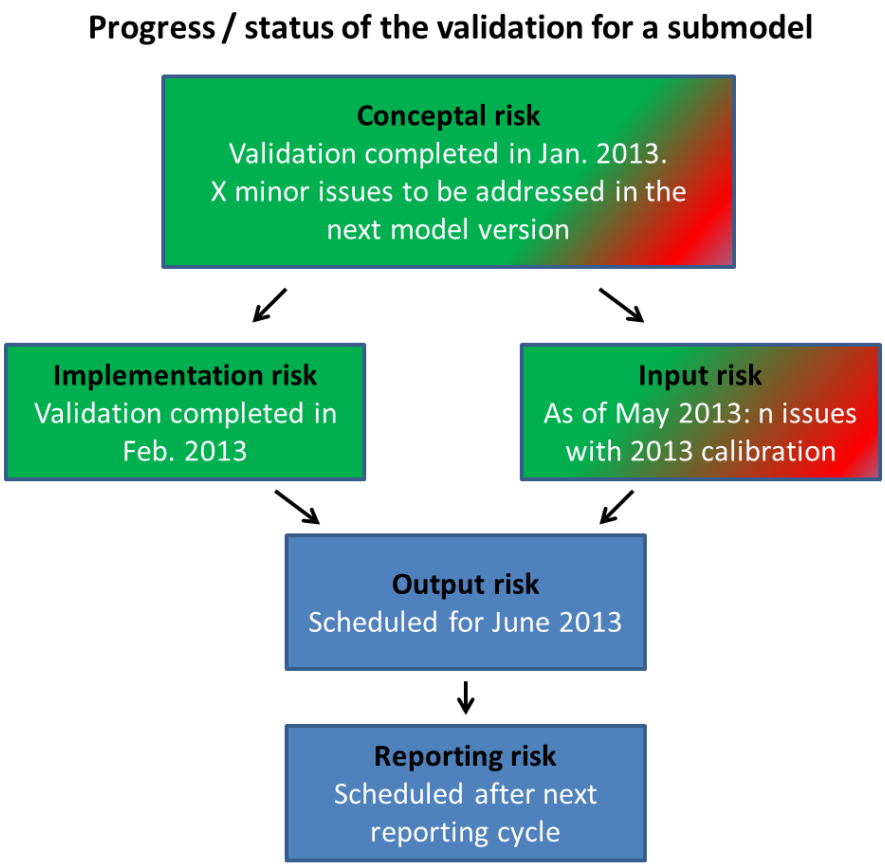
Management of a validation process in a branching environment involves managing several dimensions or aspects simultaneously. This is very hard in a normal, static management reporting environment. Issues are usually associated with several aspects, for example, the submodel, the model version, and the software version (note: these latter two versions are usually not identical), whether it is a new feature request, an improvement of an existing one or a fix for a problem with an existing feature, the severity, the deadline, the reporting line, etc. Depending on the purpose of reporting and to whom the report is addressed, we have to be able to slice and dice the validation issues according to multiple criteria. In addition, the resolution of issues has to be documented and saved with an audit trail, otherwise the validation of future incremental model changes needs to start from the beginning again. We strongly recommend the use of an issue-tracking system for this purpose. Ideally the issue-tracking system is shared among the development team and the validation team. This would provide both useful information and an excellent audit trail from the point of feature requests to implementation and finally to validation.

The validation of vendor models should follow the same lines and standards as internally developed models. Vendor models can be chosen to save implementation time, but to avoid conceptual risk a vendor model needs to be understood. A validation from another party should not simply be accepted. The intended use of the model output plays a central role in the validation, and it might be different than anticipated by the vendor. So, if a third party validated such a model, it must have done it assuming a well-described use. In the context of natural catastrophe models, it is quite possible that one of the uses may be pricing rather than economic capital modeling. Catastrophe models are by far not the only submodels of an economic capital model that might have different applications elsewhere in the firm. Interest rate models suffer from the same problem: models used for product pricing may have to be arbitrage free, while those used for economic capital modeling might not.

Presentation and Communication

The deliverable of the validation process is the validation report. If the model is big, it is likely to be composed of submodels, each of which will be validated separately, most likely by different teams: for example, asset models, natural catastrophe models, or economic scenario generators. Therefore it is important that a common terminology, liked the one suggested below, is used and that the structure of the various reports are similar. The validation report is not to be confused with the model documentation. It is part of the validation process to check the model documentation.

We suggest that the overview of the progress or the state of a validation is being reported to management by using annotated graphs like the one below for each submodel.



The validation report has to document which areas have to be validated more carefully in subsequent validation cycles. We recommend classifying the depth of the validation in three categories:

1. Superficial, further validation required: In this category, it should be mentioned whether the reason is a time constraint or a skills constraint.
2. Adequate, no further validation effort required: In this category it should be mentioned under which condition the validation has to be renewed, such as a major model change or recalibration of its input parameters. If skill is required to run a model, then the condition may also be that the team that operates such a model does not change.
3. Adequate, but ongoing validation required: This could result from a situation in which the model is fed with different data in every reporting cycle, for example, macroeconomic scenarios. This can also be the case for an ad hoc model that is likely to be changed in every reporting cycle, such as models for small portfolios of highly customized products for which it would not be worthwhile going through the time-consuming process of revising the whole model.

This classification is not to be confused with the assessment, the result of the validation. The results should be classified as one of the following:

- ✓ Inadequate, requiring change or improvement, or
- ✓ Accepted.

We believe that any validation issue reported with these classifications is rich, yet simple enough, and can easily be interpreted by management. This should not be understood as a scoring system that produces an easy way to aggregate the individual assessments to yield a single score for the model.

Governance

Over time we believe that internal audit will be in charge of verifying that the model validation policy is implemented. This does not mean that internal audit carries out the model validation. It will be involved in formulating a validation policy and oversee that it is kept. The precursor of the pan-European regulator EIOPA, CEIOPS [8], developed a short list of what a validation policy should contain. It is noteworthy that the validation policy of CEIOPS mentions under the heading of governance that the risk management function should be responsible for model validation. In the same paper, CEIOPS stresses that independence within the model validation process is essential. This could be a potential conflict because risk management is often heavily involved in modeling. Consequently, if risk management is involved in both activities, then it has to be ensured clearly at any point in time which of the two activities people perform. In addition, the model documentation and the model validation report have to be kept separate.

A Case Study of AIG's Model of Credit Default Swaps

The following is a historical example of model risk failure of governance: insufficient validation that is linked to an incomplete statement of purpose. AIG is a prominent international insurance conglomerate that played a well-known role in the 2008 financial crisis. One of the main causes of AIG's problems was its credit default swap (CDS) portfolio. When AIG originally entered this business, it did so with reliance on models developed by Gary Gorton, a finance professor at Yale School of Management. Mr. Gorton's models were used to determine the level of fees charged for the CDS contracts. The models were based on large volumes of historical data concerning defaults and used a sophisticated mathematical simulation process.

So what went wrong? Why did the CDS business fail? What was missing in the model validation process that could have led AIG to make better decisions?

