The quantification and aggregation of model risk: perspectives on potential approaches

Michael Jacobs Jr.

Accenture Consulting,
1345 Avenue of the Americas,
New York, NY 10105, USA
Email: michael.a.jacobs@accenture.com

Abstract: The field of Model Risk Management (‘MRM’) continues to evolve. To date, industry MRM efforts focused primarily on MRM for individual models. Now, more institutions are shifting focus towards aggregating firm-wide model risk. Regulatory guidance specifically focuses on measuring risk in individual and in aggregate. In this study, we will discuss various approaches to measuring and aggregating model risk across an institution. We also present an example of model risk quantification in the realm of stress-testing, where we compare alternative models in two different classes, Frequentist and Bayesian approaches, to modelling stressed bank losses.

Keywords: financial crisis; model risk; SR 11-7; stress-testing; Bayesian.


Biographical notes: Michael Jacobs Jr. is a Principal Director in the Risk Analytics practice of Accenture. He specialises in risk model development and validation across a range of risk and product types. He has 25 years of experience in financial risk modelling and analytics, having worked 5 year for Deloitte and PwC as a Director in the risk modelling and analytics practice, with a focus on regulatory solutions; 7 years as a Senior Economist and Lead Modelling Expert at the OCC, focusing on ERM and Model Risk; and 8 years in banking at JPMC and SMBC, developing wholesale credit risk and economic capital models.

1 Introduction and motivations

Modern risk modelling (e.g. Merton, 1974) increasingly relies on advanced mathematical, statistical and numerical techniques to measure and manage risk in bank portfolios. This gives rise to model risk (OCC and BOG-FRB SR 11-7, 2011), defined as the potential that a model used to assess financial risks does not accurately capture those risks,1 and the possibility of understating inherent dangers stemming from very rare yet plausible occurrences perhaps not in reference data sets or historical patterns of data (Jacobs, 2013; Jacobs et al., 2015).
In recent years, there has been a trend in financial institutions towards greater use of models in decision-making, not only driven to some extent by the evolution of prudential regulation, but also manifest in all aspects of management. In this regard, a high proportion of bank decisions is automated through decision models, which can be either statistical in nature or a methodology that constitutes a rule-set (Davenport and Harris, 2005).

There are several areas in which we see this trend manifest. First, as of late there has been an increased use of electronic trading platforms known as algorithmic trading that executes trade commands which have been pre-programmed (e.g. through timing, price or volume), and can be initiated with manual intervention. A recent example of this is such an automated trade command that occurred on 6 May 2010 resulted in a $4.1 billion ‘flash crash’ of the New York Stock Exchange, which fell more than 1000 points, subsequently recovering to its previous value in only 15 minutes time.

Second, partly encouraged by the Basel regulations (BCBS, 2005; BCBS, 2006; BCBS, 2009a; BCBS, 2009b; BCBS, 2010), banks are increasingly using decision models in their origination, monitoring and credit recovery processes. Therefore, the viability of a loan commitment is determined by the client’s probability of default (‘PD’), and the loss given default (‘LGD’) of the loan instrument. Similarly, banks monitor customer accounts and anticipate credit deterioration using automated alert models, which pre-classify customers and determine their credit limits. In the case of credit collections, they develop statistical profiles of delinquent customers in order to implement different recovery strategies.

In the commercial arena, customers are able to select a product’s characteristics (e.g. loan amount, term, purpose, etc.) and the system in turn makes a real-time decision on the viability and price of the transaction. In many case, the model asks the customer a number of questions and proactively makes the offer that best suits the customer, which as a manual process would be cumbersome and impractical.

The use of valuation models for products and financial instruments has become widespread in financial institutions, in both the capital markets and the Asset-Liability Management (‘ALM’) businesses. Some classic examples of such models include Black–Scholes option pricing, the Capital Asset Pricing Model (‘CAPM’) and Monte Carlo simulation valuation models. Furthermore, customer on boarding, engagement and marketing campaign models have become increasingly prevalent. Such models are used to automatically establish customer loyalty and engagement actions, both in the first stage of the relationship with the institution and at any time in the customer life cycle. Actions include the cross-selling of products and services that are customised to meet the needs of the client, within a customer relationship management framework.

Other examples include the calculation of economic capital (‘EC’) charges for various types of exposures, subject to different risk types (e.g. credit, market, operational, ALM, etc.) through their individual components. This includes the quantification of the bank’s current liquidity position, projected under alternative scenarios, as well as the projection of balance sheet and income statements for the use in stress-testing (‘ST’). Furthermore, this includes the modelling of many key components in business planning and development, such as optimal bundling, customer or non-customer income or churn.

The use of models brings many palpable benefits, including:

- Automated decision-making, which in turn improves efficiency through reducing analysis and manual decision-related costs.
• Objective decision-making, ensuring that estimated results are the same in equal circumstances and that internal and external information is reused, thus leveraging historical experience.

• Ability to synthesise complex issues such as a bank’s aggregate risk.

However, model use also gives rise to risks and costs, including:

• Direct resource costs (e.g. human capital and economic opportunity cost), as well as time expended in model development and deployment.

• The risk of relying upon output of an incorrect or misused model, for which there exist many recent examples of specific instances of large losses.

Model error encompasses phenomena such as simplifications of or approximations to reality, incorrect or missing assumptions, incorrect design processes, and measurement or estimation error. On the other hand, model misuse includes applying models outside the use for which they were designed and so suited.

Model risk so defined is potentially very significant and has captured the attention of regulators and institutions, whose approaches range from mitigation via model validation to the establishment of comprehensive frameworks for active Model Risk Management (MRM). In the case of advanced institutions, such active management has been formulated into an ‘MRM’ framework, which sets out the guidelines for the entire design, development, implementation, validation and inventory and use process. Such developments have been fortified by the fact that prudential supervision and regulation are now requiring such frameworks, as stated in guidance issued by entities such as the Federal Reserve System and the Office of the Comptroller of the Currency (OCC and BOG-FRB SR 11-7, 2011) in the USA, which are serving as stating points for the industry.

However, the prudential supervisory guidance fails to provide detail regarding model risk quantification. An exception to this are rather specialised situations, such as the valuation of certain instruments, in which cases there may even be a requirement for Asset Valuation Adjustment (‘AVAs’) that may result in a larger capital requirement or in the possible use of a capital buffer for model risk as a mitigating factor in a broader sense, without its calculation being specified.

Having set this context, this study aims to provide a holistic view of model risk quantification, including a survey of approaches to model risk aggregation. This paper shall proceed as follows. In Section 2, we address the conceptual issue of defining a model. In Section 3, we review the supervisory framework with respect to MRM. Section 4 addresses the model risk quantification question. Section 5 presents some perspectives on the aggregation of model risk. In Section 6, we provide an empirical example of model risk quantification in the context of stress-testing bank credit risk portfolios. Finally, Section 7 summarises this study and provides direction for future research.

