

MONTE CARLO SCENARIO GENERATION FOR RETAIL LOAN PORTFOLIOS

JOSEPH L BREEDEN AND DAVID INGRAM*

Abstract. Monte Carlo simulation is a common method for studying the volatility of market traded instruments. It is less employed in retail lending, because of the inherent nonlinearities in consumer behavior. In this paper, we leverage the approach of Dual-time Dynamics to separate loan performance dynamics into three components: a maturation function of months-on-books, an exogenous function of calendar date, and a quality function of vintage origination date. Of these three, the exogenous function captures the impacts from the macroeconomic environment. As such, we might naturally want to generate scenarios for the possible futures of these environmental impacts.

To generate such scenarios, we must go beyond the random walk methods most commonly applied in the analysis of market-traded instruments. Retail portfolios exhibit autocorrelation structure and variance growth with time that requires more complex modeling than a random walk. This paper describes work using ARMA and ARIMA models for scenario generation, rules for selecting the correct model given the input data, and validation methods on the scenario generation. We find when the goal is capturing the future volatility via Monte Carlo scenario generation, that model selection does not follow the same rules as for forecasting. Instead, test more appropriate to reproducing volatility are proposed, which assure that distributions of scenarios have the proper statistical characteristics. These results are supported by studies of the variance growth properties of macroeconomic variables and theoretical calculations of the variance growth properties of various models. We also provide studies on historical data showing minimum training lengths and differences by macroeconomic epochs.

Key words. Economic Capital, Retail Lending, Monte Carlo Simulation, Dual-time Dynamics

1. Introduction. When modeling retail loan portfolios, making a forecast for what one expects to happen in the future is confounded by external drivers on performance. Every forecast, whether implicitly or explicitly, is incorporating a scenario for what the future environment will be. Intuitively, we know that the macroeconomic environment affects consumers and therefore delinquency and default rates. Although we might explicitly incorporate a scenario for the future macroeconomic environment, our uncertainty in that scenario is naturally quite high. Predicting the economy is significantly more difficult than predicting the retail loan portfolio.

Scenario-based forecasting models are better suited to modeling retail loan portfolios. Rather than attempting to predict both the economy and the portfolios's delinquency response in a single composite model, we accept a scenario for the economy and model the portfolio's response. One can use in-sample / out-of-sample tests to validate the forecasting model, but measuring the uncertainty of the macroeconomic scenario presents greater challenges. In this paper, we investigate the use of a Monte Carlo scenario generator to create a distribution of plausible scenarios centered around the "most-likely" scenario over the forecast period provided by the analyst. From this distribution we can measure probability bands and the likelihood of extreme events. This is a point-in-time approach to creating the distribution. Methods that randomly sample the performance of any historic period are described as through-the-cycle models, and are not considered here.

The recent emphasis on computing economic capital has also brought a renewed emphasis on scenario-based forecasting. The dynamics underlying retail portfolios are far from simple linear systems. For example, a model to predict Net Default Loss for purposes of measuring economic capital might employ the following key rates:

*Strategic Analytics Inc., 2935 Rodeo Park Drive East, Santa Fe, New Mexico, 87505 (breeden@strategicanalytics.com, dingram@strategicanalytics.com).

$$(1.1) \quad DR(t) = DefaultAccounts(t)/ActiveAccounts(t-1)$$

$$(1.2) \quad EAD(t) = Exposure(t)/DefaultAccounts(t)$$

$$(1.3) \quad LGD(t) = (Exposure(t) - Recoveries(t))/Exposure(t)$$

$$(1.4) \quad AAR(t) = ActiveAccounts(t)/ActiveAccounts(t-1)$$

Where DR is the loan default rate (the aggregate version of probability of default, PD), EAD is the exposure at default, LGD is the loss given default, and AAR is the active account rate. The latter is not a member of the standard Basel II set, but we find it useful in creating long range forecasts incorporating attrition or prepayment. Each of these metrics is measured and modeled at the vintage level using the Dual-time Dynamics (DtD) approach described in Breeden, 2007 [6]. That approach has strong similarities to Age-Period-Cohort models [20, 15], Two-Way Proportional Hazards Models [11], and Generalized Additive Models (GAM) [18].

Following the method of DtD, we analyze all of the available vintages simultaneously in order to decompose each of the key rates using the following formulation

$$(1.5) \quad r(a, t, v) = f_m(a) * f_g(t) * f_q(v),$$

where r is the vintage performance rate, $f_m(a)$ is the maturation function of months-on-book, a , $f_g(t)$ is the exogenous function of calendar date, t , and $f_q(v)$ is the quality function of vintage. We have the further relationship that $a = t - v$. The functions $f_m(a)$, $f_g(t)$, and $f_q(v)$ are estimated from a set of vintages by assuming the relationship in equation 1.5 and solving for the unknown functions non-parametrically. The decomposition does not require any external data such as macroeconomic data or credit scores. Instead, we leverage the structure present in the vintage performance data.

For the rates defined in Equations 1.4, the result is four exogenous curves $f_g(t)$. Forecasting occurs by extrapolating those exogenous curves over the desired forecast period. Extrapolation can occur via business intuition, models related to macroeconomic factors, or in this context via Monte Carlo scenario generation. Once a scenario is provided, it is combined with the known maturation and quality functions to forecast each vintage.

Forecasts for the key rates are combined via a simulation layer to obtain forecasts for Net Default Losses.

$$(1.6) \quad ActiveAccounts(t) = \widetilde{AAR}(t) * ActiveAccounts(t-1)$$

$$(1.7) \quad DefaultAccounts(t) = \widetilde{DR}(t) * ActiveAccounts(t-1)$$

$$(1.8) \quad Exposure(t) = \widetilde{EAD}(t) * DefaultAccounts(t)$$

$$(1.9) \quad NetDefaultLosses(t) = \widetilde{LGD}(t) * Exposure(t)$$

Because the maturation and quality functions have significant nonlinear structure, the final Net Default Loss forecast is a combination of many nonlinear functions. For that reason, we cannot determine easily what the final loss distribution will look like just by considering the input distributions for the exogenous curves. Instead, we use a Monte Carlo approach where we generate a scenario for the exogenous curves, forecast the portfolio given that scenario, and collect the distribution of the final Net Default Loss forecasts. Economic capital can be read off of the cumulative distribution function (CDF) for Net Default Losses at the desired solvency level.

The scenarios being generated also have other uses beyond computing economic capital. Portfolio managers will often ask to see a "worst case" scenario. Using this framework, we can pull individual scenarios from the CDF to show the environment required to create a one-sigma (84%), two-sigma (97.7%), or three-sigma (99.9%) loss event.

In stress testing applications, economic data providers will often provide a "worst case" scenario, but without providing a specific probability of occurrence or severity measure. By comparing the impact of the economic scenario to the Monte Carlo distribution, we can derive a calibration of how severe the scenario is relative to the range of possible scenarios.

2. Monte Carlo Simulation. Monte Carlo simulation is a standard approach in derivatives pricing, mortgage pricing [5], many other areas of finance [21] and beyond [10]. In the majority of these applications, the models can be quite complex, but the Monte Carlo Scenario Generator (MCSG) is little more than a random number generator. Emphasis is placed upon creating the correct distribution from which to sample, but overall the MCSG is trivial.

Market traded instruments illustrate well why the MCSG can be simple and effective. Models of tradeable instruments are usually driven by scenarios for the series of returns for the instruments being modeled. Analysis consistently shows that although the distribution of returns can be skewed and fat-tailed, the returns are temporally independent, i.e. no autocorrelation structure exists above the transaction cost floor. Of course, not all studies take the simple approach. Bollerslev [4] and French, Schwert, and Stambaugh [14] were among the first to consider correlated noise processes, such as MA(1) and ARCH / GARCH, for modeling stock returns.

Our situation is different from the tradeable instruments examples. The simulation model in Equations 1.5 and 1.9 is trivial when compared to derivative pricing models, but the scenarios for the exogenous curves that will drive our simulation are more complex. Simple tests show that the exogenous curves typically exhibit significant autocorrelation structure. The exogenous curves are primarily capturing the response of consumers to external macroeconomic conditions and account management policy changes. As such, their behavior is not "traded" and therefore no arbitrage models are in place to trade-away the autocorrelation structure as occurs in market traded instruments. Consumer behavior will be at least as autocorrelated as the economy and management to which they are responding. Observation shows that consumers are even more autocorrelated than the related macroeconomic indicators, because consumer balance sheets carry a certain amount of cushion and therefore change more slowly than related measures such as GDP. For a discussion of the relationship between consumer behavior and macroeconomic factors, see [7].