In this case, the model was used as justification for the ongoing financial operation of the business. However, the model was not validated in that context. An October 31, 2008, article in the *Wall Street Journal* [16] included this observation:

> AIG didn't anticipate how market forces and contract terms not weighed by the models would turn the swaps, over the short term, into huge financial liabilities. AIG didn't assign Mr. Gorton to assess those threats, and knew that his models didn't consider them.

Another observation from the same article:

> Mr. Gorton's models harnessed mounds of historical data to focus on the likelihood of default, and his work may indeed prove accurate on that front. But as AIG was aware, his models didn't attempt to measure the risk of future collateral calls or write-downs, which have devastated AIG's finances.

These statements make it clear that any validation of AIG's CDS model was done with reference only to potential claims, and not as a model of the full financial operation of the CDS business. The CDS contracts required AIG to post collateral if the value of the insured loans fell. The model did not include this collateral requirement, nor did it reflect the way that fluctuations in loan value could dramatically increase the short-term cost of meeting that requirement. The model used to run the business was incomplete because it did not simulate all of the material risks of the business. As a result, management did not maintain either the liquidity or the capital required to post collateral when the need arose.

A model validation process in this context could have alerted management to serious model risk due to the following considerations (or steps in the validation process):

- *Incomplete statement of the purpose of the model:* The model was not being used simply to set fees for the CDS contracts, but was being relied upon as justification that the full business was financially sound. The article quotes statements to this effect made by AIG management to investors.
- *Material risks not reflected in the model:* The model focused on fee collections and on potential payments for default. It did not reflect fluctuating collateral requirements or the effect of fluctuations in the reported value of the insured loans.

The lesson in this case is that the statement of how a model is to be relied upon is critical to effective model validation. An incomplete statement of purpose can lead to an ineffective validation and significantly increased model risk.

A Case of Failed Model Governance: JPMorgan Chase and the London Whale

JPMorgan Chase is the largest U.S. bank and is widely viewed as an expert in risk management. The bank's Chief Investment Office (CIO) plays a key role in the bank's risk management. In early 2012, the CIO created an unexpected loss of over $6.2 billion due to activities that were purportedly intended to reduce and manage risk.

The Permanent Subcommittee on Investigations of the U.S. Senate published a report on this incident, titled "JPMorgan Chase Whale Trades: A Case History of Derivatives Risks and Abuses" [15]. Although a great many issues are discussed in the report, it does make clear that the use and abuse of models played a significant role in activities that led to the loss.

The CIO used models and risk metrics (which are based on models) extensively. According to the Subcommittee report:

> The CIO used five metrics and limits to gauge and control the risks associated with its trading activities, including the Value-at-Risk (VaR) limit, Credit Spread Widening 01 (CS01) limit, Credit Spread Widening 10% (CSW10%) limit, stress loss limits, and stop loss advisories.

Unfortunately, it seems that the CIO often either ignored these limits or was able to manipulate the underlying models to change them. Consider these two excerpts from the Subcommittee report:

> During the first three months of 2012, as the CIO traders added billions of dollars in complex credit derivatives to the Synthetic Credit Portfolio, the SCP trades breached the limits on all five of the risk metrics. In fact, from January 1 through April 30, 2012, CIO risk limits and advisories were breached more than 330 times.
>
> Traders, risk personnel, and quantitative analysts frequently attacked the accuracy of the risk metrics, downplaying the riskiness of credit derivatives and proposing risk measurement and model changes to lower risk results for the Synthetic Credit Portfolio. In the case of the CIO VaR, after analysts concluded the existing model was too conservative and overstated risk, an alternative CIO model was hurriedly adopted in late January 2012, while the CIO was in breach of its own and the bankwide VaR limit.

It seems clear from this narrative that although models ostensibly played a major role in risk management, they were routinely either ignored or manipulated. Apparently the CIO was allowed to change a model if model results would constrain its activities. The CIO was in a dual role—as both user and validator of the model. This lack of independence between the user and validator contributed to a significant abuse of the models, and a financial loss ensued.

A Case Study in Model Risk: Universal Life with Secondary Guarantees (ULSG)

The Universal Life contract with secondary guarantees has been one of the most controversial life insurance products introduced in the United States. The reasons why these contracts were invented, and the pricing controversy surrounding them, can best be viewed as a case study in conflicting models, which is evidence of model risk. Model risk arose both for regulators, because the regulatory reserving model could be abused, and for insurers, because pricing assumptions that have proven to be aggressive could be characterized under some models in a way that sounded reasonable rather than aggressive.

ULSG products are basically permanent life insurance with flexible premiums, that is, "universal life." Under traditional universal life insurance contracts, premiums add to an account that accumulates with interest. Each month the cost of life insurance protection is deducted from the account value, and the contract remains in force as long as the account value remains sufficient to pay for the cost of insurance. Universal life with secondary guarantees is a variation on this product, adding a guarantee that the contract will remain in force as long as either specified premiums are paid or a "shadow" account value remains positive, regardless of the accumulated (main) account value.

Often these contracts are designed so that the minimum premium needed to keep the contract in force under the secondary guarantee is at a much lower level than the premium required to maintain a positive account value. One such design involves use of a secondary or "shadow" account value that is accumulated using higher interest rates and/or lower charges for the cost of insurance. The contract remains in force as long as the "shadow" account value remains positive.

The practical equivalent of these contracts is a term life insurance contract that never expires as long as the premium is paid. The difference between such a contract and a permanent life insurance contract is the practical absence of a cash surrender value. For many ULSG contracts, the primary account value (not the "shadow" account value) is accumulated using charges for the cost of insurance that are so high that the account value goes to zero fairly quickly, leaving a zero surrender value.

Again, the reasons why these contracts were invented, and the pricing controversy surrounding them, can best be viewed as a case study in conflicting models, which is evidence of model risk. Model risk arose both for regulators, because the regulatory reserving model could be abused, and for insurers, because very aggressive pricing assumptions could be characterized under some models in a way that sounded reasonable rather than aggressive.

Model Risk in the Regulatory Framework

One doesn't typically think of the regulatory arena as a place where model risk might arise. However, regulatory reserve and capital requirements are based on models, and these models are sometimes inadequate for their stated purpose. That is where the saga of ULSG products began.

Life insurance regulatory reserve requirements in the United States were (and remain) largely based on the concept of a "net premium reserve." This is a simple formula that was designed for products with fixed premiums and fixed benefits, and that works well in that context. However, ULSG is a contract with flexible premiums, and one in which policy owner behavior can play a significant role.

The model for calculating a net premium reserve is simply to subtract the present value of future premiums from the present value of future benefits. Both premiums and benefits are projected using mortality rates from a standard table as the single decrement, and a regulatory discount rate is prescribed.

This model served its purpose well for over a century. During that period, life insurance products generally remained within the limits of the model, having fixed premiums and benefits. However, late in the 20th century more flexible products were introduced, and mortality experience changed (improved) dramatically. Both of these effects led to model risk because the regulatory reserving model was inadequate to deal with them.

When flexible premium products such as ULSG were introduced, the net premium model needed to be enhanced to define the premiums to be used for valuation. The enhancement was to specify that for valuation purposes, the premiums used would be the minimum future premiums required to keep the contract in force. The idea was to maximize the reserve, because in theory the use of minimum premiums should minimize the present value of premiums that are subtracted from the present value of benefits.

