Predicting Risk at Short Horizons
A Case Study for the USE4D Model

Jose Menchero, Andrei Morozov and Andrea Pasqua

Jose.Menchero@msci.com
Andrei.Morozov@msci.com
Andrea.Pasqua@msci.com

January 2013
Contents

Introduction.................................................................................................................. 3

Evaluating the Accuracy of Risk Forecasts ......................................................... 3

  Bias Statistics............................................................................................................ 4

  Mean Rolling Absolute Deviation (MRAD)......................................................... 4

  Adjusted MRAD...................................................................................................... 7

  Cross-sectional Bias Statistics............................................................................. 7

  Q-statistics................................................................................................................ 7

The USE4 Model and Data Set............................................................... 10

Estimating Volatility......................................................................................... 11

  Exponentially Weighted Moving Averages (EWMA).................................... 11

  GARCH(1,1).......................................................................................................... 13

  Volatility Regime Adjustment (VRA)................................................................. 13

Empirical Results.............................................................................................. 16

Setting the Correlation Half-Life................................................................. 22

Conclusion............................................................................................................. 24

References............................................................................................................. 25

Client Service Information is Available 24 Hours a Day............................. 26

Notice and Disclaimer.......................................................................................... 26

About MSCI........................................................................................................... 26
Introduction

Equity factor models are a fairly recent invention. Barr Rosenberg pioneered the use of multi-factor risk models as a robust way to estimate the asset covariance matrix (1974). In 1975 he founded Barra, which developed the first commercially available risk model for US equities, dubbed USE1.

Initially, the USE1 Model was estimated from quarterly data. Later, as data become more widely available, the observation frequency was increased to monthly. For many years, using monthly observations to estimate the factor covariance matrix was standard practice. For instance, the Barra USE3 Model, released in 1998, used monthly factor returns with a half-life of 90 months in the estimation process.

The Internet Bubble presented a serious challenge for models estimated with monthly data. The crux of the problem was that volatility changes were too rapid and extreme to be reliably captured using low-frequency observations. In response to this challenge, Barra researchers developed the USE3S Model, which used daily factor returns for estimating the covariance matrix. The higher frequency of observations allowed the model to adapt more rapidly to changing levels of volatility.

Although the USE3S Model employed daily factor returns, it maintained a prediction horizon of one month. This was accomplished by explicitly accounting for the effects of serial correlation in factor returns, which can cause significant deviations from the familiar “square-root-of-time” scaling. For instance, the returns to the Momentum factor typically exhibit positive serial correlation. This makes the factor significantly more volatile at the monthly horizon than would be suggested by applying square-root-of-time scaling to daily volatility.

For many institutional investors, however, the relevant horizon may be much shorter than one month. In this paper, we focus on one-day prediction horizons. One possible approach for predicting risk at one-day horizons would be to take the forecasts from a monthly model and simply apply square-root-of-time scaling to bring the prediction horizon to a single day. However, there are two shortcomings with this approach. First, the monthly model is calibrated for a longer horizon and will not have the appropriate responsiveness for a one-day forecast. Second, the serial correlation adjustments that were used to provide accurate forecasts at a monthly horizon now represent sources of error at a one-day horizon. In other words, if the observation frequency is synchronized with the prediction horizon, then serial correlation adjustments should not be incorporated.

In this paper, we highlight some of the modeling issues that must be addressed when constructing a model with a one-day prediction horizon. Central to this challenge is the identification of a reliable metric to evaluate the accuracy of risk forecasts.

Evaluating the Accuracy of Risk Forecasts

Building a sound risk model first requires a reliable means of evaluating the accuracy of risk forecasts. This provides an essential guide for setting model parameters and evaluating performance. Ideally, our measure of forecasting accuracy will provide both time and portfolio resolution, meaning that the measure can be used to evaluate risk forecasts for a single portfolio across time, or for a collection of portfolios within a single time period.
Bias Statistics

One commonly used measure to evaluate the accuracy of risk forecasts is the bias statistic, which conceptually represents the ratio of realized risk to forecast risk. To compute the bias statistic for a portfolio \( n \), we first use the risk model to predict the portfolio volatility \( \sigma_{nt} \) at the start of every period \( t \). We then observe the out-of-sample return of the portfolio \( R_{nt} \) over the subsequent period. The standardized return is defined by

\[
z_{nt} = \frac{R_{nt}}{\sigma_{nt}},
\]

and expresses the portfolio return as a z-score. The bias statistic is given by the standard deviation of standardized returns,

\[
B_n = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} (z_{nt} - \bar{z}_n)^2},
\]

where \( T \) is the number of days in the testing window.