2 Review of the literature and supervisory developments

Currently, the academic or technical practitioner literature on quantitative approaches to MRM is rather recent and somewhat limited. Millet and Wedley (2002) propose methods for modelling risk and uncertainty with the Analytic Hierarchy Process (AHP), showing
why benefit/risk ratios might be an improper modelling approach. The authors then introduce prototypical case studies where risk plays a role in multi-criteria decision-making to demonstrate how the AHP can be used to derive relative probabilities, multiple criteria outcome measures, risk criteria, and risk adjustment factors in order to measure model risk and model uncertainty. Cont (2006) studies the uncertainty on the choice of an option pricing model that can lead to model risk in the valuation of a portfolio of options. After discussing some properties which a quantitative measure of model uncertainty should verify in order to be useful and relevant in the context of risk management of derivative instruments, the author introduces a quantitative framework for measuring model uncertainty in the context of derivative pricing, and also discusses some implications for the management of model risk in this context. Rai (2008) studies probabilistic risk characterisations commonly used to model risk and exposure, identifying sources of variation and characterising risk as a function of risk factors. The author shows that the distribution of any risk factor depends on two sources of heterogeneity; on the one hand, uncertainty representing the degree of ignorance about the precise value of a particular parameter of a distribution, and, on the other hand, variability representing the inherent variation within the population of interest, further giving an example to demonstrate uncertainty and variability that are based on a multiplicative model of risk due to contaminated drinking water. Ammann and Verhofen (2008) use Bayesian model averaging to analyse industry return predictability in the presence of model uncertainty. Their posterior analysis shows the importance of inflation and earnings yield in predicting industry returns, and the out-of-sample performance of the Bayesian approach is found to be superior to that of other statistical model selection criteria; and further develops a variance decomposition into model risk, estimation risk, and forecast error, demonstrating that model risk is less important than estimation risk. Crouhy et al. (2008) argue that models are an inevitable feature of modern finance, and model risk is inherent in the use of models, thus they stress the technical elements of model risk. The authors emphasise that models are susceptible to errors and discuss how to mitigate model risk. Glukhov (2012) studies model risk quantification through the probabilistic decision process, starting from the Bayesian inference process, which through the incorporation of priors influence posterior confidence intervals for the model parameters. Alexander and Sarabia (2012) develop a methodology for quantifying model risk in quantile risk estimates, arguing that the application of this technique to risk assessment has become a common practice in many disciplines, and is particularly important in finance where quantile estimates have been the cornerstone of banking risk management since the mid-1980s (e.g. Value-at-Risk – ‘VaR’). Boucher et al. (2013) illustrate and estimate model risk, and focus on the evaluation of its impact on optimal portfolios at various time horizons, for standard risk measures such as VaR as the key tool for asset allocation. Based on a long sample of US data, the authors find a non-linear relation between VaR model errors and the horizon that impacts optimal asset allocations. Barrieu and Ravanelli (2015) study capital requirements when the bank’s econometric model only approximately describes the dynamics of portfolio returns, deriving a simple formula for capital requirements based on a first-order Taylor expansion of the VaR around a ‘model confidence’ parameter that quantifies model uncertainty. Most recently, Jacobs et al. (2015) contribute to the literature on measuring model risk through the application of a Bayesian model in a stress-testing model application to the US Federal Reserve’s Comprehensive Capital Analysis and Review (CCAR) program. The authors compare the proportional model risk buffer measure of
the severely adverse cumulative nine-quarter loss estimate, a common way to estimate being measure of statistical uncertainty generated by a model, from their empirical implementation of the Bayesian to the Frequentist model and find it to be 40% higher in the former versus (vs.) the latter.

Turning to the supervisory guidance on MRM, we note that at this point in time it is limited, and the regulations tend to be rather principles-based – that is there is a lack of specificity in how to define and treat model risk. In addition to the guidance issued by the OCC and Fed, there exist some other regulatory references to model risk, which we may categorise into three types:

- **Asset Valuation Adjustments (AVA):** These are strictures governing a requirement to conservatively adjust the values of certain instruments to account for any potential model risk. These regulations are particularly applicable to financial derivatives.

- **Internal Capital Adequacy Assessment Process (‘ICAAP’) capital buffer:** The second Pillar of Basel II, as well as some local supervisors’ version of this, governs the need to hold capital for all risks deemed material to the institution, as part of the ICAAP. This also holds for the aforementioned annual stress-testing exercise in the USA known as the CCAR, which is a component of ICAAP. In this setting, supervisors may require an institution to hold additional capital for model risk, in spite of the fact that many entities already do so.

- **Other references:** These include other lower profile mentions of model risk, which often consider such to be subordinate or already a component of other supervisory aspects. An important case is the imposition by the **Basel Committee for Banking Supervision** (BCBS) of the leverage ratio as a mitigant to model risk.

However, the principal regulatory occurrence was the publication of the **Supervisory Guidance on Model Risk Management** (OCC and BOG-FRB SR 11-7, 2011) but the Office of the Comptroller of the Currency (OCC) and Board of Governors of the Federal Reserve System (FED) in the USA. This is the first such guidance in which model risk is clearly defined, and in which entities are urged to have an established framework and processes in place to manage such risk. In order to meet this objective, the guidance lays out a set of guidelines or principles that will guide the management of model risk, which we summarise here:

- **Model risk is to be managed in the same manner of any other risks faced by an institution.** That is, model risk managers should identify the sources, and assess the likelihood of occurrence and the severity given occurrence of a model failure.

- **As with other risk types, model risk can never be completely eradicated, rather than only controlled by effective management.** While comprehensive model development combined with rigorous validation can mitigate model risk, such processes will never eliminate it.

- **Given that MRM is to be treated likewise with respect to other risk types, institutions should establish an MRM framework, which is approved and overseen by senior management and the Board of Directors.**

- **While a well-documented, prudent and structured approach to model development and validation aspects (i.e. inputs, outputs and design) can be efficacious in managing model risk, this is not a substitute for the continual process of improving models.**
The quantification and aggregation of model risk

- A prudent use of models is likely to involve elements such as conservative adjustments, stress-testing or even possibly capital buffers. However, an over-reliance on these conservative elements may lead to the misuse of models.

- Model risk emanates from many sources; consequently institutions should consider the interaction of these factors, and try to quantify aggregated model risk resulting from their combination.

- While the organisational structure of the MRM framework is clearly at the discretion of individual institutions, there are certain strictures of good governance that should be observed, namely the separation of the following elements:
  - **Model ownership**: Identification and measurement of the model risk to which the institution is exposed
  - **Model control**: Limit-setting, follow-up and independent validation
  - **Model compliance**: The set of processes that ensure that the ownership and control rules are performed in accordance with established policies

- Ultimate responsibility, oversight and approval of the MRM framework reside with the **Board of Directors**. Furthermore, the Board should be notified of any significant emerging model risk to which the firm is exposed.

- **Effective challenge** is the principal form of model risk mitigation. This is defined as critical analysis by objective parties who have the following capabilities:
  - Experienced experts in the lines of business in which the model is used
  - Have the ability to identify model limitations and assess the validity of assumptions
  - Are capable of suggesting model risk mitigants and model improvements

We gather from this that supervisors are expecting institutions to architect MRM frameworks that feature model development and validation criteria which are formalised, promote careful model use, set criteria to assess model performance, and define policy governance and applicable standards of documentation. Such a holistic approach to model risk is a relatively novel occurrence in the industry. We expect this to become increasingly prevalent as a wide swath of institutions eventually adopt this approach, as have the more advanced institutions by this point in time.

### 3 Defining a model and the quantification of model risk

In the analysis of model risk, the primary question which arises is how we define a model. According to the supervisory guidance, a model could constitute any quantitative method, methodology or rule-set. What any of these have in common is that these approaches apply statistical, economic, financial, or mathematical theories, techniques or assumptions that function to transform input data into quantitative estimates. Broadly speaking, this construct has three components:
• Inputs: These constitute either data (which could be ‘hard’ or an expert opinion-based), hypotheses, assumptions or other model output.

• Processing apparatus: This is a method, technique, system or algorithm for transforming model inputs into model outputs. This may be a statistical, mathematical or judgemental.

• Reporting component: This is the system for converting model outputs into a form that is useful for making business decisions.

According to the supervisory guidance, it is always the case that the definition of a model also encompasses quantitative approaches in which the inputs are partially, or completely, qualitative or expert-based. The only proviso that distinguishes such a non-hard data input system as a model is that the output is somehow quantitative in nature. Therefore, the concept of a model as understood herein is far broader than the concept of a mathematical algorithm, which can be seen in this context to be an incomplete interpretation. In particular, this interpretation admits expert judgemental constructs as well. Nevertheless, it can be easily seen that such a definition is still subject to interpretation, as according to the above the following would for sure have to be considered as models:

• An algorithm for the calculation of VaR or Credit VaR (C-VaR) for credit or market risk using either a structural model, Monte Carlo or a historical simulation.

• A scoring model that can be calibrated to estimate a PD for a loan portfolio, using techniques such as logistic regression.