Adapting to the problem of improving mortality was more difficult, because there were two aspects to the issue. Not only were general mortality rates improving, making standard tables out of date, but advances in medical underwriting and variances in company underwriting practices made the range of mortality experience vary ever more widely between insurers. One standard table could no longer be viewed as a reasonable approximation to experience for every insurer. Those insurers with the strictest underwriting and best mortality experience found that statutory reserves based on standard tables were higher than necessary, and the need to hold such reserves was adding unnecessary cost for their customers.

Unnecessary costs for the insuring public are of concern to both regulators and to insurers. The products were designed so that a literal application of the net premium reserve calculation led to very low reserves.[4] The regulatory reserving model was therefore inadequate. Regulators in some states agreed with insurers in allowing alternate reserve calculations. There was great controversy over this, and it eventually led to an effort to change the regulatory reserving model. Much effort has been expended in developing the new "principle-based" model, which is now well on its way toward adoption.

Model Risk in Pricing

In states where regulators allowed it, the pricing and reserving for ULSG products was based directly on estimates of the future cash flows and investment returns, using anticipated experience with a margin. The real model risk was in setting the assumptions regarding future claims and investment returns. Models play a role in setting assumptions, and model risk came into play.

A Limited Model for Future Mortality Rates

As previously discussed, mortality experience had improved, and current mortality rates experienced by many insurers were substantially lower than those in standard tables. If mortality rates for the current insured population averaged 75 percent of a standard table, it did not sound unreasonable to assume it would remain so in the future. Both pricing and reserving might be based on an assumed mortality rates equal to 75 percent of the standard table.

Such a model for mortality rates is limited because it does not reflect variation by age (or by sex or other variables). In this case, the ratio between current experience and the standard table was age dependent, with the ratio being lower during the working ages and higher during the retirement (older) ages. If this dependence were reflected in reserving and pricing assumptions, the ratio for a closed group of policyholders should be expected to increase over time as the group aged. A simple model that did not

[4] Recall that the net premium reserve for a flexible premium product is based on paying the lowest premium possible while keeping the contract in force. For Universal Life products, this amounts to keeping the accumulated account value near zero. Some ULSG products were designed so that if the account value got near zero, the charges against the account value increased dramatically. This meant that paying the minimum premium in early years led to the need for extremely high premiums in later years to keep the contract in force. Since the present value of future premiums was deducted when calculating the reserve, this made the reserve very low. Of course, any reasonable policyholder would pay more than the minimum in early years, thereby avoiding the high charges in later years. But that is an aspect of policy owner behavior that was not considered in the regulatory reserving model.

reflect age dependence could lead to inadequate pricing, especially for contracts issued to older individuals.

Of course, we have not yet discussed the likelihood that mortality rates will continue to improve over time. Recent improvement in overall mortality rates had been at up to 2 percent per year. It is not unreasonable to assume that some improvement would continue into the future. Pricing might reflect an assumption of continued improvement in mortality rates at perhaps 1.5 percent per year, a lower rate of improvement than recent experience in order to be conservative.

Such a model for mortality improvement is limited because again it does not reflect variation by age. The trend of mortality improvement has been studied extensively over time, and the rate of improvement has often been demonstrated to be age dependent, with faster improvement in the working ages and slower improvement at the advanced ages. To the authors' knowledge, all mortality projection scales developed in connection with standard tables share this characteristic. Therefore a simple model that projects 1.5 percent improvement at all ages might understate future mortality rates at older ages—the ages when most claims under ULSG contracts are expected to occur. Therefore such an assumption in a pricing model could lead to inadequate pricing.

There is some evidence that pricing using mortality models such as those discussed here actually took place. When ULSG products were first introduced, the prices at which they were offered were lower than competing products, especially at older issue ages. And some of the earliest revisions to such product offerings were to withdraw them from the market only for older issue ages, or to increase the premium rates offered to older customers.

Conflicting Models for Future Interest Rates

Although the premiums for a ULSG contract are flexible, in most cases they are not paid all up front but are spread out over many years. This means that the pricing must anticipate the investment return that will be earned on money to be received years in the future.

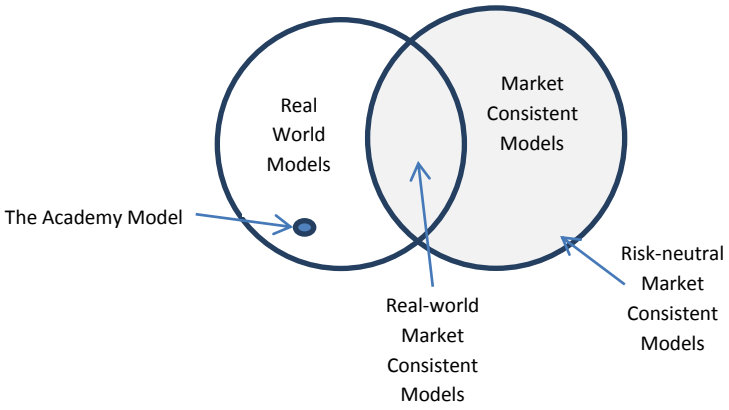
In recent decades a great many models of interest rate behavior have been developed. It can be tempting to simply adopt a widely used model and apply it in pricing, without an understanding of the original purpose for which the model was developed. This is model risk in one of its more subtle forms.

Consider one widely used model—the interest rate model developed by the American Academy of Actuaries (the Academy model) for use by U.S. regulators in measuring interest rate risk for purposes of setting minimum capital requirements. This is a stochastic model and was carefully calibrated so that the distribution of scenarios that it produces closely matches the historical distribution of interest rates over the last 50 plus years in terms of the proportion of high and low interest rates and by other measures.

One might be tempted to use this interest rate model in pricing. If one believes that the future will be like the past, then a model carefully calibrated to reproduce the distribution of historical interest rates seems to make sense.

Or perhaps not. The Academy model is intended to be used to measure the volatility of insurer financial results in the outlier scenarios—the ones at the tails of the distribution. The bulk of scenarios in the middle of the distribution do not matter in that exercise, and even the level of the middle of the distribution is not a primary issue. However, in pricing, the greatest probability weight is placed on the bulk of scenarios in the middle of the distribution, thereby putting the greatest emphasis in model results that are furthest from the intended purpose of the Academy model. Using the Academy model for pricing (rather than for estimating capital requirements) is not in keeping with the intended purpose of the model and therefore introduces model risk.

The universe of models for future interest rates fall into three main categories: real-world models, market-consistent models, and risk-neutral models. These categories overlap as shown in the diagram below.



In this diagram, the universe of market-consistent models includes risk-neutral models and some real-world models, but not all real-world models. Only the market-consistent models are appropriate for pricing financial instruments, or, for that matter, insurance. One characteristic that separates a market-consistent real-world model from other real-world models is its calibration. A market-consistent model is calibrated for consistency with market prices on a particular date.[5] A real-world model can sometimes be turned into a market-consistent model by recalibration.