Conceptually, the bias statistic measures whether the risk forecasts were accurate, on average, for a single portfolio across time. For accurate forecasts, we expect the realized bias statistic to be close to 1. However, due to sampling error, the bias statistic will never be exactly 1 — even for perfect risk forecasts. Instead, it is customary to identify a confidence interval. Assuming normally distributed returns and perfect forecasts, the 95-percent confidence interval is approximately \( 1 \pm \sqrt{2/T} \).

Two attractive features of the bias statistic are that it is simple to interpret and provides portfolio resolution. Unfortunately, it does not provide time resolution, meaning that it is possible to under-predict risk for some sub-periods and over-predict it for others while nonetheless obtaining a bias statistic close to 1. In other words, the bias statistic may allow cancellation of errors across time. This becomes especially problematic over long sample periods encompassing many years and multiple market regimes. A risk model user must be confident that volatility forecasts are reliable for all market regimes — not just “on average.”

Mean Rolling Absolute Deviation (MRAD)

One way to partially mitigate the time-resolution problem is to divide the long sample period into smaller sub-periods, and to compute the bias statistics over these sub-periods. For instance, if we select 12 days as the length of our sub-period, we obtain,

\[
B_n^\tau = \sqrt{\frac{1}{11} \sum_{t=\tau}^{\tau+11} (z_{nt} - \bar{z}_n)^2},
\]

where \( \tau \) denotes the start of the 12-day sub-period. The window is then rolled forward one day at a time until reaching the end of the entire sample.

The conventional Mean Rolling Absolute Deviation, or MRAD, is defined as the mean absolute deviation of the bias statistics from their ideal value of 1,
\[ MRAD = \frac{1}{N(T-1)} \sum_{n} |B_n^T - 1|, \]  

where \( N \) is the number of portfolios and \( (T - 1) \) is the total number of (overlapping) 12-day sub-periods. By virtue of the absolute value, MRAD penalizes both under-prediction and over-prediction of risk. Assuming perfect forecasts, normally distributed returns, and 12-day sub-periods, it may be shown that the expected value of MRAD is approximately 0.17.

Real financial data, of course, tend to have fat tails. In Figure 1, we plot the expected value of MRAD versus kurtosis for perfect risk forecasts. These results were generated via simulation by drawing 12 randomly generated returns from a fat-tailed distribution with standard deviation equal to 1. The fat tails were obtained using a student \( t \)-distribution with the appropriate number of degrees of freedom. We see that as the kurtosis rises to even modest levels, the expected MRAD for perfect forecasts increases significantly.

**Figure 1:** MRAD versus kurtosis for perfect risk forecasts. Kurtosis levels were varied using a student \( t \)-distribution with the appropriate number of degrees of freedom.

![Graph](image)
In Figure 2, we present the results of the simulation study. Whereas the true optimal half-life is infinite, Figure 2 spuriously points to an optimal half-life of only six days. Furthermore, the minimum MRAD is approximately 0.16, well below the theoretical lower bound of 0.17 for perfect forecasts. How can these puzzling results be explained?

This conundrum is resolved by recognizing that the bias statistic — and by extension the MRAD — is not a true out-of-sample measure. This may be surprising at first glance, since the bias statistic is computed using out-of-sample z-scores. Nevertheless, because risk forecasts are updated daily, the bias statistic uses information in the volatility forecasts that was not known at the start of the 12-day period.