• Valuation constructs with respect to instruments (i.e. derivatives, credit exposures, insurance contracts, etc.) or portfolios of such.

The converse under this definition would call into question whether the following are models:

• Data aggregation algorithms such as statistical summary statistics with respect to a data series (e.g. sums, averages, standard deviations, etc.), financial ratios or calculators that combine model output (e.g. expected loss calculation for the Allowance for Loan and Lease Losses – ALLL).

• A growth projection, or a discounted cash flow calculation, based on a single or limited number of years historical rate of change, or market discount rates.

• A rule-set for decisioning that is based upon a single variable and either a judgementally or data-based threshold (e.g. an acceptance vs. rejection decision for a real estate loan based upon a Loan-to-Value (LTV) ratio).

However, institutions have the latitude to decide themselves what is in scope as a model, and would be subject to model risk, therefore affected by model risk policies and supervisory guidance. Such scope will often not be precisely defined; hence, the application of expert subjective judgement will necessarily be a component of the process. However, once such scope is crystallised, a companion decision will be at hand regarding which models to analyse (e.g. risk, commercial, financial, etc.).
3.1 Measurement of model risk

Apart from some special cases, such as the valuation of derivatives (e.g. model risk AVAs), the quantification of model risk is not at this time an explicit requirement on the part of prudential supervisors. It is for this reason that there has not been significant progress in developing such analytics, as institutions have not had the regulatory impetus to do so.

Nevertheless, from the standpoint of prudent risk management, some institutions have found it to be in their interest to research means of analysing model risk in a quantitative framework. The purpose of this is to support a qualitative MRM framework as promoted by the supervisory guidance. Therefore, we can develop a quantitative MRM life cycle, which would have three phases (see Figure 1):

- **Identification** of model risk sources and classification according to the following taxonomy:
  - Data errors or contamination, missing values, insufficient time series or samples
  - Estimation uncertainty or model error, computational complexity, invalid assumptions
  - Model not recalibrated in a timely manner, used for an incorrect purpose or execution error
Quantification of the model risk inherent to each source. This can be accomplished using a methodology based upon model output sensitivity to potential fluctuations in the inputs, such as estimation uncertainty in parameters, which can characterise the uncertainty associated with such source.

- Sensitivity of output to exclusions of variables or data point/time periods
- Measures of statistical estimation model, benchmark models or market benchmarks
- Measure decay in predictive power between re-estimations or the impact of not using the model

Mitigation of model risk identified and quantified by applying the appropriate measures, which will depend on the nature of the source.

- Data quality assurance processes
- Quantitative capital buffer for model risk; conservativism in inputs, estimates and outputs; model back-testing and stress-testing
- Strict model control environment & governance; ongoing monitoring (limits, alerts, etc.).

Although the elimination of model risk is not possible, implementing an approach that combines rigorous risk management structures, as prescribed by the supervisory guidance, with prudent detailed quantification such as described, may be an effective strategy to mitigate it.

3.2 Aggregation of model risk

Measuring an aggregation of model risk is a new and evolving field. Multiple methodologies are being considered but none established as ‘standard’. The following are three general approaches for discussion:

- Approach 1 – model risk scorecard: A qualitative scorecard approach that considers inherent Model Risk Factors (MRFs) and model risk mitigation activities to present an aggregate view of model risk across an institution.

- Approach 2 – advanced operational risk: A quantitative approach to model risk aggregation based in part on the event and scenario modelling techniques of the Advanced Measurement Approach (AMA) to operational risk.

- Approach 3 – model uncertainty: A bottom-up, quantitative approach to measuring model risk based on model risk sensitivity and model risk control activities such as benchmarking, back-testing, consideration of alternative modelling techniques, and corresponding impact.

Borrowing from the traditional Enterprise Risk Management (‘ERM’) frameworks, the Model Risk Scorecard Approach (‘MRSA’) considers inherent MRFs and model risk mitigation activities to present an aggregate view of model risk across an institution. To measure model risk using the MRSA, each model in the institution’s model inventory is assessed based on the following factors (see Figure 2).
Inherent model risk represents the starting point of the MRM process for an individual model, and can be a function of a number of factors (similar to the approach for tiering models), including:

- **Complexity:** Mathematical structure or estimation computational overhead is high.
- **Uncertainties:** Doubts regarding data, inputs or model assumptions.
- **Ubiquity:** Model is in broader use or is interconnected with other models.
- **Impact:** The model directly impacts earning/capital or has regulatory/accounting uses.

An example of a model scorecard is shown in Figure 3 in which we have two categories of factors, inherent model risk vs. model risk mitigants; and within each category two sub-categories, complexity vs. impact for the latter, and management vs. validation activities in the former. Model A scores low on both inherent model risk dimensions, and high for both model risk mitigants, so has a blended overall model risk score of low. On the other hand, this rating pattern is reversed for Model B, which scores low on both risk mitigants, and high on both inherent model risks, to yield a grand score of high total model risk. Finally, while Model C is similar to Model A in that is scores low for both aspects of inherent model risk, it fails on the model risk activities dimension by getting a score of low (while still getting a high on model validation activities), so that it earns a medium score for total model risk.
Inherent model risk is mitigated (or ‘hedged’) by a combination of model risk control activities within the business units and by dedicated MRM control functions:

- **Ongoing MRM by model owners (‘First Line of Defense’).**
  - **Pre-implementation model validation:** Measuring predictive or discriminatory power on a model development holdout sample
  - **Model performance monitoring:** Measuring exceptions or breaches of confidence bounds between model redevelopments or full model validations
  - **Ongoing model benchmarking:** Comparison of model output to internally generated, vended challenger models or relevant market quantities
  - **Ongoing model back-testing:** Measuring predictive or discriminatory power on incremental post model-development period data

- **Activities of the MRM control functions (‘Second Line of Defense’).**
  - **Quantitative model validation:** Initial and periodic analysis around mathematical or statistical assumptions, input data, variable or model selection process, estimation/calibration, model fit diagnostics, model predictive power, characteristics of model output
  - **Qualitative model validation:** Initial and periodic analysis around business or economic assumptions, evaluation of conceptual soundness or model theory, assessment of intuitiveness of results and quality of reporting
  - **Model control activities:** Controls around model use, model governance, version or change control, model operating or IT environment, model limit monitoring.

**Figure 4** Illustration of the aggregation of model risk scorecards (see online version for colours)

<table>
<thead>
<tr>
<th>Model A</th>
<th>Risk Weight</th>
<th>Total Model Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model B</td>
<td>Risk Weight</td>
<td></td>
</tr>
<tr>
<td>Model C</td>
<td>Risk Weight</td>
<td></td>
</tr>
<tr>
<td>Model D</td>
<td>Risk Weight</td>
<td></td>
</tr>
</tbody>
</table>

The overall model risk is a function of inherent model risk and model risk mitigation activities, as illustrated in Figure 2. This process can be used to measure and monitor model risk for individual models, as well as in the aggregate, and for periodic reporting to senior management and board. To measure aggregate model risk for the entire model
The quantification and aggregation of model risk

inventory, an aggregation methodology can be applied. Aggregation can utilise weights to arrive at a more representative overall score. The resulting total is a subjective score; however, it can be utilised as a relative measure for period to period monitoring. Figure 3 is an illustration of how we would aggregate model risk across models. This involves assignment of a Model Risk Weight (MRW) to the score derived from each individual scorecard. There are various ways in which this MRW can be determined, for example we may have an aggregate RWA for clusters or types of models that is determined by expert judgement, and then within such clusters we may choose to weigh equally.

Advantages of the MRSA include that it is capable of:

- effectively aggregating model risk across institution;
- leveraging existing model inventory and risk classifications;
- clear, easy to understand reporting tools for senior management; and
- could be rolled out quicker than some alternative approaches.

Disadvantages of the MRSA include that:

- it is capable of aggregation but does not attempt to quantify model risk;
- its scores, weights, and overall results are subjective and based on judgement; and
- it possesses limited ability to capture model interdependencies.