The Academy model, as currently used for measuring C-3 risk (interest rate risk), is not periodically recalibrated. It is designed to replicate the general behavior of interest rates, but is not periodically recalibrated to reflect the market's perception of that behavior as it changes over time. Therefore the Academy model is not market consistent and should not be used for pricing of any contract where consistency with other prices in financial markets is important. At the time ULSG contracts were introduced, market interest rates were notably lower than the average of the future distribution of interest rates produced by the Academy model. Use of the Academy model for pricing would have been like making a bet that interest rates would rise to historical average levels—a bet that the market was putting long odds against.

In the case of ULSG, one might ask why the pricing of a life insurance contract would need to be consistent with pricing of other financial instruments. In response, one can cite the activity undertaken by

[5] When pricing insurance, the calibration may be rougher to cover a range of dates covering the period during which contracts may be sold at a given price.

investment bankers when they discovered an inconsistency between pricing of ULSG contracts and lifetime income annuities.

Lifetime income annuities are single-premium contracts, so all premiums are paid at inception, and the insurer therefore calibrates the price to be reasonably consistent with market interest rates available on the date of purchase. In contrast, most premiums under a ULSG contract are paid in the future, and the insurer must estimate the future interest rates at which they might be invested.

Investment bankers discovered that they could invest a lump sum in a lifetime annuity on an individual and use the annuity payments to pay the premiums on a ULSG contract on the same individual. Doing so converted the ULSG contract into something equivalent to a single-premium life insurance contract. The death benefits on such an arrangement were high because the ULSG premium rates were low. Investment bankers used a reasonable mortality assumption to calculate that death benefits would almost assure them of returns of over 7 percent on their initial investment, with potentially much higher returns on early death. This was in a market where interest rates were generally under 7 percent.

The investment bankers took advantage of the inconsistency in pricing between single-premium annuities and ULSG contracts. At the time many of these contracts were being written, market interest rates were notably lower than the average of the future distribution of interest rates produced by the Academy model, which had not been recalibrated and was not market consistent.[6] The required use of the Academy model for regulatory capital measurement may have led to its unintended use in pricing. This may have contributed to the situation where ULSG contracts were offered at rates that seemed inconsistent with the rest of the market, the situation that investment bankers exploited.

This discussion would not be complete without mentioning the relationship between risk-neutral models and real-world market-consistent models for interest rates. This relationship is often misunderstood, and this misunderstanding can be a source of model risk. There are some practitioners that argue that a model that is not risk neutral cannot be market consistent, but this is not the case.

Without going too far into technical details, the issue is that the probability-weighted average projected level of the short-term interest rate is higher in a risk-neutral model than it is in a real-world market-consistent model, because of the use of "risk-neutral" probabilities rather than "real" probabilities.

All market-consistent models must assign a value to the risk of adverse events. Real-world models do this explicitly, without changing the probability of adverse events. Risk-neutral models do this by using "risk-neutral" probabilities that assign extra probability weight to adverse events.

In the context of pricing or valuing securities, an adverse event is one that reduces the value. For fixed income securities, a rise in interest rates is an adverse event that reduces value. Therefore, risk-neutral probabilities are biased toward increases in interest rates. This means that the average probability-weighted path of short-term interest rates in a risk-neutral model tends to be higher than in a real-world

[6] The Academy model was still appropriate for its intended use, which was estimating risk arising from the tails of the distribution of interest rates. One can view the fact that the center of the distribution keeps changing based on current market prices as not materially relevant when the main purpose is to estimate the extreme tails of the distribution.

market-consistent model that is calibrated to the same market prices. This difference arises not because higher interest rates are expected in the risk-neutral model, but because the probability weights in that model are not the real probabilities. Using them as if they were the real probabilities introduces model risk.

Risk-neutral models of interest rates were designed for the purpose of valuing securities in a market-consistent way, without the need to derive an explicit market price for risk. Such models serve that purpose admirably well. However, extending the use of such models to project future interest rates is to go outside the intended purpose of the model, and that always introduces model risk.

References

[1] Model Validation Principles Applied to Risk and Capital Models in the Insurance Industry, North American CRO Council, 2012.

[2] Solvency II—Model Validation Guidance, Lloyd's, June 2012.

[3] Supervisory Guidance on Model Risk Management, Board of Governors of the Federal Reserve System, OCC 2011-12, April 2011.

[4] C. Kaner, J. Bach, and B. Pettichord. *Lessons Learned in Software Testing: A Context Driven Approach*, Wiley Computer Publishing, 2001.

[5] Commission on Models in the Regulatory Decision Process of the National Research Council, *Models in Environmental Regulatory Decision Making*, 2007.

[6] Basel Committee on Banking Supervision, Update on Work of the Accord Implementation Group Related to Validation under the Basel II Framework, *Basel Committee Newsletter*, No. 4, January 2005.

[7] Raymond R. Panko, What We Know about Spreadsheet Errors, *End User Computing*'s Special Edition on Scaling Up End User Development, Volume 10, No 2. Spring 1998, pp. 15–21.

[8] CEIOPS' Advice for Level 2 Implementing Measures on Solvency II: Articles 120 to 126, Tests and Standards for Internal Model Approval, CEIOPS-DOC-48/09, 2009.

[9] Frederick P. Brooks, *The Mythical Man Month: Essays on Software Engineering*, Addison-Wesley, 1995.

[10] N. Nagappan et al., The Influence of Organizational Structure on Software Quality: An Empirical Case Study, Proceedings of the 30th Annual Conference on Software Engineering, 2008, pp. 521–530.

[11] C. Franzetti, *Operational Risk Modeling and Management*, Chapman & Hall, 2011.

[12] A. Inselberg, Multidimensional Detective, Proceedings of the IEEE Symposium on Information Visualization, 1997.

[13] CEIOPS' Advice for Level 2 Implementing Measures on Solvency II: Article 86f, Standards for Data Quality, CEIOPS-DOC-37/09, 2009.

[14] M. Stricker, D. Ingram, and D. Simmons, Economic Capital Model Validation, White Paper, Willis Economic Capital Forum, 2013.

[15] Report of the Senate Permanent Subcommittee on Investigations: JPMorgan Chase and the London Whale: A Case History of Derivatives Risks and Abuses.

[16] Carrick Mollenkamp, Serena Ng, Liam Pleven, and Randall Smith, Behind AIG's Fall, Risk Models Failed to Pass Real-World Test, *Wall Street Journal,* October 31, 2008.

Appendix 1: Some Approaches to Graphical Reporting of Risk

In this appendix we discuss a few data visualization techniques that we consider suitable for a broad audience. First, some of the data presented in the report can be reformatted using the graphics suggested below. Thus, the new figures contain no new information; this is very important for the following test. These new graphics are then presented to the intended audience, and the question is whether this new presentation of the existing information allows them to draw new or different conclusions. If it does not, then this is an indication that the users are close to a reflective equilibrium, which means that their judgment was based on the content rather than the presentation of the content.

Second, these graphics can be used as suggestions for the report developers. We think that one should be careful not to view validation guidelines as model development guidelines, although model developers can certainly profit from a deeper understanding of how a validation team looks at a model. This may be different for the category of reporting risk, because a good presentation of the model results is almost independent from the technical modeling approach being used. In fact, we consider it an advantage to use a presentation independent of the technical modeling approach because this allows the presentation to remain the same over various model development cycles and also across different models that may be used to benchmark the bespoke model.