That this tends to favor overly responsive risk forecasts may be seen by the following thought experiment. Assume for simplicity that returns are drawn from a standard normal distribution. Suppose that, by chance, the portfolio experiences several large standardized returns. In this case, the 12-day bias statistic is likely to exceed 1. However, by over-reacting to the event and dramatically increasing risk forecasts, the subsequent z-scores will be artificially reduced. Although the risk forecasts following the event are biased upward, the realized bias statistic is closer to 1. Conversely, a string of small returns is likely to lead to a bias statistic less than 1. By reducing risk forecasts, the subsequent standardized returns will be artificially inflated, again leading to a bias statistic closer to 1. In other words, excessive responsiveness produces noisy and inaccurate risk forecasts, although the realized bias statistics may be deceptively close to 1.
Adjusted MRAD

A straightforward way to remove the in-sample effect is to simply use risk forecasts taken from the start of the 12-day period, and to hold these forecasts constant over the next 12 days. This leads to the adjusted MRAD measure. An advantage of the adjusted MRAD measure is that it uses no information in the volatility forecasts that was not known at the start of the 12-day sub-period, and therefore represents a true out-of-sample measure. A disadvantage of the adjusted MRAD is that it requires using somewhat stale forecasts.

Cross-Sectional Bias Statistics

Another measure of forecasting accuracy is the cross-sectional bias statistic,

$$B_T = \sqrt{\frac{1}{N-1} \sum_{n=1}^{N} (z_{nt} - \bar{z}_t)^2}.$$  

(5)

This is similar to the time-series bias statistic of Equation 2, except that now the standard deviation is computed across portfolios for a single day $t$. Since the cross-sectional bias statistic uses no information for volatility forecasts that was unknown at the start of the sub-period, it represents a true out-of-sample measure.

Conceptually, the cross-sectional bias statistic measures whether the risk model forecasts were accurate, on average, for a collection of portfolios on a single day. It cannot, however, answer whether the risk model forecasts were accurate for individual portfolios. In other words, the cross-sectional bias statistic provides time resolution, but not portfolio resolution. Consequently, it is possible to obtain a cross-sectional bias statistic close to 1, even if risk forecasts for individual portfolios were poor. In particular, this may occur if over-prediction errors for some portfolios are canceled by under-prediction errors for others.

$Q$-statistics

Patton (2011) describes measures of forecast accuracy in terms of “loss functions.” He defines a loss function as “robust” if the ranking of any two volatility forecasts by expected loss is the same whether the ranking is done using the true variance (unobservable) or some unbiased variance proxy (e.g., squared return). One example of a robust loss function is the $Q$-statistic, defined for portfolio $n$ and time $t$ as

$$Q_{nt} = z^2_{nt} - \ln\left(z^2_{nt}\right).$$  

(6)

Patton further shows that the $Q$-statistic is the unique loss function (up to trivial additive and multiplicative constants) that depends solely on standardized returns (i.e., $z$-scores). This makes the $Q$-statistic ideal for evaluating risk model accuracy, because it places every observation on an equal footing (whether the volatility is high or low).

Another key property of robust loss functions is that they are minimized in expectation when the predicted volatility equals the true volatility. For other loss functions, this is not true. That is, a biased volatility forecast can minimize the loss function. This would obviously be problematic for risk model calibration purposes. The MRAD is an example of a loss function that is not robust.
Also note that the $Q$-statistic clearly satisfies both time and portfolio resolution. That is, it is not based on averaging, so it is not possible to offset an over-forecasting error for one observation with an under-forecasting error for a different observation. Of course, the $Q$-statistic can be averaged either across time or across portfolios to obtain the desired resolution.

Intuitively, we can think of the $Q$-statistic as being comprised of two penalty functions. The first term, $z_n^2$, becomes large when risk forecasts are too low and therefore represents an under-forecasting penalty function. The second term, $-\ln\left(z_n^2\right)$, represents the over-forecasting penalty function and dominates when risk forecasts are too high.

It is important to understand the expected value of the $Q$-statistic for perfect risk forecasts. If returns are normally distributed, the expected value of the $Q$-statistic is approximately 2.27. As with the MRAD, however, it is a rather sensitive function of kurtosis. In Figure 3 we plot the expected value of the $Q$-statistic versus kurtosis, assuming perfect risk forecasts.