To generate scenarios consistent with the observed historical structures, we will consider here the ARMA and ARIMA classes of models, although others are certainly possible. We employ a set of test criteria to verify that the model chosen was sufficient to generate scenarios statistically equivalent to the observed history.

3. Monte Carlo Scenario Generator. To address the problem described above, we created a univariate Monte Carlo scenario generator. We begin by fitting the model parameters over the historic data, preferably using an AICC statistic [8] to prevent the inclusion of too many lags and thus prevent overfitting. The in-sample model residuals, $\epsilon(t)$ are collected into an error distribution to be used in scenario generation.

Scenarios are generated month-by-month from

$$(3.1) \quad \hat{f}_g(t+1) = F(f_g(t); p) + \epsilon(t+1)$$

where $F(f_g(t); p)$ is the ARMA or ARIMA function applied to the exogenous curve values up to the current month, p are the model parameters, and $\epsilon(t+1)$ is the stochastic term sampled from the distribution of model residuals. The forecast for $\hat{f}_g(t+1)$ is then included in the model input when forecasting time $t+2$, etc. This process iterates until the desired horizon is reached, thus generating a single scenario. Additional scenarios are created by rerunning the process with a new series of stochastic terms $\epsilon(t)$.

Properly capturing the stochastic terms is important. To make sure that we are allowing for extreme values that may not be present in the historic data, it is reasonable to create a functional fit to the residuals distribution and sample from that function rather than directly from the tabulated residuals. One obvious choice is a normal distribution, and we usually find that this is sufficient to explain the distribution. For generality, we also compute skew and kurtosis for the distribution and use a Normalized Inverse Gaussian distribution [2, 19] when we find skew and kurtosis to be statistically significant. Other distributions are, of course, possible, but there is rarely any empirical need for exotic distributions to explain the observed residuals. Still, given the limited data, some practitioners may wish to consider other hypotheses.

Another important consideration is the choice of form for the function $F(f_g(t); p)$. In our case, the question is whether an ARMA or ARIMA class model should be used. This is a non-trivial question, which is handled at length in the next section.

4. Model Selection and Validation. Key to any modeling exercise is validation. A suitable set of validation tests guides us in model creation. In the case of modeling the volatility of retail loan portfolios, we do not have sufficient data to conduct in-sample / out-of-sample tests of the generated variance. Most training datasets are five to seven years long, capturing a single recession at most. All this data is needed for model calibration. A one-year out-of-sample test of the Monte Carlo generated distribution would require one-year data segments to create a distribution. Just to validate the standard deviation of the distribution would require dozens of years of data. A simple one-year out-of-sample snapshot amounts to a single realization and, although a reasonable test of a forecasting model, it is insufficient to validate a distribution.

The same situation prevails for all volatility models of retail lending. The Basel II guidelines pursue validation via multiple models. [3] By creating a separate stress test model, they seek to show that the results of the stress test model simulating various recessions will fall in the appropriate region of the volatility distribution. The second proposed validation method is to create a data-driven model of economic capital to compare against the results of the regulatory capital model. These methods are certainly appropriate for this Monte Carlo model of volatility. In fact, a major use of Monte Carlo simulation model described here is as the second model against which regulatory capital is validated.

Beyond just comparing against other models, our only options are component validation and verifying that the scenarios generated are statistically similar to the historic data.[21] To this end, we can employ separate tests of the forecast model and the MCSG. The following list of tests cover the range from testing the modeling algorithm to testing the specific application to a given data set. As such, Tests 1, 2,

3, 6, and 7 are run just to verify the algorithms. The others should be conducted to verify the specific results.

Forecast Model Validation Tests.

1. Extracting structure from known data sets
2. Alternating vintage
3. Bootstrap validation of the estimation process
4. Old vintage / new vintage
5. Ideal Scenario Validation

Monte Carlo Validation Tests.

6. Extracting structure from known data sets
7. Bootstrap validation of the estimation process
8. No significant auto-correlation in the in-sample residuals
9. Forecast scenarios match the historic monthly changes
10. Forecast scenarios match the historic autocorrelation behavior
11. Variance growth rate of scenarios matches the historic data

One test not listed is the Dickey-Fuller test [12]. We mentioned earlier that we have a choice in our MCSG of whether to use an ARMA or ARIMA model. The Dickey-Fuller test or other related tests are often employed in order to determine if a time series should be modeled directly or modeled against the first differences. These unit root tests are specifically designed around choosing the best approach for creating forecast models. Since the application here is not forecasting, but rather scenario generation for purposes of creating the distribution of possible futures, we need to consider whether other tests are more suited to the task. Rather, we suggest choosing between ARMA and ARIMA type models on the basis of which generates scenarios with the best match to the statistical properties of the historic series.

4.1. Test 1: Extracting Structure from Known Data Sets. When data sets are created artificially to possess maturation, exogenous, and vintage quality dimensions, we expect DtD to be able to capture this structure. This is an obvious test for any modeling procedure, and an example of this test was published by Breeden, 2007 [6]. Any model implementation should recreate this test.

4.2. Test 2: Alternating Vintages. DtD assumes that all the vintages within a given segment share the same maturation and exogenous curve. Therefore, by splitting the segment in half and analyzing each subsegment independently, we should observe the same shape maturation and exogenous curves to within the estimation. The alternating vintage approach ensures that the same time period is analyzed in both subsegments, but that no data is shared between them. This test confirms that the vintages are coming from an underlying process that can be reliably observed with DtD.

A visual inspection of the curves including the uncertainties is often sufficient to identify clear structural problems. However, statistical tests are also available to confirm this. We can test whether the two subsegments are creating models of equivalent accuracy by employing the Granger-Newbold or Diebold-Mariano tests.

Granger and Newbold [16] developed an approach for determining if two models are significantly different that applies for series which may have cross-correlated errors. To apply their test, let $f_m^o(a)$ be the curve obtained from the "odd" vintages, $f_m^e(a)$ is the curve from the "even" vintages, and $f_m(a)$ is the curve obtained from the full data set. In a well-formed problem, $f_m^o(a) = f_m^e(a) = f_m(a)$, so we want to determine if the deviations of the odd and even curves from the full curve are significant.

To apply their approach, let

$$(4.1) \quad \epsilon_m^o(a) = f_m^o(a) - f_m(a)$$

$$(4.2) \quad \epsilon_m^e(a) = f_m^e(a) - f_m(a)$$

and

$$(4.3) \quad x_a = \epsilon_m^o(a) + \epsilon_m^e(a)$$

$$(4.4) \quad z_a = \epsilon_m^o(a) - \epsilon_m^e(a)$$

Then they compute

$$(4.5) \quad GN = \frac{r_{xz}}{\sqrt{(1 - r_{xz})(N - 1)}}$$

where r_{xz} is the correlation between the two series $x(a)$ and $z(a)$. The GN statistic has a t-distribution with degrees of freedom $N - 1$ where N is the length of the series. If this measure is statistically different from zero, then the models are not equivalent in their accuracy and the stability test fails.

The GN test statistic uses a quadratic function of the errors to compare accuracy. This test has been generalized to any accuracy measure by Diebold and Mariano, a common choice being absolute error[9]. Given an objective function $g(\epsilon)$ the mean modeling error is

$$(4.6) \quad \bar{d} = \frac{1}{N} \sum_{i=1}^N |g(\epsilon_m^o(a)) - g(\epsilon_m^e(a))|$$

For models of equivalent accuracy, $\bar{d} = 0$. To test whether \bar{d} is significantly different from zero, we compute

$$(4.7) \quad DM = \frac{\bar{d}}{\sqrt{(\gamma_0 + 2\gamma_1 + \dots + 2\gamma_q)(N + 1)}}$$

where the correction for forecast horizon drops out since we are comparing explanatory models.

If we let γ_i equal the i -th auto-covariance of the sequence,

$$(4.8) \quad d_i = g(\epsilon_m^o(a)) - g(\epsilon_m^e(a)),$$

then the DM statistic is a t-distribution with $N - 1$ degrees of freedom. If the result is significantly different from zero, then the models are not equivalent.

Either the GN or DM tests can be used to compare the maturation or exogenous curves to determine if they are statistically equivalent.