Since we are talking about economic capital models, the amount of economic capital has to be presented. Although this is a single key figure, we encourage a presentation in which the total economic capital and the stand-alone capital for various segments of the business is visible. Of course, this assumes that the model includes a bottom-up aggregation, but in our experience this is generally the case.

The first kind of presentation will help to identify the risk drivers and their relative importance. This kind of information can be displayed using "waterfall" approach.

Figure A1-1 illustrates presentation of risk drivers using the "waterfall" approach. In this case we recommend displaying them from smallest to largest. The proximity of the largest risk drivers next to the total economic capital makes it easier to compare the largest risk drivers to the total and thus adds relevant risk management information.

Note that the bar for each segment in figure A1-1 represents the stand-alone economic capital for that segment, before reflecting diversification. The entire effect of diversification is shown as a separate bar. One can show the degree of diversification benefit that is attributable to each segment by changing the waterfall graphic as shown in figure A1-2. This latter approach may be more consistent with capital allocation by segment if capital allocation reflects the diversification benefit contributed by each segment.

Figure A1-1

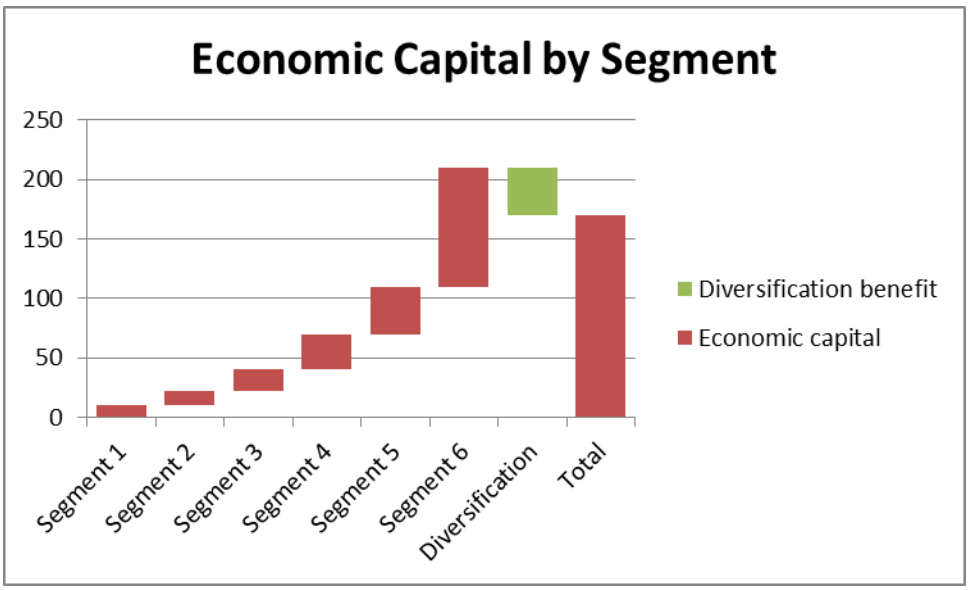
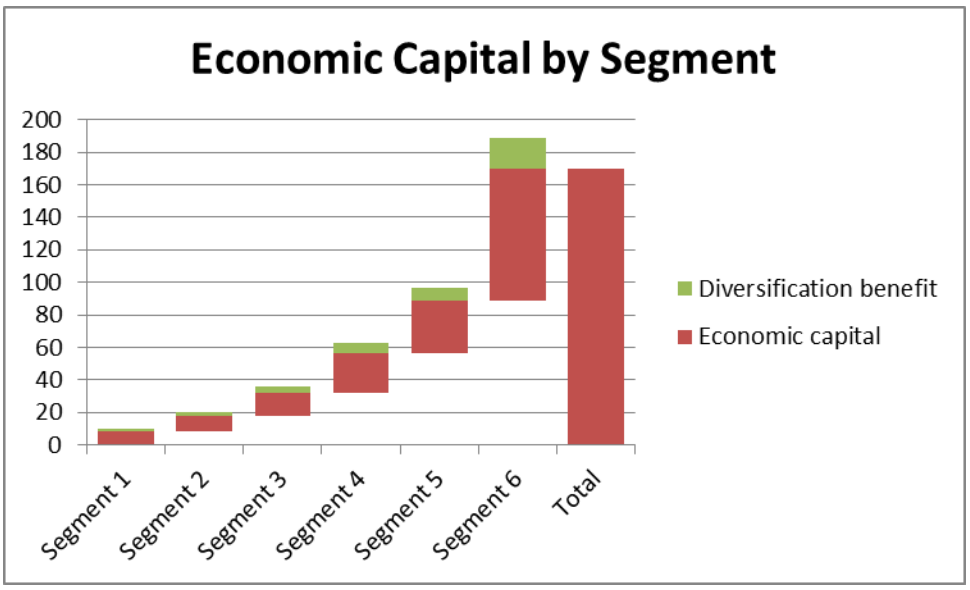


Figure A1-2



Similar graphics can be used to decompose the economic capital into different segmentations, for example, by class of business, by legal entities, etc. Doing so makes it easier to detect anomalies of a model and or an allocation. Strictly speaking the allocation is not required from a regulatory viewpoint. But it is frequently stated as a necessary step in application of the model to steer the business.

Although very valuable, the above graphics provide no insight into the dynamical behavior of the model. It is difficult to visualize the dynamics of economic capital models because these models—deterministic

or stochastic—have many parameters. As a first step, it must be determined to which parameters the model is very sensitive. This is a difficult task, and we will give some guidance about how to do this in the technical section. For the following, we assume that the most important parameters have been chosen, and we need to visualize their effect on the economic capital.

Displaying probability distributions can hardly be the solution, let alone joint distributions in higher dimensions. Most people do not even understand the units in which a probability density distribution is displayed. They might recognize a few characteristic shapes, such as the normal distribution, but are very puzzled to hear that if the random variable is denoted in U.S. dollars, then the *y* axis has the unit 1/U.S. dollar. On the other hand, cumulative probability distributions are easier to understand, since quantiles can easily be shown, but they do not have very characteristic shapes. In addition, we are concerned about the tails of these distributions, and those can definitely not be visualized with either probability density functions or cumulative distribution functions. Technically oriented people use copulas to separate the marginal distributions from the dependency structure of multidimensional probability distributions. But copulas are nothing other than multidimensional probability distributions on the unit cube. And hence, if probability distributions cannot be used for communicating results, then copulas cannot either. Another simple reason why we believe that probability distributions should not be used in reports is that they can obviously be applied only to stochastic models. Yet some submodels in a large model or a benchmark model might be deterministic factor models, which require a different visualization technique. We are left with a difficult task in communicating the dynamic behavior of a model. But how should users of a report compare results if they are not comparable in their presentation format?

Parallel coordinates can be used to visualize and analyze multivariate data. The underlying idea is to map a point in an *n*-dimensional space onto a line in the normal two-dimensional plane.

A point in *n*-dimensional space is specified with n coordinate values. In the context of a stochastic model of an enterprise, the point can represent the results of one scenario in a set of stochastic scenarios. The *n* dimensions correspond to *n* business segments of the enterprise. The *n* coordinate values of a point are the *n* operating gains for each segment in that scenario.