Figure 3: $Q$-statistic versus kurtosis for perfect risk forecasts. Kurtosis levels were varied using a student $t$-distribution with the appropriate number of degrees of freedom.

It is also interesting to consider how biases in volatility forecasts affect changes in the $Q$-statistic. Assume that returns are drawn from a standard normal distribution (i.e., with true volatility of 1). Let the volatility forecasts be denoted by $\sigma_p$. One may show that the expected loss relative to perfect forecasts is given by

$$E[\Delta Q] = \frac{1}{\sigma_p^2} + 2 \ln\left(\sigma_p\right) - 1.$$  \hspace{1cm} (7)

This function is plotted in Figure 4. If the predicted volatility is equal to the true volatility ($\sigma_p = 1$), then the expected loss is zero. Note also that the loss function is asymmetric. That is, it penalizes under-
prediction of risk more heavily than over-prediction of risk. For instance, if the predicted volatility is half the true volatility, then the increase in $Q$-statistic is approximately 1.61. However, if the predicted volatility is double the true volatility, then the increase in $Q$-statistic is only about 0.64. This is another attractive feature of the $Q$-statistic, since most investment managers regard under-prediction of risk as more problematic than over-prediction of risk.

**Figure 4:** Increase in $Q$-statistic versus predicted volatility when the true volatility is 1. Results were obtained assuming a normal distribution. The asymmetry indicates that the $Q$-statistic penalizes under-forecasting of risk more heavily than over-forecasting of risk.

It is instructive to repeat the simulation exercise of Figure 2, except now using the adjusted MRAD and $Q$-statistics. The results are presented in Figure 5. We see that neither of these measures exhibits a spurious minimum, thus correctly pointing to an infinite optimal half-life.
Figure 5: Q-statistic and adjusted MRAD versus EWMA volatility half-life. Returns were generated using a standard normal distribution. Neither quantity exhibits a spurious minimum, thus correctly pointing to an infinite half-life.

The USE4 Model and Data Set

The empirical data set used for this study consists of the history of daily factor returns from the Barra US Equity Risk Model (USE4). The history of daily factor returns starts in 1993, while the volatility forecasts in our empirical study begin on July 19, 1995. The data set contains 4,328 trading days, ending on September 21, 2012.

The USE4 Model is available in long horizon (USE4L) and short-horizon (USE4S) versions. Both models share the same factor structure and factor returns, but differ in their responsiveness. The USE4S Model is designed to provide the most accurate forecasts at a one-month prediction horizon. The USE4L Model is tailored for longer-term investors who are willing to trade some degree of forecasting accuracy for greater stability in the risk forecasts. With the launch of USE4D, the USE4 Model is now available in a third version for daily horizon forecasts. Again, the USE4D Model shares the same factor structure and factor returns as the two other USE4 Model variants.

Here, we briefly review some highlights of the USE4 Model. Empirical results can be found in Yang, Menchero, Orr, and Wang (2011), while methodology details are described in Menchero, Orr, and Wang (2011). The USE4 Model contains 73 factors, comprised of: (a) the Country factor, which is the regression intercept, (b) 60 industry factors, and (c) 12 style factors. Factor returns are estimated by performing daily cross-sectional regressions of stock returns against the start-of-day factor exposures. The estimation universe is the MSCI USA IMI, a broad index representing the US market. The model
employs square-root of market capitalization as the regression weights. The exact colinearity between industry factor exposures and the Country factor is removed by constraining the industry factor returns to be cap-weighted mean zero every period.

As described by Menchero (2010), the estimated factor returns may be interpreted as the returns of factor-mimicking portfolios. The Country factor portfolio essentially represents the cap-weighted estimation universe (i.e., MSCI USA IMI). The industry factor portfolios are dollar-neutral and capture the performance of the industry net of the market and other styles. The style factor portfolios are dollar-neutral and have unit tilt on the particular style, with zero exposure to industries and other styles.