The Advanced Operational Risk Approach (‘AORA’) is based in part on the event and scenario modelling techniques of the ‘AMA’ to operational risk. In this framework, model risk viewed as a function of the frequency and severity of potential model risk ‘events’, which can be analysed based on a combination of historical experience and scenario analysis. A database of historical model risk events is used to estimate the historical component of model risk. Model-specific risk considerations include common loss measure such as profit and loss (P&L), capital impact, etc. Data gathering considerations include that:

- sparsity of external data, as industry loss event databases, may not include model risk events or may not identify them as such;
- internal data may not be tagged as ‘model risk’ related; and
- consideration may be given to using back-testing or benchmarking data as proxies.

Similarly to AMA scenario process, workshops with model owners and users can be used to determine model risk event types, frequency, and severity of loss events. One may have workshops focus on capturing the universe of model-specific risk events, such as what could go wrong, and given and model risk event what could be the range and frequency of potential impacts.

All of this is highly subjective and such judgement-based process is common to the rest of AMA framework. In the industry, we note that attempts to capture potential model risk events have not yet occurred. We actually have such a database of model loss events shown in Figure 5 where we present a conceptual framework for the AORA modelling process. The left-most panel represents the data input component, which includes the hard data in our hypothetical model risk loss event database (which will contain the dates, frequencies, types and severities of the events – which may be both internal and
external), as well as the soft component of management expert judgement. The second panel from the left represents the frequency and severity of loss distributions, which may be segmented into ‘cells’ homogenous with respect to model risk, along the lines of loss event type and business unit. Typically we would fit a parametric probability distribution to these, as appropriately given the characteristics of the empirical data distribution (e.g. Poisson or compound Poisson for frequency; lognormal, Hyperbolic or Generalised Pareto for severity) and furthermore we may estimate separate distributions for the body vs. the tail of the distributions. The second from right-most panel of Figure 5 represents the simulation engine, whereby we combine the estimated distributions of frequency and severity, across model risk segments or cells, which is typically some kind of Monte Carlo simulation. A key modelling challenge in this regard is the specification of a dependency structure for the different model risk segments – we may leverage methodologies such as the Method of Copulae, with an eye to choosing a conservative copula, or we may be ultraconservative and simply choose to assume that the risks are co-monotonic and perform simple summations. Finally, the convoluted and combined total model risk distribution is shown in the right-most panel of Figure 5, from which we may derive a risk measure, such as an expected value or some high quantile of our preference.

Figure 5  Illustration of the advanced operational risk to model risk aggregation (see online version for colours)

Advantages of the AORA include that:

- it is empirically grounded, therefore subject to model validation, and likely to be favoured by regulators;
- it aggregates and quantifies model risk in a coherent manner;
- institutions already using AMA can leverage existing infrastructure and processes; and
- it may be used as input in estimating economic capital for model risk.
Disadvantages of the AORA include that:

- historical model risk event data tend to be sparse;
- scenario-based frequency and severity estimates highly subjective and based on judgement (which is common to all AMA models);
- distributional assumptions are subjective and can affect outcomes (which is also common to all AMA models); and
- more complex to implement, especially without existing AMA infrastructure.

**Figure 6** Illustration of the model uncertainty approach to model risk aggregation (see online version for colours)

The *Model Uncertainty Approach* (MUA) is a bottom-up, quantitative approach to aggregating model risk. The MUA is based on the principle of *model risk sensitivity*. Model uncertainty measurements are derived from benchmarking, back-testing, consideration of alternative modelling techniques, and the corresponding impact. Figure 6 illustrates the MUA framework from a conceptual overview vantage point. In the left-most panel, we depict the two types of output, those from our internally developed models and also from any benchmarking models that may be available. As we vary model inputs (e.g. data, assumptions, parameters, etc.), we can measure the variation in model output, which is depicted by the horizontal lines in the grid of the middle panel, resulting in the model uncertainty measurement in the right-most region of the graphic. We can see from this that the MUA approach utilises information from various sources and translates model risk into metrics, such as model output sensitivity to key modelling assumptions (i.e. model benchmarks, parallel run results or alternative models), alternative approach/methodologies, output benchmarks (internal or external data, including industry data or counterparty), back-testing results and model validation results and observations.

Consider an example of P&L and capital impact to demonstrate how model output variability can be traced to downstream metrics. We could conduct sensitivity analysis to compare outputs of a model and an alternative benchmark model. An example of this could be a historical VaR model compared to a simpler alternative methodology such as variance-covariance method. In terms of input benchmarks, we could perform sensitivity
analysis to compare model outputs based on data from external sources (e.g. alternative source, aggregated/consensus data). As an example, we could look at alternative prepayment speeds as obtained from an external source and used to assess the impact of varying this input. In terms of model validation findings, we could consider recalculation during the validation process that indicated a range of variability of model results based on key modelling assumptions, for example the impact on model outputs of interpolation settings such as using linear vs. spline.

Advantages of the MUA are that it:

- can achieve bottom-up quantification of model risk for individual models;
- enables a more rigorous and less subjective measurement of model risk;
- may be used to capture model interdependence, in combination with a detailed taxonomy of models and inputs; and
- may be used as input in estimating economic capital for model risk.

Disadvantages of the MUA are that:

- measurement and aggregation assumptions are highly subjective and based on judgement;
- development of individual, model-specific approaches can be more time and resource intensive; and
- maintenance relies on the ongoing performance of model control activities (i.e. benchmarking, back-testing, external data comparisons, etc.).

While there are many ways to measure and aggregate model risk, it is useful to define how such measurement system would respond to changes in key ‘MRFs’. Holding everything else constant, the following factors can be expected to increase or decrease aggregate model risk. Some hypothetical MRFs are illustrated in Figure 7, for example introduction of a new or change to an existing model augments model risk. Similarly, an increase in model exposure may be considered to increase model risk, as would a change in model use or an error identified model error. Issues in back-testing, such as exceptions outside normal range or having not performed during a certain period, would also be thought to result in greater model risk, as would a higher divergence from benchmarks or having not been validated for some time. On the other hand, positive risk factors believed commonly to diminish model risk include a model validation has recently been performed, periodic back-testing process initiated, additional model controls put in place, or neither a model retired nor having been in use for some time.

One challenge in aggregating total model risk is assessing the model interdependencies, which impacts the total model risk for a firm. Regulators expect firms to consider such interdependencies as part of the process, as the guidance clearly states that aggregate model risk is affected by interaction and dependencies among models; reliance on common assumptions, data, or methodologies; and any other factors that could adversely affect several models and their outputs at the same time (OCC and BOG-FRB SR 11-7, 2011). A simple approach would look at making ‘conservative’ assumptions in adding model risk outputs from different models, where appropriate
The quantification and aggregation of model risk

model risk across models as if the risks were perfectly correlated. A more advanced approach may involve expanding model inventory to create a detailed taxonomy of identify model and input dependencies:

- **Downstream models:** Determine impact originating from base model’s outputs.
- **Common inputs:** Assess the impact on multiple models sharing those inputs.

**Figure 7** Illustration of the model risk factors (see online version for colours)

4 **An example of model risk quantification:** alternative stress-testing methodologies

Specification of a model, definition of parameters or quantities of interest or specification of the parameter space all require justification which is an important part of the model validation procedure expected of financial institutions (see OCC and BOG-FRB SR 11-7, 2011). However, estimation of parameters after these judgements is made typically proceeds without regard for potential non-data information about the parameters, in an attempt to appear completely objective. Nevertheless, subject matter experts typically have information about parameter values, as well as about model specification. For example, a loss rate should lie between zero and one, or a dollar loss estimate for an institution should be no more than the value of assets, or the definition of the parameter space. However, if we are considering a loss rate for a particular portfolio segment, we in fact have a better idea of the location of the rate. The Bayesian approach allows formal incorporation of this information, i.e. formal combination of the data and non-data information using the rules of probability. In the context of stress-testing, we may take an institution’s or the regulators’ base scenarios or their adverse scenarios to represent such
non-data information as our Bayesian prior. Note that often when building a stress-testing model, the developer would be given this information exogenously with respect to the reference data at hand.