4.3. Test 3: Bootstrap Validation. DtD as described in [6] uses a hybrid nonlinear / non-parametric, iterative estimation process. This complexity means that the properties of the estimator cannot be solved in closed form. In such cases, a bootstrap estimation process is standard [13]. The goal is to show that the analysis of the full data set falls in the center of the results obtained from the bootstrapped data sets. The Alternating Vintage test is a simple case of this, but the full bootstrap

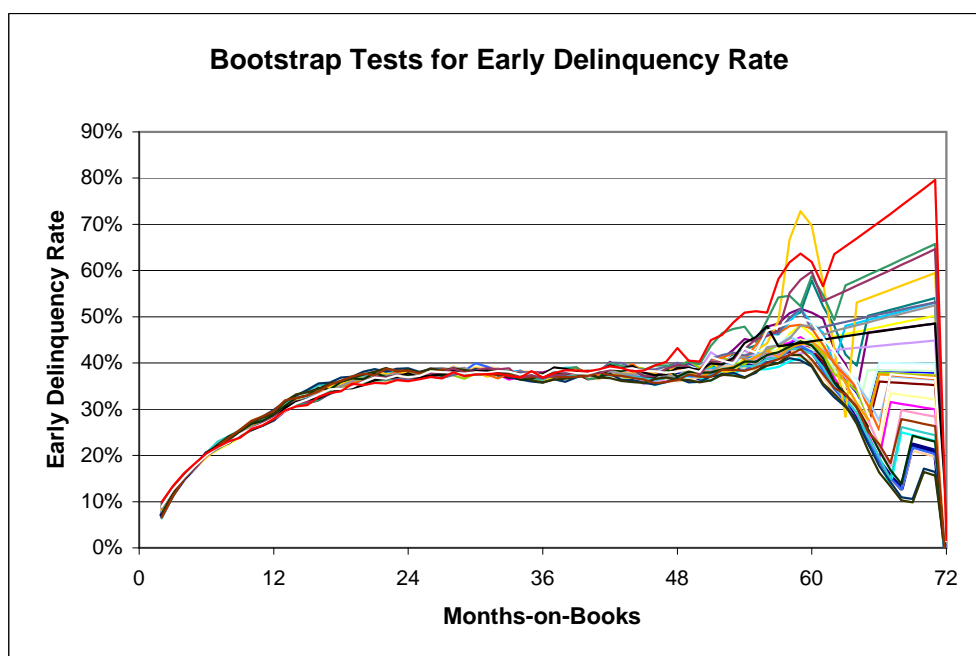


FIG. 4.1. *Independent analyses of 100 bootstrap runs with a 50% sampling rate.*

test is more powerful at determining that the estimation process provides unbiased results.

A bootstrapped data set is created in this context by randomly sampling with replacement from the original data set. Each bootstrap data set is therefore a sub-sample of the original data set. As with the alternating vintage test, selecting a subset should not alter the maturation or exogenous curves in shape, although the scaling of the curves can change.

4.4. Test 4: Old Vintage / New Vintage Comparisons. The Alternating Vintage test (Test 2) and Bootstrap test (Test 3) compare a random sampling of vintages to verify estimation stability. These tests cannot check for segment drift over time. One of the most common reasons for poor performance of DtD is that the segmentation is not stable. If a bank significantly shifts their distribution of new originations, but leaves all the accounts in a single segment, the true maturation curve for the old vintages might be different from that for the new vintages.

Examples include shifting from prime to subprime, going from 5-year term loans to 7-year terms, or introducing a new no-payments-for-12-months program. In each of these cases, the possibility of a significant shift in originations mix means that the segment should be split along the dimension in question.

The test is trivial to conduct. The data is split into two parts separating the old vintages from the new vintages. Each group is analyzed independently. The overall scaling of the maturation curves can shift during the time periods, but with a stable segmentation, each subsegment should show the same shape maturation curve.

Since the maturation curves were estimated from temporally distinct data segments, they may have different calibration levels. We are trying to determine if the nonlinear structure of the curves is equivalent, but simple linear scalings between the

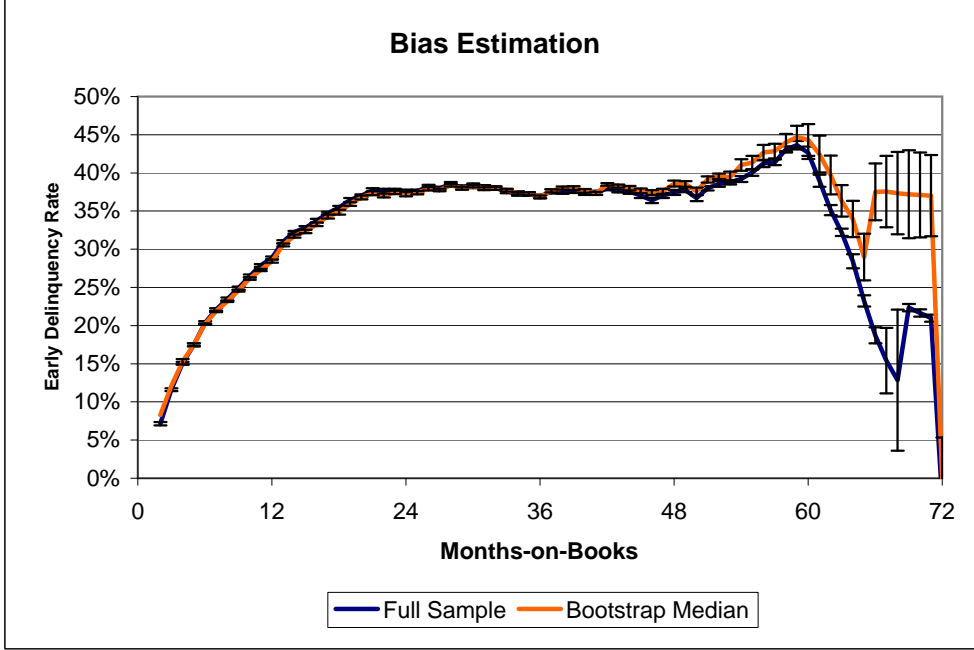


FIG. 4.2. Comparisons of the full data set analysis to the median of the bootstrap samples. Divergences in the tail of the curve occurs because the bootstrap samples run out of data in that region. When the data thins out, the curves are extrapolated, leading to estimation errors.

curves are unimportant. If we first normalize the curves to an equivalent range, we can apply the Granger-Newbold or Diebold-Mariano tests described previously.

We can also take a somewhat different perspective by computing a distance metric between the two curves. Measuring the distance between curves can be very helpful when trying to find the optimal segmentation for a portfolio. At each point along the maturation curve, we estimate both the value $f_m(a)$ and the uncertainty in that value $\sigma_m(a)$. We can compute error-normalized distance between the old vintage maturation curve $f_m^o(a) \pm \sigma_m^o(a)$ and the new vintage maturation curve $f_m^n(a) \pm \sigma_m^n(a)$, as

$$(4.9) \quad d_{on} = \frac{f_m^o(a) - f_m^n(a)}{\sqrt{\sigma_m^o(a)^2 + \sigma_m^n(a)^2}}$$

This is also the formula for the Unequal Variance Students t-Test, and in effect we are testing whether these points are statistically different.

The component distances, d_{on} , can be combined into a single distance between the two curves using one of the standard metrics, such as the L_1 , L_2 , or L_∞ -norm. The L_1 -norm will likely be most robust to noise, which is often important in our application.

As before, the same set of tests can be applied to the exogenous curves to verify that those components are also statistically equivalent.

4.5. Test 5: Ideal Scenario Validation. DtD was developed as the basis for scenario-based forecasting. When creating a forecast, we take the maturation and vintage quality as estimated from the historic data. The exogenous curve, however, is

a measure of the macroeconomic environment. As such, we expect this to differ over the out-of-sample period. To create a forecast, the analyst would supply a scenario for the future exogenous curve, as well as scenarios for the volume and quality of future vintages.

With a simple time series model, the standard measure of accuracy is to conduct an out-of-sample test. Train the model on one time interval and test the forecast over another time interval. With any scenario-based forecasting model, the situation is more complex. We want to be able to distinguish between model error associated with estimating maturation and quality in-sample, and scenario error related to the exogenous scenario used out-of-sample. Any model that uses macroeconomic data as an input would have a similar need to separate sources of error.

The out-of-sample testing method employed with DtD is the Ideal Scenario Validation (ISV). For an ISV, an in-sample data set is used to measure all the model components. A second out-of-sample data set is analyzed to obtain the exogenous curve during that period. Vintage-level forecasts are then run using the known maturation and quality functions and the out-of-sample exogenous curve as the ideal scenario. DtDs assumption that a segment can be decomposed into maturation, exogenous, and vintage quality is dependent upon the data being provided. A poor segmentation can be unstable with respect to the maturation or exogenous curves, resulting in poor out-of-sample validation. The ISV measures the accuracy of the decomposition and forecasting given a known scenario for the environment. In essence, it measures model accuracy separately from scenario accuracy. This validation approach is appropriate for any scenario-based forecasting method.