Estimating Volatility

The simplest way to estimate volatility is to compute the sample standard deviation over a trailing window. If stock return distributions were stationary, then using the maximum sample size and equally weighting every observation would minimize sampling error and hence produce the most accurate forecast.

Stock return distributions, however, are not stationary. Events that occurred ten years ago have little to do with current volatility levels. Therefore, to reflect current market conditions, we must give more weight to recent observations. The crucial question is how much more? On the one hand, if we give too much weight to recent observations, then we base our estimates on very few data points and we are strongly penalized with sampling error. By contrast, if we give too much weight to distant observations, then we are harmed by incorporating stale data into our volatility estimates. Making good risk forecasts for real financial data requires an optimal tradeoff between these two effects.

Exponentially Weighted Moving Averages (EWMA)

A simple, yet effective, technique for attaching more weight to recent observations is exponential weighted moving averages (EWMA). In this approach, the variance is written as a weighted average of lagged squared returns.

\[ \sigma^2_t = \sum_m w_m r_{t-m}^2, \]  

(8)

where the weights \( w_m \) sum to 1, and decrease by fixed proportion every period. More specifically, \( w_{m+1} = \lambda w_m \), where \( \lambda \) is the decay factor. The EWMA estimate can be conveniently rewritten in recursive form as a weighted average between yesterday’s squared return and yesterday’s variance forecast,

\[ \sigma^2_t = (1 - \lambda) r_{t-1}^2 + \lambda \sigma^2_{t-1}. \]  

(9)

The relative weight we place on yesterday’s variance forecast determines the responsiveness of the model and is related to the EWMA volatility half-life parameter as follows,

\[ HL = \frac{\ln(0.5)}{\ln(\lambda)}. \]  

(10)
The EWMA volatility half-life is a critical model parameter. To guide us in selecting an appropriate value, we compute the average adjusted MRAD and $Q$-statistic for USE4 daily factor returns, with the averages taken over all factors and time periods. In Figure 6, we plot these quantities versus the EWMA volatility half-life. Unlike the simulated results in Figure 5, we see that real financial data do exhibit a clear minimum. Figure 6 shows that the optimal volatility half-life by both the $Q$-statistic and the adjusted MRAD is about 21 trading days, or one month. Reassuringly, both out-of-sample statistics point to the same optimal half-life. Note that a 21-day volatility half-life corresponds to a decay factor of 0.97 according to Equation 10.

**Figure 6:** $Q$-statistic, adjusted MRAD, and conventional MRAD versus volatility half-life for EWMA forecasts using USE4 daily factor returns. The optimal half-life is about one month (21 trading days) by either the $Q$-statistic or the adjusted MRAD. The conventional MRAD spuriously points to an optimal half-life of about five trading days.

For comparison purposes, we also present the conventional MRAD in Figure 6. Note that this measure suggests an optimal half-life of just five days. Also observe that the minimum MRAD is well below the minimum value of the adjusted MRAD. These results illustrate — this time using real financial data — the same effect seen in the simulated results of Figure 2. That is, in-sample effects spuriously point to excessively short half-life parameters, which in turn produce deceptively attractive MRAD values.
GARCH(1,1)

The method of Generalized Auto Regressive Conditional Heteroskedasticity, or GARCH, was developed by Engle (1982), Bollerslev (1986) and others in the 1980s. By now, there are many variations of GARCH. An excellent survey of these techniques is provided by Bauwens, Laurent, and Rombouts (2006).

In this paper, we focus on the most widespread variant of GARCH, namely GARCH(1,1). In this approach, variance estimates are given by

\[ \sigma_t^2 = \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma \sigma_{t-1} \]  

This expresses predicted variance as a weighted average of three terms: (a) yesterday’s squared return, with weight \( \alpha \), (b) yesterday’s variance forecast, with weight \( \beta \), and (c) a long-run variance, with weight \( \gamma \). The stability condition requires that the weights sum to 1,

\[ \alpha + \beta + \gamma = 1 \]  

We determine the parameters using maximum likelihood estimation with a two-year look-back window. We estimate GARCH parameters separately for every factor in the model.