The *n* coordinate axes are all arranged as equally spaced parallel vertical lines in the plane. Each vertical line corresponds to a business segment, and the line is an axis with a range of values that include (but may extend beyond) the range of operating gains for that business segment. For each point in *n*-dimensional space (that is, each scenario), a line is drawn across all the vertical lines such that it intersects each vertical line (each business segment) at the value of the coordinate on that axis (the segment's operating gain in that scenario).

Figure A1-3 shows how the results of one scenario could be shown using parallel coordinates. The green line crosses each vertical axis at a point corresponding to the operating gain for that segment in that scenario.

Figure A1-3

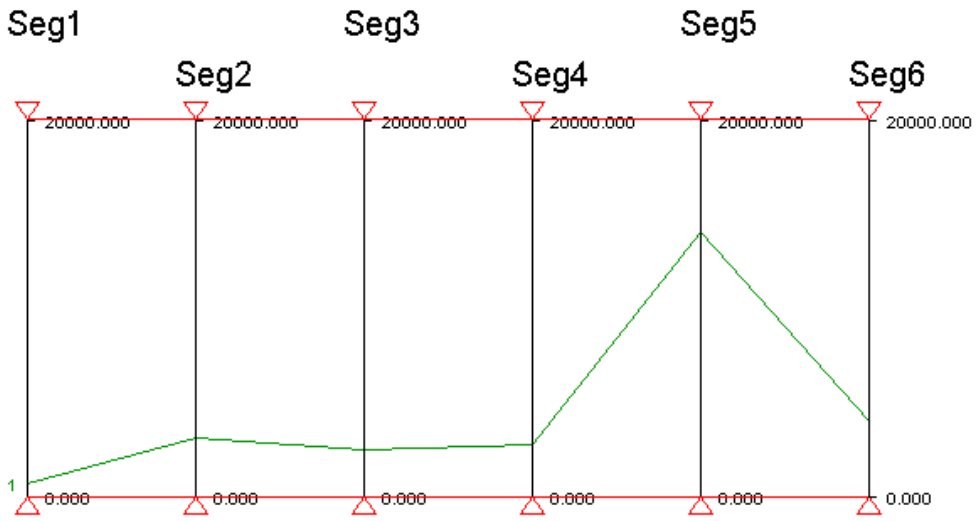
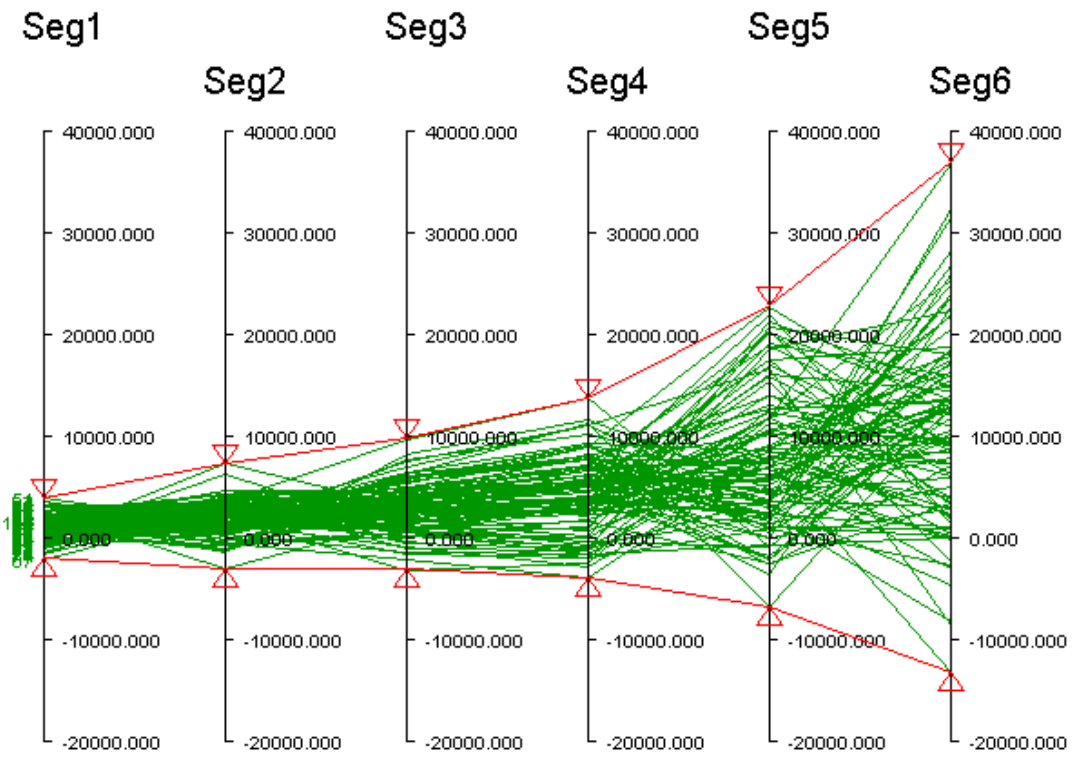


Figure A1-4 shows how this graph might appear when 100 lines (100 scenarios) are included. Note that results for some business segments do not appear to vary much. That happens because the range on the vertical scale is the same for all business segments, and some business segments are much smaller than others.

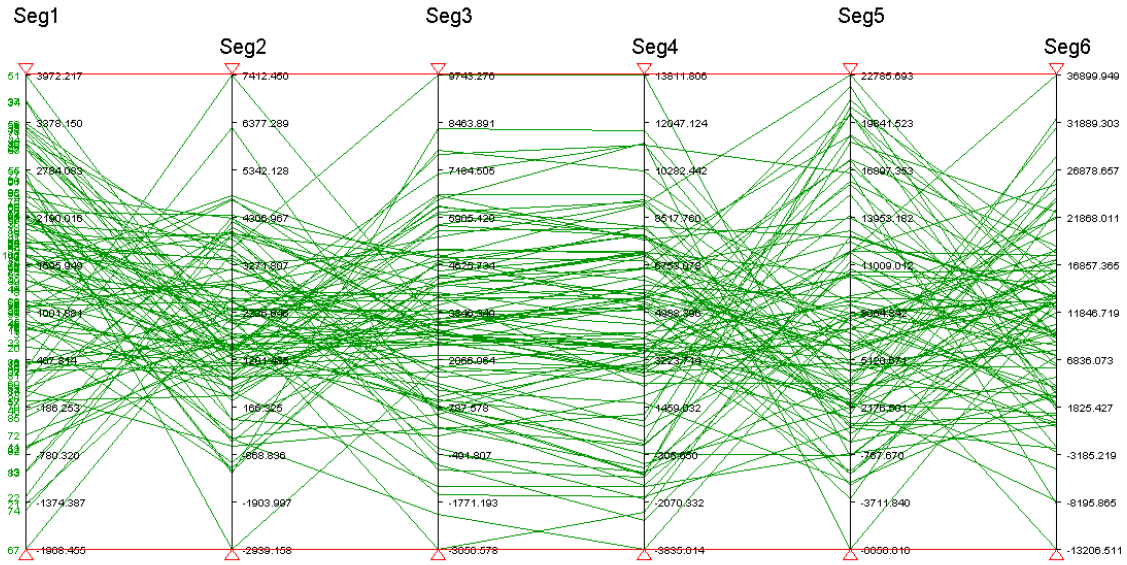
The software that generates these graphics places a red marker at the top and the bottom of each axis as a means of allowing the user to filter results. The only lines (scenarios) that are shown are those that intersect each axis within the range of the filter. In this case the range of the filter range has been set wide enough that no scenarios are excluded. The red line at the top and bottom indicates the maximum and minimum values for each axis across all scenarios.