Of course, the analyst will not be able to recreate the ideal scenario during an actual forecast. A good analyst will save past scenarios originally created by studying macroeconomic trends and business policy changes and compare them to what actually happened in order to assess scenario creation accuracy.

ISVs are typically measured on a 12-month out-of-sample period for each segment. This results in a single error sample for each segment. Routine data anomalies can cause unusually high or low values for the ISV result. Therefore, one should consider collecting ISV results across all available segments and comparing. Model error has been found experimentally to scale with

- length of training data
- number of training vintages
- event rate
- forecast horizon

In cases where the ISV result is poor, Test 4 (old vintage / new vintage) is an appropriate diagnostic tool. If the ISV is good, one may assume that the segmentation is acceptable. In some cases, businesses will accept a somewhat worse ISV result even when Test 4 shows some weaknesses in order to maintain alignment with internal business processes or to avoid increasing the modeling error by over-segmenting already thin segments.

4.6. Test 6: Extracting Structure from Known Data Sets. Just as was done in Test 1, the MCSG should always be tested to verify that it can learn the structure from known test series prior to application to unknown series. In many cases this will be a trivial test, but with the presence of autocorrelation structure in the exogenous curves, the MCSG will necessarily be more complex and tests such as this become more valuable. We routinely perform such tests. Passing simply requires that the ARMA and ARIMA models be properly constructed.

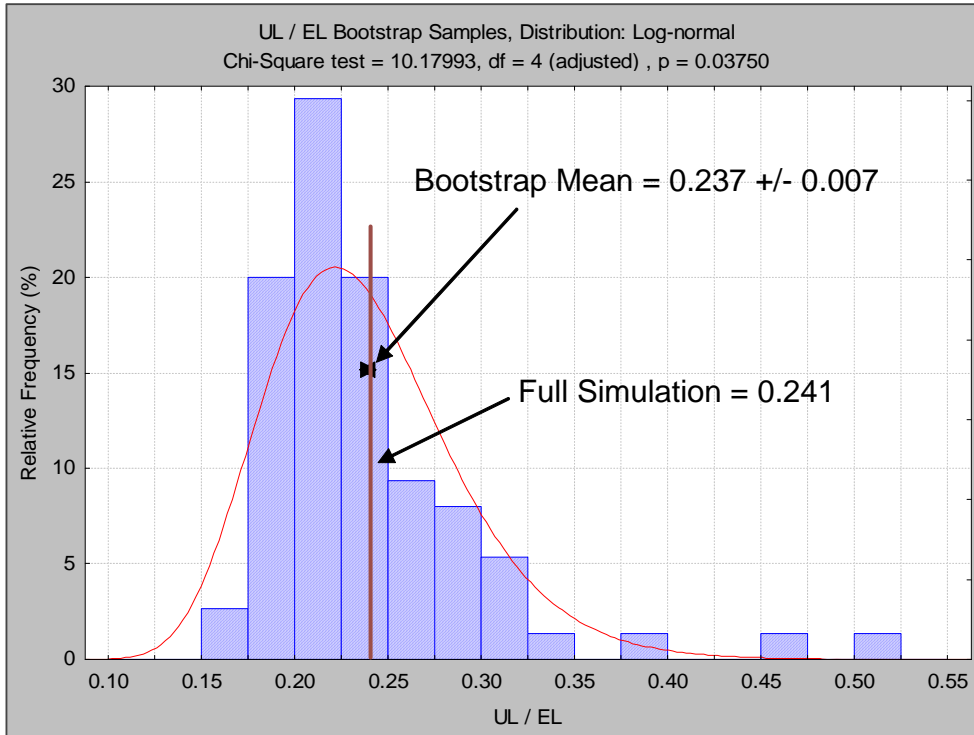


FIG. 4.3. *Bootstrap tests for UL / EL. Sub-segments represent a 50% sampling of the original data set. The bootstrap mean and full simulation result agree to within estimation error. The distribution represents 75 runs of 5,000 scenarios each. The red line is a LogNormal fit to the distribution.*

4.7. Test 7: Bootstrap Validation. The Monte Carlo Bootstrap test follows directly from the DtD Bootstrap test. Each bootstrapped subsegment is fed into DtD to create a baseline scenario as discussed earlier. Those files are then fed into the MCSG to measure the spread of volatility that results and verify that the result for the full data set is in the center of the distribution created from the bootstrapped subsegments. Because the sub-segments have less data than the full segment, the uncertainty will be higher, i.e. the distribution of measured volatility can be quite wide, but passing the test simply requires that the full result be centered.

Passing this test verifies that the MCSG is producing unbiased results relative to the original baseline scenario. The following is one such example.

4.8. Test 8: Monte Carlo Model Residuals Show No Auto-Correlation Structure. The MCSG is configured specifically to capture the statistically significant auto-correlation structure in the external series being simulated. Assuming that the model was properly created, there should be no significant auto-correlation present in the model residuals. The Durbin-Watson statistic [17] can be used to test for a correlation with a one-month lag. More generically, computing the full Auto-correlation function will help verify that no significant model errors occur for any lag.

4.9. Test 9: Monthly Differences of the Generated Scenarios Match the Historic Data. After the Monte Carlo scenarios have been generated, they can be compared to the historic data to make sure the the model is accurately reproducing the

observed structures out-of-sample. The most basic test is the distribution of monthly changes in the scenarios as compared to that observed historically. These distributions need not be Normal, but they should agree to within estimation error. A Kolmogorov-Smirnov (K-S) statistic [22] is one such suitable test of agreement between the history and the scenarios. As mentioned earlier, we do not have sufficient data to validate the distribution over a one-year forecast, but this test is equivalent to validating the distribution for a one-month horizon.

4.10. Test 10: Autocorrelation Function of the Generated Scenarios Match the Historic Data. The scenarios are generated with the intent of capturing the significant in-sample autocorrelation structure. Comparing the aggregate ACF for the generated scenarios to the in-sample ACF provides a test of the effectiveness of the scenario generator. In doing this comparison, we must pay attention to the model used for the scenario generator. For an ARMA model, the ACF test is performed on the original series. When an ARIMA model is employed, the ACF must be computed on the differenced series.

4.11. Test 11: Time Scaling of Variance. To choose an appropriate class of models for scenario generation, we considered several properties of the historic data. Although we want to create scenarios that match as many properties of the historic time series as possible, the ultimate goal is to create a distribution of possible futures consistent with the historic distribution. Being able to reproduce the historic time scaling of variance via the generated scenarios is essential to ensuring that measurements such as economic capital will be correct.

Choosing the correct model for the Monte Carlo scenario generator does not require that the scenario generator be run. By measuring the rate of growth of the variance in the historic external series, the appropriate model can be selected. If the variance grows faster than the square-root-of-time, then ARIMA is the required. If the variance grows as the square root of time, then a random walk model is prescribed. If the variance grows slower than the square-root-of-time, then an ARMA model is appropriate.

To investigate this further, we studied the time scaling of variance for a range of ARMA and ARIMA models theoretically and in numerical experiments, and compared those to available retail loan data and macroeconomic data. To begin with, we derive the time scaling of variance for some ARMA and ARIMA models. The detailed calculations of variance growth rates for these models are included in the Appendix.

MA(1). If $T = 1$, then

$$(4.10) \quad E\left((x_{t+T} - x_t)^2\right) = \sigma_T^2 = 2\sigma^2(\phi^2 - \phi + 1)$$

where σ_T^2 is the variance across multiple periods, and σ^2 is the one-period variance for a series where the monthly differences are IID. ϕ is the MA(1) parameter.

Similarly, for $T > 1$, we find

$$(4.11) \quad \sigma_T^2 = 2\sigma^2(\phi^2 + 1)$$

Except for $T = 1$, σ_T is constant as a function T .

AR(1).

$$(4.12) \quad \sigma_T^2 = 2\sigma^2 \frac{1 - \alpha^T}{1 - \alpha^2}$$

where α is the AR(1) parameter.