GARCH has several conceptually appealing attributes. First, the model provides for mean reversion in volatility forecasts through the long-run variance term in Equation 11. Also, GARCH replicates the volatility clustering often found in financial time series. Finally, by setting the \( \gamma \) parameter to zero, GARCH(1,1) encompasses EWMA as a special limiting case.

Volatility Regime Adjustment (VRA)

The USE4D Model utilizes the Volatility Regime Adjustment (VRA) technique to estimate factor volatilities. This technique was first introduced with the launch of the Barra US Equity Model (USE4), as described by Menchero, Orr, and Wang (2011). The central concept behind this approach is to use cross-sectional observations to calibrate the model to current volatility levels.

Risk models base forecasts on historical observations. However, since volatility levels vary across time, this may lead to certain biases in the risk forecasts. For instance, entering a period of financial crisis, volatility levels tend to rise. In this case, risk models use the lower-volatility past to predict the higher-volatility future, thus causing a tendency to under-predict risk during these periods. Conversely, immediately following a period of financial turmoil, there are many extreme events in the estimation window, which creates a tendency to over-predict risk at such times. Since risk models are not crystal balls, there is no way to completely eliminate these biases. The VRA technique, however, is designed to mitigate these biases by using cross-sectional observations to calibrate the model to current volatility levels.

The first step in the VRA technique is to estimate factor volatilities using EWMA, as in Equation 8. This estimate is characterized by the volatility half-life parameter, defined by Equation 10. Next, we compute the cross-sectional bias statistic, as in Equation 5. The computation is performed over the z-scores of daily factor returns, using the EWMA volatility forecasts from the first step. We can think of the cross-sectional bias statistic as providing an instantaneous measure of risk forecasting bias. The factor volatility multiplier, at time \( t \), is defined in terms of the trailing cross-sectional bias statistics,
\[ \lambda_t^F = \sqrt{\sum_{i} \mu_m B_{t-i}^2} , \]  

(13)

where \( \mu_m \) is the exponential weight characterized by the VRA half-life parameter.

If the factor volatility multiplier is greater than 1, it says the original EWMA estimates were too low, and risk forecasts should be adjusted upward. Conversely, when \( \lambda_t^F \) is less than 1, EWMA estimates were too high and risk forecasts should be adjusted downward. The adjusted volatility forecast for factor \( k \) at time \( t \) is given by

\[ \tilde{\sigma}_{k_t} = \lambda_t^F \sigma_{k_t} . \]  

(14)

This adjustment is designed to remove the average bias, with all factor volatilities scaled by the same proportion.

Volatility estimation with the VRA technique therefore requires two half-life parameters: the EWMA volatility half-life defined in Equation 10, and the VRA half-life used in Equation 13. To find the optimal values for these two parameters, we perform a search for the global minimum Q-statistic over the two-dimensional space. In Figure 7, we plot curves of Q-statistics (averaged across factors and time periods) versus the VRA half-life for several values of the EWMA volatility half-life. We find that the combination that minimizes the Q-statistic is a 42-day EWMA volatility half-life in conjunction with a two-day VRA half-life.

Figure 7: Q-statistic curves versus VRA half-life for three values of EWMA volatility half-life. The Q-statistic is minimized at a VRA half-life of two days and an EWMA volatility half-life of 42 days.
At first glance, a two-day VRA half-life may seem surprisingly short. Recall, however, that it is primarily the penalty of sampling error that imposes limits on the responsiveness of the model. In the EWMA approach, where volatilities are estimated one factor at a time, sampling error is relatively large. In the VRA approach, by contrast, we largely mitigate sampling error by virtue of using 73 cross-sectional observations every day.

Although a two-day VRA half-life is optimal in terms of forecasting accuracy, it may lead to rather large daily fluctuations in predicted volatilities. To reduce such variability, the USE4D Model combines the 42-day EWMA volatility half-life with a four-day VRA half-life. This provides considerably more stability in volatility forecasts, while producing only a modest increase in the Q-statistic, as evident from Figure 7.