Figure A1-4



This graph can be adjusted to highlight outlier results in smaller lines of business. In Figure A1-5 the range of each vertical axis is different because it is the range of operating gains for that business segment alone, not the range of total enterprise gains. This magnifies the visual impact of the differences between scenarios in small business segments and can help illustrate how fluctuations in one segment are related to the adjacent ones, without regard to size. In statistical terms, this approach can help illustrate the degree of correlation between the variability of results in adjacent business segments.

Figure A1-5



Anyone viewing Figure A1-5 will notice that the lines between segments 3 and 4 are mostly horizontal, whereas those between other segments contain a mixture of horizontal and sloped lines. This illustrates that results for segments 3 and 4 are highly correlated, whereas those for other segments appear much less correlated in this graph.

Sometimes changing the order of the vertical axes can yield important insights. To illustrate this, Figure A1-6 uses the same data, but includes only business segments 1, 5, and 6. One can see that high results in segment 1 are always connected to low results in segment 5. However, segments 5 and 6 exhibit no clear negative correlation.

This can be made even more visually obvious by reversing the scale of the axis for segment 1, as shown in Figure A1-7. When this is done, it is quite clear that there is some sort of connection between results for segments 1 and 5.

Figure A1-6

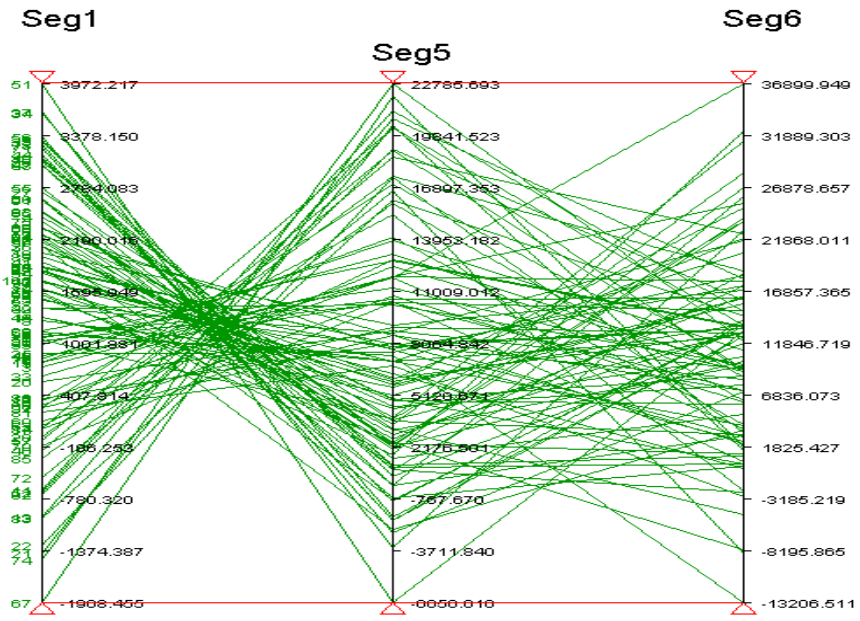
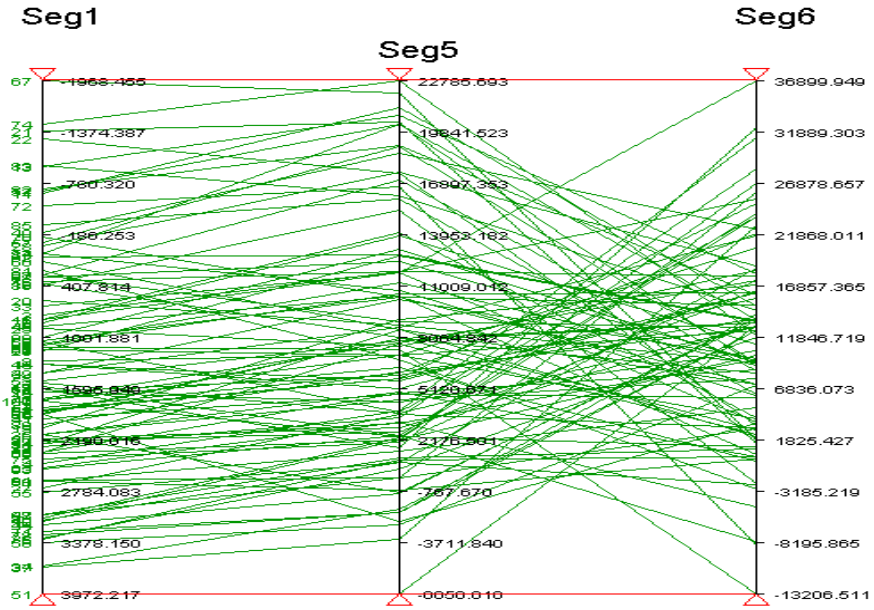


Figure A1-7



This technique was already used in the late 19th century, but it was mostly forgotten and then surfaced again around 1960. Here we are not concerned with how geometric properties transform from *n* dimensions to *n* parallel coordinates. It is the data analysis aspect of parallel coordinates that we would like to promote.[7]

Interactive features can be used to highlight the relationship between two variables. But at the same time this points to a weakness of parallel coordinates. If we order the coordinate axis differently, then this changes the picture and the insights completely. It is generally agreed that more insight into the multidimensional data can be obtained by ordering the axis such that the number of crossovers are minimized. No easy and automated way exists to achieve this. In our context we have one axis with a special meaning: the economic capital. Following the same argument as in the waterfall graphics, we recommend putting this axis on the right. From left to right, ordering the risk factors with increasing influence on the economic capital is a good starting point for the visualization.

All the graphics in the examples above were created using the open source parallel coordinates tool xdat (http://www.xdat.org). The included graphics may not reveal the full power of parallel coordinates, because the dynamical behavior of the model can best be explored interactively. The data for this example were created stochastically using prespecified correlations.

It would be helpful to add quantile markers on each axis; unfortunately xdat does not support this feature. Risk management is not only about detecting outliers and preventing or mitigating them. A good understanding of normal business volatility is part of a sound risk management. After all, outliers can be defined only once we know what is considered to be normal.

A useful way to display the normal business volatility is to create a data set that contains all the scenarios that yield a target value, in our example the total claims, between the 25th and the 75th percentile. Typically one will find that the range of results for each business segment in such a data set includes points outside the 25 to 75 percent range for the segment, even when the total is limited to that range. A demonstration of that fact can lead to better understanding of how diversification leads to a reduction in risk.

We want to emphasize that the above visualization techniques can be used for output of deterministic, as well as of stochastic, models. The results of a single scenario under deterministic assumptions can be placed in the same graphic (perhaps using a different color) to show how it falls within the range of stochastic results.