Random Walk.

$$(4.13) \quad \sigma_T^2 = T\sigma^2$$

ARIMA(0,1) or MAI(1).

$$(4.14) \quad \sigma_T^2 = \sigma^2 (T\phi^2 + 2(T-1)\phi + T)$$

ARIMA(1,0) or ARI(1).

$$(4.15) \quad \sigma_T^2 = \sigma^2 \left(T + \frac{T\alpha^2}{1-\alpha^2} + 2 \left(\frac{\alpha^2}{1-\alpha^2} + \alpha \right) \sum_{i=1}^{T-1} \alpha^i (T-i) \right)$$

Past experience with retail lending portfolios has shown that the variances involved for default rates typically do grow faster than the square-root-of-time, so the expectation is that ARIMA models will be required. However, this is not a requirement as the other types may be selected when appropriate. The ideal situation would be to test the available history for each variable and segment directly. Unfortunately, the typical bank time series is too short to support these time-scaling-of-variance tests.

One alternative is to leverage the econometric stress testing models created for the exogenous curves [7]. These models are typically used to extrapolate the exogenous curve given a macroeconomic scenario. However, they can also be run backward, as shown in Figure 4.5, to create an estimate of what the impact of past macroeconomic environments would have been on a portfolio such as exists today. We can then study the properties of this index as a proxy for the behavior of exogenous curves through those economic cycles.

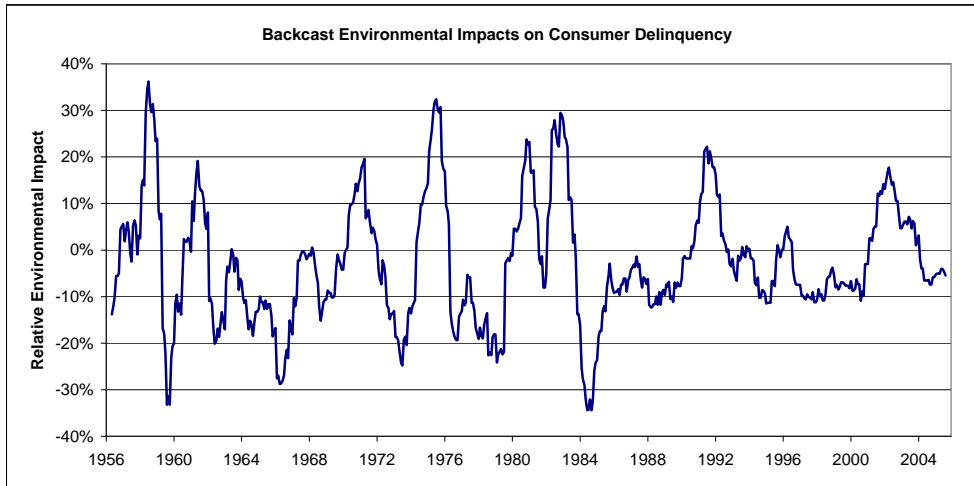


FIG. 4.4. *Exogenous curve capturing environmental Impacts for a US Default Rate was calibrated to macroeconomic factors (Unemployment, GDP, and Interest Rates) and backcast to 1955 to show what environmental impacts on a modern consumer portfolio would have looked like in previous macroeconomic environments.*

The following figure shows the results of one such test. This example is for a US Default Rate index backcast to 1955, Figure 4.4. The standard deviation growth is greater than \sqrt{T} , which implies a greater-than-random-walk growth rate and thus requiring an ARIMA class model.

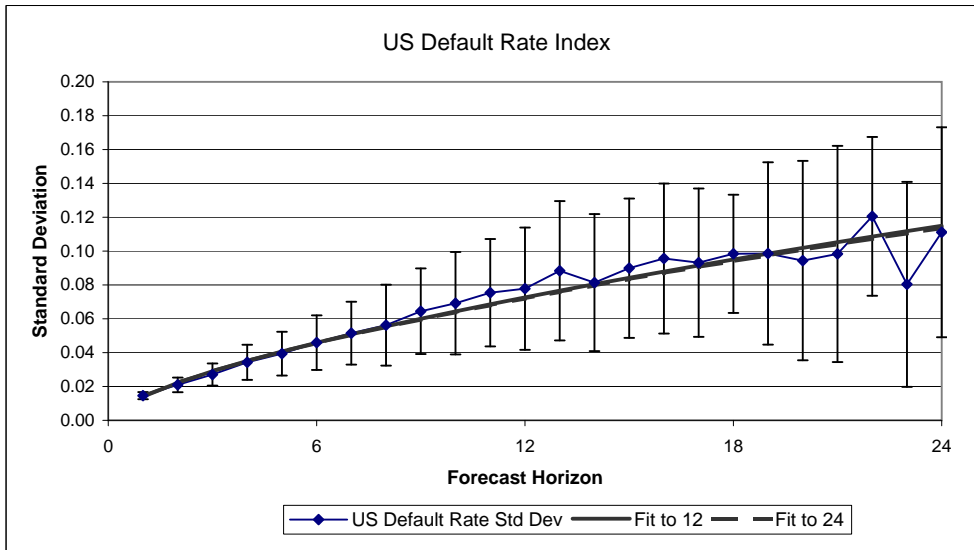


FIG. 4.5. Example of a Variance Growth Test. The weighted power law fit to the observations results in a relationship indicative of ARIMA models.

5. Example of GDP by Country. Since we often find the GDP growth rate series to correlate well to the default rate exogenous curve, we tested series from multiple countries to see which Monte Carlo model would be selected. Table 5.1 lists the countries tested. For each, we computed the year-over-year percentage growth in GDP. Some of these series are plotted in figure 5.1.

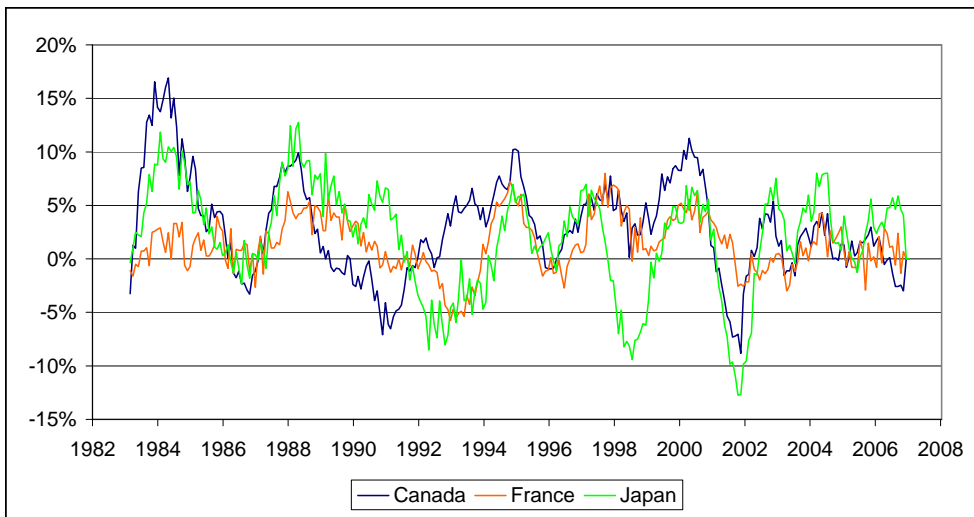


FIG. 5.1. Time series of GDP year-over-year growth rates for some countries analyzed.

The series plotted clearly do not have any significant long term trend. Therefore, when the Dickey-Fuller test is applied, it would fail to require differencing for the modeling. Most practitioners would then default to an AR model of the original

series rather than differencing as is done with ARIMA models.

Conversely, our task is to choose a model that creates the best distribution over the next year (for economic capital) or few years (for lifetime capital needs). Using these series, we computed $E\left((x_{t+T} - x_t)^2\right)$ for $T \in [1, 24]$ to produce a data for each series analogous to what is shown for the US in figure 4.5. The time scaling of variance for the different models as shown in section 4.11 is not exactly a power law for any but the random walk. However, fitting a power law to the data provides an approximate metric for selecting between the models. Table 5.1 lists the exponent when fitting $\sigma_T = aT^b$. δb is the estimation error in b at 95% confidence.

Country	Data Range	Power (b)	Uncertainty (δb)	DF Prob.
Australia	Jan-1980 to Dec-2006	0.588	0.06	0.0054
Canada	Mar-1982 to Dec-2006	0.715	0.06	0.0197
France	Mar-1982 to Dec-2006	0.774	0.03	0.0394
Germany	Mar-1982 to Dec-2006	0.54	0.038	0.0083
India	Apr-1994 to Dec-2006	0.435	0.11	0.0694
Japan	Mar-1982 to Dec-2006	0.688	0.078	0.0286
Korea	Jan-1989 to Dec-2006	0.57	0.054	0.004
Mexico	Mar-1982 to Dec-2006	0.548	0.053	0.0052
South Africa	Jan-1998 to Dec-2006	0.168	0.59	0.0064
United States	Mar-1982 to Dec-2006	0.685	0.034	0.0091

TABLE 5.1

A comparison of the power law growth rate of the variance to the Dickey-Fuller test result. These were applied to GDP growth rate series for various countries. In all but two cases, the variance growth rate suggests an ARIMA model is required, even though a Dickey-Fuller test would indicate that modeling the differenced series is not necessary for forecasting.