The financial crisis period of 2008 and 2009 affords a prime example to illustrate the behavior of the VRA technique. In Figure 8, we plot the factor volatility multiplier $\lambda^F_t$ together with the factor cross-sectional volatility (CSV), defined as

$$CSV_t = \sqrt{\frac{1}{K} \sum_k f_{kt}^2}.$$  

(15)

Figure 8: Factor volatility multiplier and factor cross-sectional volatility (CSV) for the USE4D model. Results were smoothed using 10-day rolling windows.
In January 2008, we see that factor CSV spiked from 60 bps to about 150 bps per day. The factor volatility multiplier responded quickly to the increased volatility levels, reaching a maximum value of 1.7 during the month. The next big spike in factor CSV came at the start of September 2008, when levels stood at about 100 bps per day. Over the next two months, these levels had more than doubled to 250 bps per day. Again, the factor volatility multiplier quickly detected the increased volatility levels, hitting a peak of 1.8 in October 2008.

By the end of April, 2009, the worst of the financial storm was over, and factor volatility levels stood at about 150 bps per day. The next eight months saw a strong secular decline in volatility levels, with factor CSV hitting a low of about 50 bps per day by the end of 2009. Over this time of rapidly declining volatility, the factor volatility multiplier was well below 1, hitting lows of about 0.6 during this period. This example demonstrates the effectiveness of the VRA technique for reducing volatility forecasts in the aftermath of a financial crisis.

### Empirical Results

In this section, we compare forecasts for six different sets of volatility estimates. The data set used for our study is the full history of USE4 daily factor returns, as described previously. The first set of volatility estimates is taken directly from the USE4D model, which uses a volatility half-life of 42 days and a VRA half-life of four days. Next, we consider a highly responsive EWMA forecast, with volatility half-life of four days (i.e., equal to the VRA half-life of USE4D). We refer to this as the EWMA(4) forecast. We also consider the EWMA(21) forecast, which uses the optimal volatility half-life of 21 days, as seen in Figure 6. The last EWMA forecast we consider is EWMA(42), which uses the same volatility half-life as USE4D. We also report results for the GARCH(1,1) model, described above. Finally, we examine the USE4S model, with volatility forecasts scaled to a one-day horizon.

Before presenting results on forecasting accuracy, we consider the stability of the volatility forecasts. To investigate this, we first compute for factor $k$ and day $t$ the absolute change in volatility forecast over the previous day,

$$v_{k,t} = \frac{|\sigma_{k,t} - \sigma_{k,t-1}|}{\sigma_{k,t-1}}. \quad (16)$$

We then define the forecast variability as the average over all factors and all time periods,

$$\overline{\sigma} = \frac{1}{KT} \sum_{k,t} v_{k,t}. \quad (17)$$

The variability for each volatility forecast is reported in Table 1. The variability ranged from a low of 63 bps for USE4S, to a high of 7.96 percent for EWMA(4). The USE4D Model had an intermediate variability at 2.82 percent.
Table 1: Comparison for six different volatility estimates. Averages were computed across all 73 factors and 4,328 trading days. Differences in Q-statistics are relative to the USE4D model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Variability (Percent)</th>
<th>Conventional MRAD</th>
<th>Adjusted MRAD</th>
<th>Q-Statistic</th>
<th>Q-Statistic (Difference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USE4D</td>
<td>2.82</td>
<td>0.2124</td>
<td>0.2478</td>
<td>2.4182</td>
<td>0.0000</td>
</tr>
<tr>
<td>EWMA (4)</td>
<td>7.96</td>
<td>0.2059</td>
<td>0.3092</td>
<td>2.5231</td>
<td>0.1049</td>
</tr>
<tr>
<td>EWMA (21)</td>
<td>1.66</td>
<td>0.2239</td>
<td>0.2562</td>
<td>2.4455</td>
<td>0.0274</td>
</tr>
<tr>
<td>EWMA (42)</td>
<td>0.85</td>
<td>0.2403</td>
<td>0.2590</td>
<td>2.4564</td>
<td>0.0382</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>4.33</td>
<td>0.2147</td>
<td>0.2622</td>
<td>2.4486</td>
<td>0.0304</td>
</tr>
<tr>
<td>USE4S</td>
<td>0.63</td>
<td>0.2507</td>
<td>0.2647</td>
<td>2.4631</td>
<td>0.0450</td>
</tr>
</tbody>
</table>