The Bayesian approach is most powerful and useful when used to combine data as well as non-data information while incorporating powerful computational techniques such as Markov Chain Monte Carlo methods. Such models are widely discussed in the economics and finance literature and have been applied in the loss estimation settings. These applications invariably specify a ‘prior’, which is convenient, and adds minimal information – there is no such thing as an uninformative prior – allowing computationally efficient data analysis. However, this approach, while valuable, misses the true power of the Bayesian approach, i.e. the coherent incorporation of expert information.

The difficulty in Bayesian analysis is the elicitation and representation of expert information in the form of a probability distribution, which requires thought and effort, rather than mere computational power; therefore, it is not commonly followed. Furthermore, in ‘large’ samples data driven information will typically overwhelm non-dogmatic prior information, so the prior is irrelevant asymptotically, and economists often justify ignoring prior information on this basis. However, there are many settings in which expert information is extremely valuable. In particular, cases in which data may be scarce, costly, or when its reliability is questionable. In the context of stress-testing, it is the case that scenarios may be hypothetical and not supported by observed historical data. Such issues more frequently arise in loss estimation, where sufficient data may not be available for low-default assets or for new financial instruments, or where structural economic changes may raise doubts about the relevance of historical data.

Empirical analysis in our paper follows the steps in a Bayesian analysis of a stress-testing model. Estimation of stressed losses rates for groups of homogeneous assets is essential for determining the amount of adequate capital under stressed scenarios. Since our goal is to incorporate non-data information in our Bayesian analysis, we utilise the supervisory stress-testing scenarios to elicit and represent expert information which is used to make inferences in the context of a simple model of loss. In this regard, we are aware that many institutions are moving away from simple linear regression frameworks for CCAR or DFAST, towards models such proportional hazards or rating migrations; nevertheless, regression-based techniques at an aggregated level are still rather prevalent in the industry, so that we think that there is value in using this as a starting point, and we can consider more advanced techniques for future directions of this research.

Our empirical analysis of the quantification of model risk stems from the implementation of a Bayesian methodology to stress-testing and model validation which follow the closely the CCAR program. As part of the Federal Reserve’s CCAR exercise, US domiciled top-tier Bank Holding Corporations (BHC) are required to submit comprehensive capital plans, including pro forma capital analyses, based on at least one BHC defined adverse scenario which is to be defined by quarterly trajectories for key macroeconomic variables over the next nine quarters or longer, to estimate loss allowances. In addition, the Federal Reserve generates its own supervisory stress scenarios, so that firms are expected to apply both BHC and supervisory stress scenarios to all exposures, in order to estimate potential losses under stressed operating conditions. Separately, firms with significant trading activity are asked to estimate a one-time potential trading-related market and counterparty credit loss shock under their own BHC scenarios, and a market risk stress scenario provided by the supervisors. In the case of the supervisory stress scenarios, the Federal Reserve provides firms with global market shock components that are one-time hypothetical shocks to a large set of risk factors.
The quantification and aggregation of model risk

Table 1 lists the macroeconomic variables used in supervisory stress-testing scenarios as part of the Federal Reserve CCAR program. Our analysis of using the macroeconomic stress scenarios to inform historical analysis is based on the collection of the last three Fed scenarios for the three macroeconomic variables over these nine-quarter periods, i.e. RGDPYY – Real Gross Domestic Product (year-to-year change), UNEMP – Unemployment Rate, and HPI – National Housing Price Index. We justify focusing on these three variables as they are the most commonly used, and the most accepted by regulators, as well as having good explanatory power for the target loss variables that we are considering in this study. We consider the two CCAR supervisory stress scenarios in 2011 and 2012, and the supervisory severely adverse scenario in 2013, focusing on Aggregate Bank Gross Charge-Offs (ABCO) from the Fed Y9 report as a measure of loss. Our historical data set covers the period from 2000 to 2013. To the best of our knowledge, this is the first study of its kind which combines data to form the prior three supervisory exercises in stress-testing within Bayesian framework. The reason why we base the prior distributions upon the supervisory scenarios is a practical one, as often modellers will use the quality of the supervisory scenarios as a criterion in model development, for example a common practice being testing the redeveloped model with the prior years’ scenarios. Of course, model developers have the option to use their own internally developed scenarios, or information gleaned from subject matter experts such as the lines of business, in order to form their priors.

Table 1 | Federal Reserve Comprehensive Capital Analysis and Review (CCAR) program macroeconomic variables used in supervisory stress-testing scenarios

<table>
<thead>
<tr>
<th>CCAR date</th>
<th>Macroeconomic variable (MV)</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>Real GDP (year-to-year)</td>
<td>RGDPYY</td>
</tr>
<tr>
<td></td>
<td>Consumer price index</td>
<td>CPI</td>
</tr>
<tr>
<td></td>
<td>Real disposable personal income</td>
<td>RDPI</td>
</tr>
<tr>
<td></td>
<td>Unemployment rate</td>
<td>UNEMP</td>
</tr>
<tr>
<td></td>
<td>Three-month treasury bill rate</td>
<td>3MTBR</td>
</tr>
<tr>
<td></td>
<td>Ten-year treasury bond rate</td>
<td>10YTBR</td>
</tr>
<tr>
<td></td>
<td>BBB corporate rate</td>
<td>BBBCR</td>
</tr>
<tr>
<td></td>
<td>Dow Jones index</td>
<td>DJI</td>
</tr>
<tr>
<td></td>
<td>National housing price index</td>
<td>HPI</td>
</tr>
<tr>
<td>2012</td>
<td>All 2011 CCAR MVs +</td>
<td>RGDPG</td>
</tr>
<tr>
<td></td>
<td>Nominal GDP growth</td>
<td>NDPIG</td>
</tr>
<tr>
<td></td>
<td>Nominal disposable income growth</td>
<td>MR</td>
</tr>
<tr>
<td></td>
<td>Mortgage rate</td>
<td>CBOE’s market volatility index</td>
</tr>
<tr>
<td></td>
<td>Commercial real estate price index</td>
<td>CREPI</td>
</tr>
<tr>
<td>2013</td>
<td>All 2011 and 2012 CCAR MVs</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 presents the summary statistics of and correlation between the macroeconomic variables for both our historical data set and the Fed scenario we use in our empirical analysis. Figure 8 displays the time series and kernel density plots of these variables for both data sets. ABCO averages 70 bps over the last 14 years since 2000, peaking at 2.72% towards the end of 2009. ABCO is extremely skewed towards periods of mild loss...
during early 2000s, having a mode of around 20–25 bps. RGDPYY historically averages 1.94%, while in the Fed stress scenarios it displays an average contraction of –0.73%, having mild positive and negative skews in historical and Fed scenario data, respectively. Figure 8 shows that while the historical distribution of RGDPYY is bimodal, having modes at around 5% and 9% which represents the historical regime shift between expansionary and contractionary periods, RGDPYY’s Fed scenario distribution has a single mode at around zero. UNEMP has a historical average of 6.4% (ranging from 4% to 10%), while in the Fed scenarios it is centred at 11.9% (ranging from 10% to 14%). As with RGDPYY, UNEMP displays a bimodal distribution, with modes of 4% and 9% (10% and 14%) considering the historical data (data from Fed scenarios). The historical average of HPI has a historical average of 150.2 (ranging from 101.6 to 199.0), while in the Fed scenarios it is centred at 129.0 (ranging from 112.8 to 142.4). As with the other macroeconomic variables, HPI displays a bimodal (unimodal) distribution, with modes of 140 and 180 (135) considering the historical data from Fed scenarios.