This table shows that in all cases except India and South Africa, the time scaling of variance is greater than square root of time. Indian and South Africa were two of the shortest data sets, and given the uncertainty intervals we cannot determine what model is required. For the other countries, to generate scenarios that will recreate the observed behavior, we must use ARIMA class models, since ARMA models have strictly less than square-root-of-time variance growth.

To reiterate a key point, the apparent disagreement between the Dickey-Fuller test and the time scaling of variance test occurs simply because they are asking different questions. In our situation, the time scaling of variance test would appear to be most directly relevant.

6. Best Practices.

6.1. Minimum Number of Scenarios. Since one of our primary goals in generating scenarios is to measure the distribution of possible outcomes, we need to explore how many scenarios are required to measure the distribution accurately. Obviously, the more scenarios generated, the more certainty we have when measuring a point along the cumulative probability distribution. To determine a suitable number of scenarios to achieve the desired accuracy, we constructed a test problem with a known LogNormal distribution. To create an example that has similar properties to what occurs in economic capital calculations for retail portfolios, we calibrated the test distribution to have a median of 1 and width such that the 99.9% point on the cumulative distribution has a distance of 1 from the median. In capital terms, this can be expressed as Expected Loss (EL) = 1 and Unexpected Loss (UL) / EL = 1.

We then began randomly sampling points from the known distribution in order to create a tabulated cumulative probability distribution. The point is to see how many samples are required in order to achieve a match between the original distribution and the tabulated distribution at various points along the distribution. This is equivalent to assuming a perfect model and testing the sampling error.

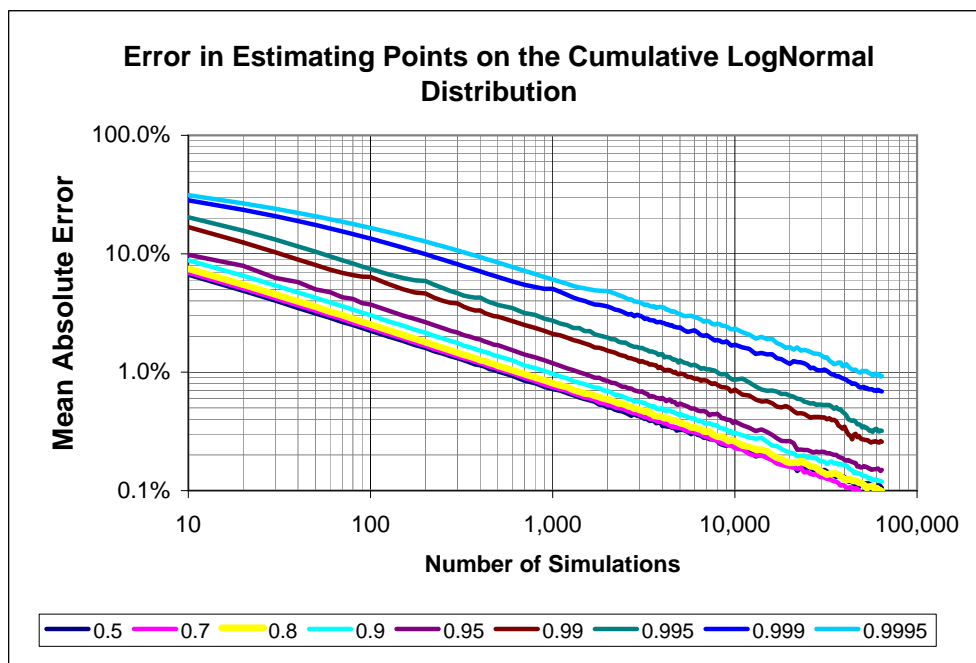


FIG. 6.1. Each line represents a point on the cumulative probability distribution, showing how error decreases as the number of scenarios increases.

Figure 6.1 shows the result of this analysis. Each line represents one point on the cumulative probability distribution. The 0.99 line shows the 99th percentile on the distribution, and the average measurement error obtained when drawing the number of random samples shown on the x-axis. For the 99% line, a 5% error is attained after only 150 scenarios, 2% error is achieved with 1,000 scenarios, and a 1% error is obtained with 5,000 scenarios. Since economic capital is usually measured at the 0.999 point on the cumulative probability distribution (a 99.9% solvency level), running 10,000 to 30,000 scenarios will provide an uncertainty of 1% to 2%. Given that the ISV errors are typically greater than 2%, 20,000 scenarios appears to be a reasonable compromise between accuracy and computational effort. Volatility measurements for business planning or portfolio optimization usually focus on the 1 in 10 year events, i.e. the 90% point on the distribution. For such studies, as few as 1,000 iterations will provide better than 1% error.

Figure 6.2 summarizes the results of Figure 6.1 by showing how many scenarios are required to achieve a 1% error in the result. In order to show this on a log-log scale, the x-axis is plotted as 1- Solvency, where the lines in Figure 6.1 represent different solvency levels.

6.2. Minimum Training Data. When insufficient training data is available, volatility measurements could underestimate the intrinsic long-run volatility of the

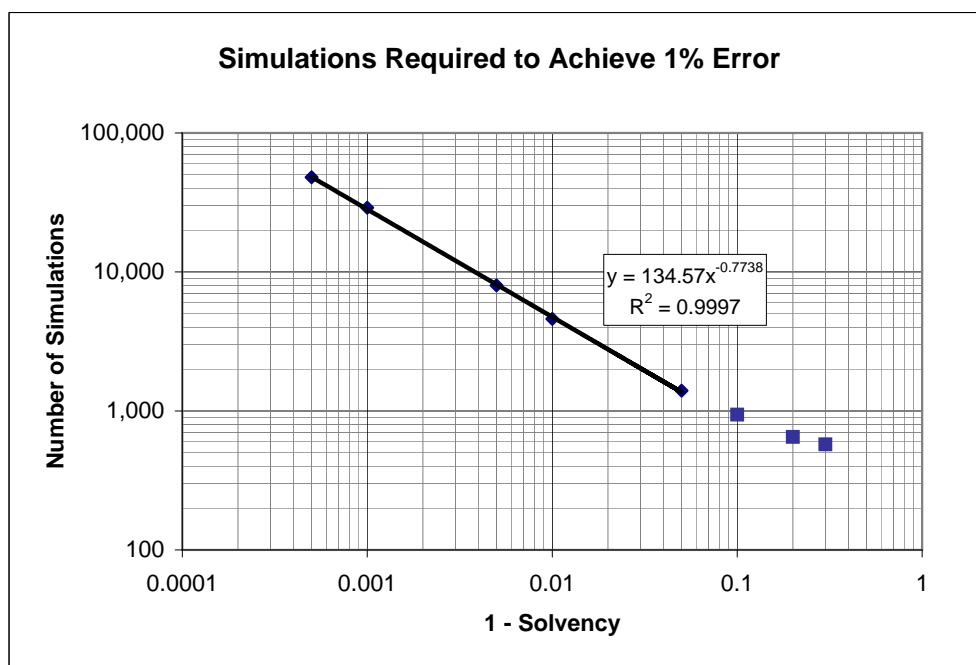


FIG. 6.2. Number of simulations required to achieve 1% error or less when measuring a given point in the tail of the cumulative probability distribution.

segment. To test this possibility, we used the same backcast series described in Test 11 that calibrated retail portfolios to macroeconomic indicators and ran those indices backward in time. By then taking many training subsets of different lengths from various environments, we created UL to EL ratios and compared them to the values obtained when all the data was available. The following figure shows that training on very short data sets (three years or less) can produce a result that is significantly biased toward underestimating the volatility. Volatility here was measured at the 99.9% level.

Aside from the bias, we also estimated the spread of the UL / EL estimates as a function of the length of the training data. The bias can be seen to be quite large again for the very short data sets. For a typical data set with seven years of training data, the bias is insignificant. Training over different macroeconomic periods is the largest contributor to uncertainty in the UL / EL ratio, so the following test is specifically constructed to investigate this term.

In contemplating this result, one should consider that five years of data represents roughly one macroeconomic cycle. Another reasonable interpretation of these results could be that at least one full macroeconomic cycle is required to obtain a stable estimate of volatility.