[7] A good overview can be found in Inselberg's paper [13], and a good example of how this sort of graphics can create insights can be found at http://eagereyes.org/techniques/parallel-coordinates.

Appendix 2: Comparison with Other Frameworks

North American CRO Council

In 2012 the North American CRO Council published a paper entitled "Model Validation Principles Applied to Risk and Capital Models in the Insurance Industry" [1]. Table A2-1 lists the eight principles discussed in that paper and explains how each is reflected in this paper.

Table A2-1

CRO Council Principles	How Reflected in This Paper
1. Model design and build need to be consistent with the model's intended purpose	Conceptual risk, as discussed in this paper, focuses on the issue of consistency between the model design and its intended purpose. It is unclear whether the term "build" in the CRO Council Principles refers to implementation risk, which is treated separately in our work.
2. Ensure that model validation is an independent process	We view this as more a matter of model governance than validation; however, we agree with the need for independence on the part of the validators, as discussed in our brief subsection on governance.
3. Establish an owner of model validation	We view this as more a matter of model governance than validation; however, we agree with the need to clearly designate responsibility for validation.
4. Ensure appropriateness of established model governance	In this paper, validation is viewed as a subset or part of model governance, rather than viewing governance as part of the model to be validated. Clearly these topics are interrelated, as discussed in our brief subsection on governance.
5. Make model validation efforts proportional to evidenced areas of materiality and complexity	We agree with this as a general principle of good management. Our focus has been more toward ensuring that the validation is complete and addresses all aspects of model risk.
6. Validate the model components: a. Input components b. Calculation components **c.** Output components	These three parts roughly correspond to three of the five elements of model risk toward which validation efforts are directed in our framework: a. Input risk b. It is unclear whether their calculation components refer to the algorithm or the implementation of the algorithm. We distinguish between the two, whereas the CRO Council Principles do not seem to distinguish them. c. Output risk, except that we treat reporting risk as a separate issue
7. Address limitations of model validation	This is refined in our work into three separate

	issues: a. Conceptual limitations b. Limitations of the implementation of the concepts and c. Managing the control cycle: the difficulty of performing model validation in a dynamic and constantly changing environment that places limitations on the completeness of validation at any point in time
8. Document the model validation	We provide a subsection that covers presentation and communication of a model validation.

Solvency II Criteria for Regulatory Approval of Internal Models

Articles 100-123 of the Solvency II draft directive include six clear criteria for regulatory approval of internal models. Table A2-2 lists these criteria alongside comments regarding how they are reflected in this paper. It should be noted that this paper does not address the same issue (criteria for regulatory approval), but since model validation is certainly needed for regulatory approval, there is significant overlap.

Table A2-2

Solvency II Criteria for Regulatory Approval	How Reflected in This Paper
1. Use test • Demonstration of internal widespread use • Role in risk management governance and decision making • Role in economic and solvency capital allocation • Frequency of use • Appropriateness of the model	Not all models are intended for regulatory purposes, so we discourage wider use than appropriate for any particular model. However, the model's role in decision making must be considered in validation, along with the appropriateness of the model, e.g., the consistency between the model design and its intended purpose. We have included these issues in the phase of validation that addresses conceptual risk.
2. Statistical quality • Adequate statistical and actuarial techniques • Consistency with technical provisions • Current and credible information • Realistic assumptions • Ability to justify assumptions to authorities • Data accuracy, completeness, and appropriateness • Yearly update of data • Nonprescription of any particular distributions	The criteria covering statistical quality span areas that we have classified as partly conceptual risk and partly input risk. The implementation risk that we relate to this issue is not as explicitly covered in Solvency II.

• Cover all material risks • Cover all risks explicitly mentioned under Solvency II framework • Allowance for diversification is permitted • Allowance for risk mitigation techniques is permitted • Special attention to options and guarantees • Allowance for future management actions • Specific allowance for nonguaranteed payments	
3. Calibration • Allowance for different time periods of assessment provided broadly equivalent • Capital requirements derived directly from probability distributions where possible or else approximations can be used if equivalent protection is demonstrated • Supervisors may request validation against external data	Calibration is discussed here both in the context of validating model inputs and in the context of the feedback loop whereby variances between forecasts and actual results are used to update calibration. The emphasis on statistical methods in Solvency II reflects the fact that its capital requirements are expressed in terms of a specific probability level rather than in other terms such as ability to survive a specific stress test.
4. Profit & loss attribution • Annual P&L attribution required • Risk classification to be related to attribution analysis	Profit and loss attribution is a helpful technique for analyzing financial results. In the context of model validation, it is part of the feedback loop whereby results are compared with forecasts and variances are identified and used to review the calibration of model assumptions or input parameters.
5. Validation standards • Regular cycle of validation required, with scope: monitoring performance, appropriateness of specification, testing results against experience • Effective statistical process to demonstrate appropriateness • Statistical methods should also apply to material new data and information • Analysis of stability of model, sensitivity of results • Accuracy, completeness, and appropriateness of data	All of these categories of validation activity have been discussed in this paper as part of the validation process. The emphasis on statistical methods in Solvency II reflects the fact that its capital requirements are expressed in terms of a specific probability level rather than in other terms such as ability to survive a specific stress test.
6. Documentation standards • Sufficient to demonstrate compliance with all above tests • Theory, assumptions, mathematical basis, and empirical basis • Limitations of model • Change history to be documented	The validation process in this paper includes checks for the existence of adequate documentation of theory, assumptions, limitations, and change history.

Guidance on Model Risk Management from the Office of the Comptroller of the Currency

In April 2011, the Office of the Comptroller of the Currency (OCC), under the direction of the Board of Governors of the Federal Reserve System, published a document titled "Supervisory Guidance on Model Risk Management" [3]. Although that document discusses many of the same issues as are addressed in here, it puts a more narrow scope on the term "validation" than is done in this paper.

The OCC paper, by its title, is focused on model risk. This paper equates the management of model risk with validation, but the OCC paper does not do so. Although we outline five sources of model risk, the OCC paper says model risk occurs primarily for two reasons:

- The model may have fundamental errors and may produce inaccurate outputs when viewed against the design objective and intended business uses.
- The model may be used incorrectly or inappropriately.

This breakdown is relatively vague, and we believe that the more detailed five-part breakdown we have provided will be helpful in identifying additional sources of model risk that need to be addressed in the validation process.

The topic of validation is viewed more narrowly in the OCC document, being just one part of model risk management. The OCC paper says that "An effective validation framework should include three core elements":

- Evaluation of conceptual soundness, including developmental evidence
- Ongoing monitoring, including process verification and benchmarking
- Outcomes analysis, including back-testing.

Such a framework for validation does not address whether the model is being used in an appropriate way, nor does it address whether the model output is being communicated in an effective way to the ultimate decision makers that depend on the model. We believe that these additional considerations should be part of a validation. The OCC paper does not ignore them, but rather places them outside the model validation process, implicitly as part of model governance.

From a top-level perspective, the OCC paper views model risk management as synonymous with model governance. Our view is that model governance focuses on when models are to be used and when they are not, and oversees their creation and validation. Validation, in our view, is the process that minimizes the risk of depending on a model, and it occurs within a governance process that determines if a model is to be used and oversees its use.