Table 2  Summary statistics and correlations macroeconomic variables and bank charge-offs

<table>
<thead>
<tr>
<th>Macroeconomic variables</th>
<th>Statistic</th>
<th>ABCO</th>
<th>RGDPYY</th>
<th>UNEMP</th>
<th>HPI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Historical data</strong></td>
<td>Count</td>
<td>56</td>
<td>56</td>
<td>56</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>0.70</td>
<td>1.94</td>
<td>6.39</td>
<td>150.21</td>
</tr>
<tr>
<td></td>
<td>Standard deviation</td>
<td>0.79</td>
<td>1.90</td>
<td>1.88</td>
<td>26.63</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>0.01</td>
<td>–4.09</td>
<td>3.92</td>
<td>101.60</td>
</tr>
<tr>
<td></td>
<td>25th percentile</td>
<td>0.11</td>
<td>1.43</td>
<td>4.81</td>
<td>136.23</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.22</td>
<td>2.10</td>
<td>5.77</td>
<td>144.45</td>
</tr>
<tr>
<td></td>
<td>75th percentile</td>
<td>1.29</td>
<td>3.12</td>
<td>8.05</td>
<td>166.60</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>2.73</td>
<td>5.27</td>
<td>9.91</td>
<td>199.00</td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>1.10</td>
<td>–1.48</td>
<td>0.56</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>–0.11</td>
<td>2.84</td>
<td>–1.09</td>
<td>–0.64</td>
</tr>
<tr>
<td><strong>Fed scenarios</strong></td>
<td>Count</td>
<td>21</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>–0.73</td>
<td>11.90</td>
<td>129.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Standard deviation</td>
<td>2.02</td>
<td>1.73</td>
<td>9.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>–4.31</td>
<td>9.58</td>
<td>112.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25th percentile</td>
<td>–1.96</td>
<td>10.54</td>
<td>120.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>–0.58</td>
<td>11.49</td>
<td>128.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>75th percentile</td>
<td>0.72</td>
<td>13.65</td>
<td>136.90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>2.14</td>
<td>14.31</td>
<td>142.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>–0.37</td>
<td>0.05</td>
<td>–0.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>–0.99</td>
<td>–1.62</td>
<td>–1.19</td>
<td></td>
</tr>
<tr>
<td><strong>Correlations</strong></td>
<td>ABCO</td>
<td>100%</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>RGDPYY</td>
<td>–54.35%</td>
<td>100%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>UNEMP</td>
<td>88.55%</td>
<td>–37.91%</td>
<td>100%</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>HPI</td>
<td>–14.91%</td>
<td>3.90%</td>
<td>–17.18%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Note: ABCO: Aggregate bank gross charge-offs; RGDPYY: year-on-year change in real gross domestic product; UNEMP: unemployment rate; HPI: national housing price index.
Figure 8  Time series and kernel density plots historical vs. Fed scenario macroeconomic variables
We estimate univariate, bivariate, and trivariate Bayesian as well as Frequentist models for each of the three macroeconomic variables by forming priors using univariate regressions. In our empirical implementation our data sample is the history and the prior sample is formed from the last three pooled Fed scenarios. Dependent variables for prior regressions are established by calculating the historical quantile of each macroeconomic variable based upon the scenario data set, and using the historical value of ABCO at that quantile as the response variable.

Our empirical results for all regressions are presented in Table 3 and Figure 9. In order to conserve space, we only display the density plots and posterior distributions of trivariate Bayesian regressions for ABCO vs. each variable, identifying corresponding macro-sensitivity.

The results for estimating the posterior distributions on macro-sensitivities in univariate regression are presented in Panel A of Table 3. The historical data (Fed scenarios) estimate of beta coefficient for RGDPYY is –0.2267 (–0.3953) resulting in a posterior estimate of –0.2500 which is slightly higher in absolute value. In the case of UNEMP, coefficient estimates are 0.3735, 0.4263, and 0.4 for historical, Fed scenario, and posterior, respectively. For HPI, historical data estimate of the coefficient is –0.0044, while for the Fed scenarios it is –0.0271 which translates into a posterior estimate of –0.0150 that is much higher in absolute value. Therefore, in general the posterior estimates are greater in absolute value, reflecting greater sensitivity as observed in the scenario data set that informs the prior. Note that there is no loss of generality, as model developers could use their own priors formed from internal views on scenarios or expert opinion in lieu of prior Fed scenarios, and sensitivities may be reduced as well.

Panel B of Table 3 presents the results for bivariate regressions and similar to univariate results Fed scenarios have greater sensitivity than historical estimates. In the case of RGDPYY and UNEMP pair, while posterior estimate for RGDPYY (–0.2147) is higher in absolute value, it is counter-intuitively lower for UNEMP (0.0062). One possible reason for this is that the Fed scenario data set shows a correlation between UNEMP and RGDPYY relative to the historical pattern historically that is such that the posterior estimate is pulled in an unintuitive direction. When we consider the UNEMP and HPI pair, we observe that the posterior estimate for unemployment lower in absolute value (0.3036 vs. 0.3738), which we also find to be counter-intuitive, explains similarly as in the case of the UNEMP vs. RGDPYY pair. The posterior estimate for the HPI sensitivity is estimated to be larger in absolute value (|–0.0073| vs. |0.0001|). In the case of RGDPYY and HPI, the historical coefficient data estimate for RGDPYY (HPI) is –0.2248 (–0.0038), and based on the prior scenarios having greater sensitivity of –0.3953 (–0.0271). For both macro-variables the posterior estimates are found to be higher in absolute value (|–0.2435| vs. |–0.0091|).

As with the univariate and bivariate regressions, we find for the trivariate model that in general the absolute value of macro-sensitivities is greater in magnitude of the Fed scenario regressions than in historical ones. The results for trivariate model estimation are presented in Panel C of Table 3. For RGDPYY, the historical data (Fed scenarios) estimate of beta coefficient is –0.1014 (–0.3953) and the posterior estimate is –0.1463, which is higher in absolute value. In the case of UNEMP, the historical data (Fed scenarios) estimate of beta coefficient is 0.3345 (0.4263, which is indicative of greater
The quantification and aggregation of model risk

sensitivity). The posterior parameter estimate for UNEMP is counter-intuitively low (0.2594). Considering the housing index (HPI), coefficients are −0.0001 and −0.0065, for historical data and posterior estimates, respectively.

In Figure 10, we present the Bayesian and Frequentist modelled as well as the historical loss rates. In addition, forecasted scenario loss rates (Bayesian vs. Frequentist modelled in the case of the severely adverse scenarios and the Frequentist modelled for the base case). We observe that while the models tend to under- (over-) predict in the stress (recent benign) period, optically the Bayesian model actually performs worse in the stress period and better in the recent period. However, in the severe adverse scenario, the Bayesian modelled losses reach more extreme levels than those Frequentist modelled ones – this is a good property from a supervisory perspective, as it reflects model monitoring or back-testing, which is not only purely based upon data but also incorporates prior views on model parameters.

Table 3  Estimation of stress-testing macroeconomic models Bayesian with Fed scenario priors vs. historical models GDP, unemployment rate, and housing index

<table>
<thead>
<tr>
<th>Modeled as</th>
<th>RGDPYY</th>
<th>UNEMP</th>
<th>HPI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: Univariate regressions – posterior distribution (β)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Historical model</td>
<td>−0.2267</td>
<td>0.3735</td>
<td>−0.0044</td>
</tr>
<tr>
<td>Fed scenarios</td>
<td>−0.3953</td>
<td>0.4263</td>
<td>−0.0271</td>
</tr>
<tr>
<td>Bayesian model</td>
<td>−0.2500</td>
<td>0.4000</td>
<td>−0.0150</td>
</tr>
<tr>
<td><strong>Panel B: Bivariate regressions – posterior distribution (β)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Historical model – RGDPYY and UNEMP</td>
<td>−0.1013</td>
<td>0.3347</td>
<td></td>
</tr>
<tr>
<td>Bayesian model – RGDPYY and UNEMP</td>
<td>−0.2147</td>
<td>0.0062</td>
<td></td>
</tr>
<tr>
<td>Historical model – UNEMP and HPI</td>
<td>0.3738</td>
<td>0.0001</td>
<td></td>
</tr>
<tr>
<td>Bayesian model – UNEMP and HPI</td>
<td>0.3036</td>
<td>−0.0073</td>
<td></td>
</tr>
<tr>
<td><strong>Panel C: Trivariate regressions – posterior distribution (β)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Historical model – RGDPYY, UNEMP and HPI</td>
<td>−0.1014</td>
<td>0.3345</td>
<td>−0.0001</td>
</tr>
<tr>
<td>Bayesian model – RGDPYY, UNEMP and HPI</td>
<td>−0.1463</td>
<td>0.2594</td>
<td>−0.0065</td>
</tr>
</tbody>
</table>