6.3. Stability Over Different Macroeconomic Eras. Since most retail lenders only have data representing the last recession, one must ask whether training on a single recession is representative of the volatility obtained if we had modeled past recessions instead. Since US institutions do not have sufficient data to extract exogenous curves for previous recessions directly, we once again use the approach of modeling the exogenous curve with macroeconomic data, and then extrapolating backward what

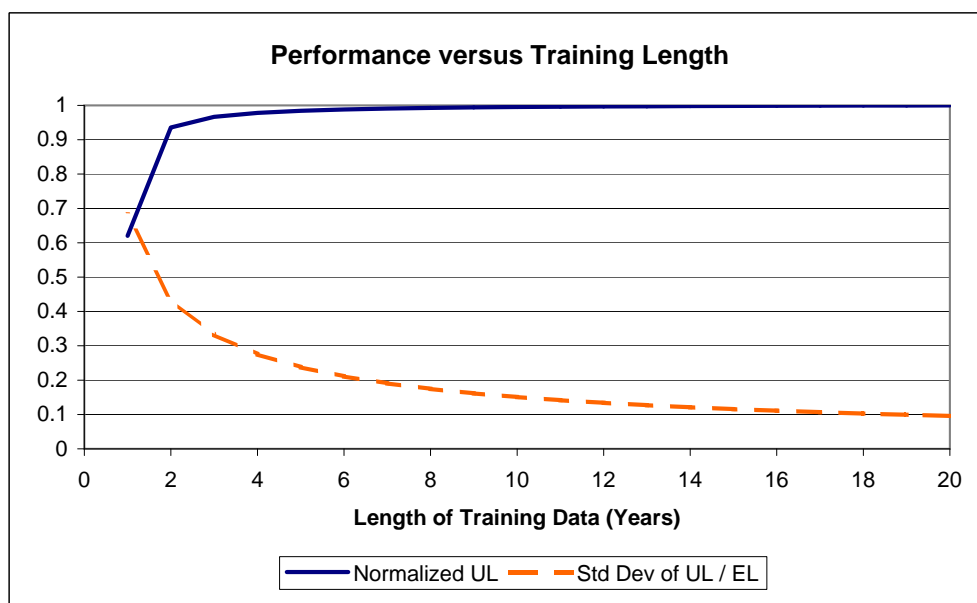


FIG. 6.3. The blue line shows the divergence of UL / EL from the long run value as the training data gets shorter. The orange line shows how the uncertainty in the UL / EL estimate increases for short data sets.

the macroeconomic impacts on retail lending would likely have been, Figure 4.4. Using the resulting index, we trained the MCSG on ten-year windows of data, predicted the volatility for the following two years, and then moved forward by two years to repeat. When this test was performed, the most striking result was the change that occurred around 1984. When data prior to 1984 was trained on, the results were very consistent across recessions. When data entirely after 1984 was employed, another internally consistent result was obtained, but significantly lower than the previous volatility.

Figure 6.4 summarizes this information by comparing the average separation between 90% and 50% for distributions derived from the pre-1984 and post-1984 epochs. Within each epoch, the results are quite consistent, but clearly the recent period has shown significantly less volatility.

The explanation for this behavior is typically expressed as the Greenspan era, and has been noted in several previous studies in other contexts [23, 1]. The essential question for financial institutions is whether volatility, and thus economic capital, should be computed solely on the last 25 years of macroeconomic cycles, or should older periods be included. This is a governance question, as the MCSG can be run as needed. Practitioners should note, however, that if only internal experience is to be used, then they are implicitly selecting just the modern era. To train on older recessions, macroeconomic data must be incorporated as well.

7. Incorporating Cross-Correlations. All of the preceding discussion involved generating scenarios for the exogenous curve time series, $f_g(t)$. Implicit in this discussion was that there were no significant correlations between time series. Of course, the real world often has such correlations, so we must consider correlations.

Two distinct situations arise. The first is that different segments can have highly

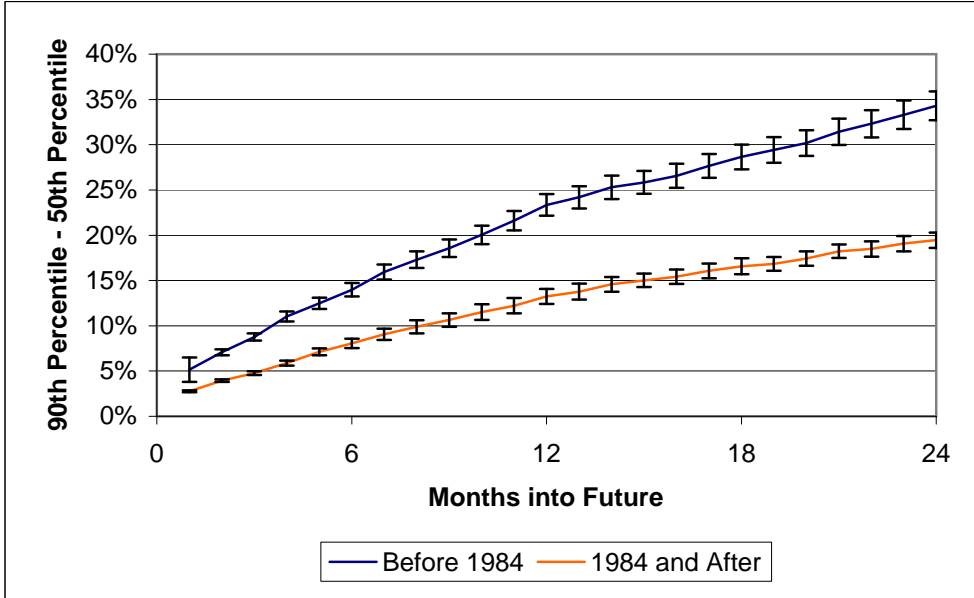


FIG. 6.4. The mean separation between the 90th and 50th percentiles for two year periods before 1984 and since 1984. Errors bars represent the amount of variability observed across different training periods.

correlated exogenous curves. This is equivalent to having the same impact from the economy across multiple similar product segments. Segment correlations are observed frequently in retail loan portfolio data.

The second situation is where variables within a single segment are correlated. A concrete example of this is to consider whether DR (or PD), LGD, or EAD are correlated. This is an important question with Basel II, because the regulatory model considers volatility only in PD. Because situations may arise where LGD stresses are correlated to PD stresses, the regulatory guidelines are adopting an approach of using extreme values of LGD when computing capital. EAD correlations are not currently included in Basel II. For our purposes, we can generically incorporate any correlations between these variables.

For either of these situations, we can express the problem as one of having a set of exogenous curves $f_g^i(t)$ that are correlated. Using a principal components approach, or even a simple average, we can compute the correlated component $f_g^C(t)$. Each of the exogenous curves can be modeled with that component as

$$(7.1) \quad \hat{f}_g^i(t) = a^i f_g^C(t) + b^i + \eta^i(t)$$

where a^i and b^i are curve-specific scaling factors and $\eta^i(t)$ is the curve-specific residual time series.

When performing Monte Carlo simulations, the target variables then become the combined factor, $f_g^C(t)$, and the idiosyncratic factors, $\eta^i(t)$. More complex forms are of course possible. When looking across a wider range of segments, we may find multiple independent factors from the PCA that can be included in the modeling, but we will always want to make sure that we include the idiosyncratic factors so that no modeling errors are lost when computing volatility.

8. Conclusion. In this paper, we have described how to leverage Monte Carlo scenario generation and the Dual-time Dynamics modeling technique to predict the future volatility of retail loan portfolios. The technique is straight-forward, but validation of the models is challenging because banks do not possess sufficient data to conduct an out-of-sample test on a complete distribution of possible futures. Therefore, we have given significant space to discussing how one might validate the scenario generation. The approach employed is common in Monte Carlo models where the generated scenarios are compared statistically to the historic data to make sure that all significant structure has been captured in the model and is being replicated in the scenarios. Given the importance of validation in the context of economic capital models, we should continue to look for additional metrics by which to validate the models.

Appendix. The section contains detailed calculations of the time scaling of variance relationships referenced earlier.

MA(1). For an MA(1) model of the form

$$(8.1) \quad x_t = \phi\epsilon_{t-1} + \epsilon_t$$

where ϵ_t is normally distributed with variance σ^2 , we have

$$(8.2) \quad x_{t+T} - x_t = \phi\epsilon_{t+T-1} + \epsilon_{t+T} - \phi\epsilon_{t-1} - \epsilon_t$$

If $T = 1$, then

$$(8.3) \quad (x_{t+T} - x_t)^2 = \phi^2\epsilon_t^2 + 2\phi\epsilon_t\epsilon_{t-1} - 2\phi^2\epsilon_t\epsilon_{t-1} - 2\phi\epsilon_t^2 + \epsilon_{t+1}^2 - 2\phi\epsilon_{t+1}\epsilon_{t-1} - 2\epsilon_{t+1}\epsilon_t + \phi^2\epsilon_{t-1}^2 + 2\phi\epsilon_{t-1}\epsilon_t + \epsilon^2$$

from which we find

$$(8.4) \quad E\left((x_{t+T} - x_t)^2\right) = \sigma_1^2 = 2\sigma^2(\phi^2 - \phi + 1)$$

Similarly, for $T > 1$, we find

$$(8.5) \quad \sigma_T^2 = 2\sigma^2(\phi^2 + 1)$$

Except for $\sigma_1 < \sigma_T$, σ_T is constant as a function T .

AR(1). For an AR(1) model,

$$(8.6) \quad x_t = \alpha x_{t-1} + \epsilon_t$$

we perform similar calculations such that

$$(8.7) \quad \sigma_T^2 = 2\alpha^2 Cov(x_t, x_t) - 2\alpha^2 Cov(x_t, x_{t+T}) - 2\alpha Cov(x_{t+T-1}, \epsilon_t)$$

Using the formulas for the covariances,

$$(8.8) \quad Cov(x_t, x_t) = \frac{\sigma^2}{1 - \alpha^2}$$

$$(8.9) \quad Cov(x_t, x_{t+T}) = \frac{\alpha^T \sigma^2}{1 - \alpha^2}$$

$$(8.10) \quad Cov(x_t, \epsilon_t) = \alpha^{T-1} \sigma^2$$

we can simplify this expression to

$$(8.11) \quad \sigma_T^2 = 2\sigma^2 \frac{1 - \alpha^T}{1 - \alpha^2}$$

This expression shows that volatility grows for with T , but flattens rapidly as T increases.

Random Walk. For a random walk process, the well known result is that

$$(8.12) \quad \sigma_T^2 = T\sigma^2$$

which is commonly expressed as the deviation of a random walk grows as the square root of time. This result is commonly used in analyzing stock market returns to estimate aggregate volatility as the number of samples per interval times the sub-sampled volatility.

ARIMA(0,1) or MAI(1). For an order 1 ARIMA process, we consider two cases. An ARIMA(0,1), which could be called an MAI(1) model, is expressed as

$$(8.13) \quad x_t = \phi\epsilon_{t-1} + \epsilon_t$$

$$(8.14) \quad y_t = y_{t-1} + x_t$$

which can also be written as

$$(8.15) \quad y_t = y_0 + \sum_{j=1}^t (\phi\epsilon_{j-1} + \epsilon_j)$$

Therefore,

$$(8.16) \quad y_{t+T} - y_t = \sum_{j=1}^T (\phi\epsilon_{t+j-1} + \epsilon_{t+j})$$

and

$$(8.17) \quad (y_{t+T} - y_t)^2 = \phi^2 \sum_{j=1}^T \epsilon_{t+j-1}^2 + \sum_{j=1}^T \epsilon_{t+j}^2 + 2\phi \sum_{j=1}^{T-1} \epsilon_{t+j}^2 + \text{cross-terms in } \phi\epsilon_{t+j}\epsilon_{t+j+1}$$

This leads to an expression for the variance of

$$(8.18) \quad \sigma_T^2 = \sigma^2 (T\phi^2 + 2(T-1)\phi + T)$$

ARIMA(1,0) or ARI(1). An ARIMA(1,0), which could be called an ARI(1) model, is expressed as

$$(8.19) \quad x_t = \alpha x_{t-1} + \epsilon_t$$

$$(8.20) \quad y_t = y_{t-1} + x_t$$

which can be rewritten as

$$(8.21) \quad y_t = y_0 + \sum_{j=1}^t (\alpha x_{j-1} + \epsilon_j)$$

From this we get the expression

$$(8.22) \quad y_{t+T} - y_t = \sum_{j=1}^T (\alpha x_{t+j-1} + \epsilon_{t+j})$$

The variance can then be expressed as

$$(8.23) \quad \sigma_T^2 = T\sigma^2 + \alpha^2 \sum_{i=1}^T \sum_{j=1}^T Cov(x_{t+i}, x_{t+j}) + \alpha \sum_{i=1}^{T-1} \sum_{j=1}^i Cov(x_{t+i}, x_{t+j})$$

which can also be written as

$$(8.24) \quad \begin{aligned} \sigma_T^2 = & T\sigma^2 + T\alpha^2 Cov(x_t, x_t) \\ & + 2 \left(\alpha^2 \sum_{i=1}^{T-1} Cov(x_t, x_{t+i}) (T-i) + \alpha \sum_{i=1}^{T-1} Cov(x_{t+i}, \epsilon_t) (T-i) \right) \end{aligned}$$

Since

$$(8.25) \quad Cov(x_t, x_{t+i}) = \alpha^i \frac{\sigma^2}{1 - \alpha^2}$$

$$(8.26) \quad Cov(\epsilon_t, x_{t+i}) = \alpha^i \sigma^2$$

we find

$$(8.27) \quad \sigma_T^2 = \sigma^2 \left(T + \frac{T\alpha^2}{1 - \alpha^2} + 2 \left(\frac{\alpha^2}{1 - \alpha^2} + \alpha \right) \sum_{i=1}^{T-1} \alpha^i (T-i) \right)$$

REFERENCES

- [1] Clifford Ball and Walter Torous. Regime shifts in short term riskless interest rates. University of California at Los Angeles, Anderson Graduate School of Management 1141, Anderson Graduate School of Management, UCLA, August 1995. available at <http://ideas.repec.org/p/cdl/anderf/1141.html>.
- [2] O.E. Barndor-Nielsen. Normal inverse gaussian distributions and stochastic volatility modelling. *Scandinavian Journal of statistics*, 24:1–13, 1997.
- [3] Basel Committee on Banking Supervision. International convergence of capital measurement and capital standards: A revised framework. Available online at <http://www.bis.org>, November 2005.
- [4] T. Bollerslev. A conditional heteroskedastic model for speculating prices and rates of return. *Review of Economics and Statistics*, 69:542–547, 1987.
- [5] J Boudoukh, R F Whitelaw, M Richardson, and R Stanton. Pricing mortgage-backed securities in a multifactor interest rate environment: a multivariate density estimation approach. *Review of Financial Studies*, 20(3):769 – 811, 2005.
- [6] Joseph L. Breeden. Modeling data with multiple time dimensions. *Computational Statistics & Data Analysis*, 51:4761 – 4785, May 17 2007.
- [7] Joseph L. Breeden, Lyn C. Thomas, and John McDonald III. Stress testing retail loan portfolios with macroeconomic scenarios. Technical report, Strategic Analytics, 2007.
- [8] Peter J. Brockwell and Richard A. Davis. *Introduction to time series and forecasting*. Springer, 2002.
- [9] F. Diebold and R. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13:253–263, 1995.
- [10] A. Dubi. *Monte Carlo Applications in Systems Engineering*. Wiley, 2001.
- [11] Bradley Efron. The two-way proportional hazards model. *Journal of the Royal Statistical Society B*, 64:899 – 909, 2002.
- [12] Walter Enders. *Applied Econometric Time Series, Second Edition*. John Wiley & Sons, 2004.
- [13] David Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, 2005.

- [14] K R French, G W Schwert, and R F Stambaugh. Expected stock returns and volatility. *Journal of Financial Economics*, 19:330, 1987.
- [15] Norval D. Glenn. *Cohort Analysis, 2nd Edition*. Sage, London, 2005.
- [16] C. Granger and P. Newbold. Forecasting transformed series. *Journal of the Royal Statistical Society B*, 38:189–203, 1976.
- [17] Damodar N. Gujarati. *Basic Econometrics, 3. ed.* McGraw-Hill, 1995.
- [18] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, New York, 1990.
- [19] Anna Kalemanova and Ralf Wernery. A short note on the efficient implementation of the normal inverse gaussian distribution. Technical report, risklab germany, November 27 2006.
- [20] W.M. Mason and S. Fienberg. *Cohort Analysis in Social Research: Beyond the Identification Problem*. Springer, 1985.
- [21] Don L. McLeish. *Monte Carlo Simulation and Finance*. Wiley, 2005.
- [22] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, 1992.
- [23] Anthony B. Sanders and Haluk Unal. On the intertemporal behavior of the short-term rate of interest. *The Journal of Financial and Quantitative Analysis*, 23(4):417–423, December 1988.