Note: T-statistics are presented in parenthesis with significant ones identified in bold; RGDPYY: year-on-year change in real gross domestic product; UNEMP: unemployment rate; HPI: national housing price index.
Figure 9  Density plots and posterior distributions of trivariate Bayesian regressions for aggregate bank gross charge-offs (2001–2013) (see online version for colours)

Figure 10  Historical, scenario and Bayesian vs. Frequentist regression modelling aggregate bank gross charge-off rates
Figure 11  Panel A: Posterior conditional distributions of Bayesian regression modelling quarterly aggregate bank gross charge-off rates; and Panel B: Posterior conditional distribution of Bayesian regression modelling cumulative nine-quarter aggregate bank gross charge-off rates
In Table 4, we summarise the conditional severely adverse loss distributions of the Frequentist regression model, for both quarterly and cumulative nine-quarter ABCO rates. Classical Frequentist Coefficients of Variation (FCV) are the simple ratio of the Frequentist 95% confidence interval (F95CI) to the mean of the sampling distribution:

\[
FCV_{95}^x = \frac{(\bar{x} + 1.96 \times s_x) - (\bar{x} - 1.96 \times s_x)}{\bar{x}} = \frac{F95CI_x}{\bar{x}}
\]

where \( S_x \) is the standard error of the forecast mean. Losses average a range of 0.9–1.7% in the first two quarters, peaking at a mean ranging in 2.3–3.6% and reverting to a mean of 1.2–1.7% in the final two quarters. We observe that the variability in the loss distribution displays a U-shape, peaking in the low loss early and end quarters: the standard error drops from a range of 0.34–0.38%, to a range of 0.28–0.34%, and then rises to a range of 0.41–0.46%. This observation on the pattern in variability over the forecast horizon holds on a relative basis as well, of 32–56%, and then rising to a range of 94–153% in the first two, middle and last two quarters, respectively. The FCV measure can be interpreted as the proportional model risk uncertainty buffer, stemming from the sampling error as inferred from the Frequentist regression model.

We contribute to model validation literature by comparing the proportional model risk buffer measures obtained from our empirical implementation of the Bayesian to the Frequentist models. One common way to estimate a model risk buffer is measuring of statistical uncertainty generated by a model, such as a standard error or a confidence interval; other means of quantifying this metric include sensitivity analysis around model inputs or model assumptions, i.e. varying the latter and measuring the variability of the model output. The model risk buffer is a valuable model validation tool, as it helps us to understand the potential expected variability in model output, e.g. when we perform model benchmarking or back-testing, we can gauge if new observation of actuals is lying in an expected range, and this can serve as a basis for remedial actions such as model overlays.

The mean of the posterior distribution in the nine-quarter severely adverse loss generated by the Bayesian model is 43.2%, with a Bayesian 95% credible interval of 11.0%, resulting in a BNCV of 25.5%. The mean of the sampling distribution in the nine-quarter severely adverse loss generated by the Frequentist model is 20.6%, with a classical 95% confidence interval of 4.1%, resulting in an FCV of 20.0%. Therefore, our Bayesian analysis suggests that a quantitatively developed model risk uncertainty buffer to account for parameter uncertainty that is 5% (20%) is higher in absolute (relative) terms than that implied by the Frequentist model.

We compare the Bayesian and Frequentist stress-testing models according to several measures of model performance, as commonly used in model validation exercises. First, we use the Root Mean Squared Error (RMSE), which measures the average squared deviation of model predictions from actual observations:

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (P_i - A_i)^2}
\]
The quantification and aggregation of model risk

where $P_i$ are predicted and $A_i$ actual. Secondly, we calculate squared-correlation (SC) between model predictions and actual observations:

$$SC = \left( \frac{\sum_{i=1}^{N} \left[ (P_i - \frac{1}{N} \sum_{i=1}^{N} P_i) (A_i - \frac{1}{N} \sum_{i=1}^{N} A_i) \right]^2}{\left( \frac{1}{N} \sum_{i=1}^{N} (P_i - \frac{1}{N} \sum_{i=1}^{N} P_i)^2 \right) \left( \frac{1}{N} \sum_{i=1}^{N} (A_i - \frac{1}{N} \sum_{i=1}^{N} A_i)^2 \right)} \right)^{\frac{1}{2}}$$

(3)

$$CPE = \frac{\sum_{i=1}^{N} (P_i - A_i)}{\sum_{i=1}^{N} A_i}$$

(4)

Figure 12 Posterior conditional distributions of Bayesian regression model quarterly aggregate bank gross charge-off rates

Finally, we consider a measure widely used in model validations of stress-testing models for CCAR or DFAST, the Cumulative Percentage Error (CPE), which is favoured by prudential regulators.
Table 4

<table>
<thead>
<tr>
<th></th>
<th>Quarter 1</th>
<th>Quarter 2</th>
<th>Quarter 3</th>
<th>Quarter 4</th>
<th>Quarter 5</th>
<th>Quarter 6</th>
<th>Quarter 7</th>
<th>Quarter 8</th>
<th>Quarter 9</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Minimum</strong></td>
<td>0.41</td>
<td>0.99</td>
<td>1.32</td>
<td>1.61</td>
<td>1.74</td>
<td>1.75</td>
<td>1.53</td>
<td>0.84</td>
<td>0.00</td>
<td>21.01</td>
</tr>
<tr>
<td>2.5th percentile</td>
<td>0.65</td>
<td>1.39</td>
<td>2.29</td>
<td>2.95</td>
<td>2.94</td>
<td>2.47</td>
<td>2.08</td>
<td>1.16</td>
<td>0.19</td>
<td>37.80</td>
</tr>
<tr>
<td>25th percentile</td>
<td>0.75</td>
<td>1.61</td>
<td>2.84</td>
<td>3.74</td>
<td>3.63</td>
<td>2.87</td>
<td>2.30</td>
<td>1.38</td>
<td>0.59</td>
<td>41.32</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>0.80</td>
<td>1.73</td>
<td>3.13</td>
<td>4.15</td>
<td>3.99</td>
<td>3.08</td>
<td>2.42</td>
<td>1.50</td>
<td>0.80</td>
<td>43.23</td>
</tr>
<tr>
<td>75th percentile</td>
<td>0.86</td>
<td>1.85</td>
<td>3.41</td>
<td>4.56</td>
<td>4.35</td>
<td>3.29</td>
<td>2.54</td>
<td>1.62</td>
<td>1.00</td>
<td>45.12</td>
</tr>
<tr>
<td>97.5th percentile</td>
<td>0.96</td>
<td>2.07</td>
<td>3.99</td>
<td>5.37</td>
<td>5.06</td>
<td>3.71</td>
<td>2.77</td>
<td>1.86</td>
<td>1.41</td>
<td>48.81</td>
</tr>
<tr>
<td><strong>Maximum</strong></td>
<td>1.12</td>
<td>2.46</td>
<td>4.77</td>
<td>6.42</td>
<td>6.04</td>
<td>4.39</td>
<td>3.31</td>
<td>2.24</td>
<td>2.01</td>
<td>55.24</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>0.12</td>
<td>0.14</td>
<td>0.09</td>
<td>0.09</td>
<td>0.10</td>
<td>0.19</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
<td>0.41</td>
</tr>
<tr>
<td>Standard error</td>
<td>0.08</td>
<td>0.17</td>
<td>0.43</td>
<td>0.61</td>
<td>0.54</td>
<td>0.32</td>
<td>0.18</td>
<td>0.18</td>
<td>0.31</td>
<td>2.82</td>
</tr>
<tr>
<td>95th percentile</td>
<td>0.30</td>
<td>0.67</td>
<td>1.69</td>
<td>2.42</td>
<td>2.12</td>
<td>1.25</td>
<td>0.70</td>
<td>0.70</td>
<td>1.23</td>
<td>11.01</td>
</tr>
<tr>
<td>95th percentile</td>
<td>37.61</td>
<td>38.97</td>
<td>54.10</td>
<td>58.42</td>
<td>53.26</td>
<td>40.57</td>
<td>28.85</td>
<td>46.31</td>
<td>153.48</td>
<td>25.49</td>
</tr>
</tbody>
</table>

Table 4: Posterior conditional distributions of Bayesian regression model quarterly and cumulative nine-quarter aggregate bank gross charge-off summary statistics and numerical coefficients of variation.
Table 5

Distributions of Frequentist regression modelling quarterly and nine-quarter cumulative aggregate bank gross charge-off rates summary statistics and parametric coefficients of variation

<table>
<thead>
<tr>
<th></th>
<th>Quarter 1</th>
<th>Quarter 2</th>
<th>Quarter 3</th>
<th>Quarter 4</th>
<th>Quarter 5</th>
<th>Quarter 6</th>
<th>Quarter 7</th>
<th>Quarter 8</th>
<th>Quarter 9</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected value</td>
<td>0.92</td>
<td>1.71</td>
<td>2.78</td>
<td>3.56</td>
<td>3.49</td>
<td>2.84</td>
<td>2.38</td>
<td>1.71</td>
<td>1.20</td>
<td>20.58</td>
</tr>
<tr>
<td>Standard error of forecast</td>
<td>0.38</td>
<td>0.34</td>
<td>0.29</td>
<td>0.29</td>
<td>0.28</td>
<td>0.30</td>
<td>0.34</td>
<td>0.41</td>
<td>0.47</td>
<td>1.05</td>
</tr>
<tr>
<td>95th percentile Frequentist confidence interval</td>
<td>1.47</td>
<td>1.33</td>
<td>1.14</td>
<td>1.13</td>
<td>1.11</td>
<td>1.19</td>
<td>1.34</td>
<td>1.60</td>
<td>1.83</td>
<td>4.11</td>
</tr>
<tr>
<td>95th percentile Frequentist coefficient of variation</td>
<td>160.67</td>
<td>77.78</td>
<td>41.16</td>
<td>31.64</td>
<td>31.95</td>
<td>42.02</td>
<td>56.48</td>
<td>93.74</td>
<td>153.44</td>
<td>19.99</td>
</tr>
</tbody>
</table>
We estimate these model performance measures, in-sample, and across the entire historical period (2001–2013) and over the 12 quarter downturn period (2006–2009). In addition, we estimate the sampling distributions of these measures using a bootstrap procedure, in order to test the statistical significance of the observed differences in model performance measures. We observe that the Frequentist model outperforms the Bayesian model according to RMSE and SC measures (10.1% and 84.5% vs. 15.4% and 75.2%, in mean, for the entire sample; 13.9% and 92.0% vs. 19.6% and 87.0%, in mean, for the downturn sample). However, the Bayesian model outperforms, over the entire sample as well as during the stressed period, according to the CPE measure (7.9% and 5.9%, in mean, for the entire sample; –13.7% and –12.1%, in mean, for the downturn sample). It is not surprising that the Frequentist model performs better when RMSE and SC measures are used for validation, since it is a model which is purely calibrated to the historical data. The reason for Bayesian approach’s superior performance, using the CPE-preferred measure of model validators and supervisors, is that this model constrains the regression coefficients to exhibit more sensitivity, so that when there are large losses, the model matches actuals to a great degree than when losses are towards the middle of the distribution – intuitively, we are able to better match the tails of the error distribution than its body. In contrast, the Frequentist regression model simply tries to minimise the total squared deviation over the entire sample, which is modelling the body but not the tail of the error distribution. Note, moreover, that we could also impose alternative priors – e.g. informed by external data, internal scenarios or expert opinion – which could either accentuate this effect or even work in the opposite direction and dampen sensitivities.

5 Conclusion and future directions

This study has contributed to the evolution of the MRM field, by contributing to a movement from efforts focused primarily on MRM for individual models, to a more advanced stage where institutions focus on the aggregation of firm-wide model risk. This recognises that MRM regulatory guidance specifically focuses on both measuring risk in individual and in the aggregate. We recognise that models can be tremendously helpful to financial services firms. As the use of models increases, however, so does model risk. This is an area in which the potential impact of flawed modelling can be great, and it is also an area receiving more regulatory attention and scrutiny. Firms can benefit from a structured approach to model risk that incorporates both a framework for developing and testing of models and effective governance of matters such as model validation, inventory and aggregation. When properly designed and implemented, models should be a valuable resource for financial institutions, but firms need well-conceived programs to improve models’ utility while identifying, quantifying and mitigating their potential risk.

This study has discussed various approaches to measuring and aggregating model risk across an institution. We also present an example of model risk quantification in the realm of stress-testing, where we compare alternative models in two different classes, Frequentist and Bayesian approaches to modelling stressed bank losses. We introduce the Bayesian approach to credit risk, present empirical results from an application of Bayesian model to stress-testing, as well as conduct a model validation exercise. Our implementation uses the Fed supervisory scenarios to develop prior distributions for the macro-factor sensitivities. We contribute to model validation literature by comparing the proportional model risk buffer measures obtained from our empirical implementation of the Bayesian versus the Frequentist models.
The quantification and aggregation of model risk

Using aggregate bank charge-offs from the Fed Y9 data set, we find that the proportional model risk buffer for parameter uncertainty measure as implied by the Bayesian model is approximately 25% greater than the analogue in the Frequentist regression model. This result indicates that our Bayesian approach results in more conservative model risk buffer metrics, since Bayesian modelling allows us to explicitly account for parameter uncertainty, and, moreover, we are able to explicitly incorporate prior information as represented by previous years’ supervisory scenarios in the estimation of a stress loss forecasting model. We find that the model risk buffer to account for parameter uncertainty in the severely adverse cumulative nine-quarter loss estimate, as inferred from the Bayesian model (25.5%), is about 25% higher than the equivalent measure in the Frequentist model (20.4%). This finding implies that, by not considering the prior information in supervisory forecasts, a stress-testing exercise may be severely understating the quantity of this kind of model risk in the buffer developed for parameter uncertainty. Furthermore, we find that relative model uncertainty metric behaves differently in the Bayesian than in the Frequentist estimations, displaying a humped shaped vs. a U-shaped pattern across quarters, respectively.

We contribute to the literature by developing a Bayesian-based credit risk stress-testing methodology which can be implemented by small- to medium-sized banking institutions, as well as presenting empirical results of methodology using data from the recent CCAR implementations. Our methodology as well as empirical findings should be of interest not only to risk managers in such financial institutions but also to domestic and international regulators.

References


Notes

1 More precisely, model risk is defined as the risk that a model is faulty because either it does not capture the correct risk factors (model misspecification), does not correctly establish the relationship between risk factors and the risk being measured, or that the model is calibrated with faulty data or implemented with error.

2 For background on special issues in LGD estimation refer to Araten et al. (2004) and Jacobs and Karagozoglu (2011).

3 In addition, large custodian banks are asked to estimate a potential default of their largest counterparty. For the last two CCAR exercises, these shocks involved large and sudden changes in asset prices, rates, and CDS spreads that mirrored the severe market conditions in the second half of 2008.

4 A best practice would be to have out-of-sample or out-of-time model performance metrics, but paucity of data precludes that in this case, as it does in most CCAR and DFAST stress-testing modelling validations. However, these in-sample measures represent the current state of the practice and are in line with supervisory expectations.