In Table 1, we also report MRAD and Q-statistics for the six volatility forecasts. By the conventional MRAD measure, EWMA(4) has the lowest score. However, as we have seen, the conventional MRAD strongly favors overly responsive forecasts which can make the MRAD appear artificially low. The true out-of-sample measures (adjusted MRAD and Q-statistic) paint a different picture of forecasting accuracy.

By either the adjusted MRAD or the Q-statistic, the USE4D model produced the most accurate volatility forecasts. We also report in Table 1 the difference in average Q-statistic relative to the USE4D Model. The average was computed over all 73 factors and the entire sample period of 4,328 trading days. The second most accurate forecasts were provided by the EWMA(21) model, which had a Q-statistic 0.0274 above USE4D. The least accurate forecasts were for the EWMA(4) model, which scored a Q-statistic 0.1049 above the USE4D model.

It is a useful exercise to translate differences in Q-statistics into forecasting errors. This is difficult to do precisely, since it depends on the forecast errors of each estimate as well as the return distributions of the test portfolios. Nevertheless, a rough sense can be gleaned from Equation 7 and Figure 4. We see that a difference in Q-statistic of 0.10 translates roughly into a predicted volatility of $\sigma_p = 0.81$ (i.e., 19 percent under-forecasting bias) or $\sigma_p = 1.28$ (28 percent over-prediction). Similarly, a Q-statistic difference of 0.03 corresponds roughly to forecasts of $\sigma_p = 0.89$ (11 percent under-forecast) or $\sigma_p = 1.14$ (14 percent over-forecast).

It is also important to consider the statistical significance of these findings. One way to gauge statistical significance is to count the number of factors for which USE4D produced lower Q-statistics than the other forecasts. Compared with GARCH(1,1), the USE4D model produced lower Q-statistics for 71 of the 73 factors. Relative to EWMA(42) or USE4S, the USE4D model outperformed for 72 of the 73 factors. Finally, versus EWMA(4) or EWMA(21), the USE4D model outperformed for all factors. The likelihood of these results occurring by mere chance is exceedingly small. Thus, we can conclude with high statistical confidence that the USE4D model provided more accurate risk forecasts than the other approaches.

Although, on average, the USE4D Model produced more accurate forecasts over the full sample period, it is important to investigate the persistence of these results across time. For instance, if the outperformance was due primarily to a specific sub-period (such as the 2008/2009 financial crisis), it may call into question the robustness of the VRA methodology. To explore this possibility, we define the cumulative Q increment,
\[ \Delta \bar{Q}(\tau) = \frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} \left( Q_{kt}^A - Q_{kt}^B \right), \]  

(18)

where \( Q_{kt}^A \) is the Q-statistic for forecast \( A \), factor \( k \), at time \( t \), and \( Q_{kt}^B \) is the corresponding quantity for forecast \( B \). We can think of \( \Delta \bar{Q}(\tau) \) as keeping a “running tally” of the difference in accuracy between forecasts \( A \) and \( B \). At the end of the sample period, \( \tau = T \), the cumulative Q increment becomes simply the average difference in Q-statistic between the two forecasts. Here we let forecast \( B \) represent the USE4D model, whereas forecast \( A \) is used to represent the alternative.

In Figure 9 we plot the cumulative Q increment for all models relative to USE4D. Note that all of the lines have a persistent upward slope, indicating that the USE4D model consistently provided more accurate forecasts across all market regimes.

**Figure 9:** Cumulative Q increment (relative to USE4D) versus time for five sets of volatility forecasts. The persistent upward slope indicates that the USE4D consistently provided more accurate forecasts across different market regimes.

In order to develop greater intuition for how these forecasts compare, it is useful to compare time-series plots of the volatility forecasts for specific factors.

In Figure 10, we plot the predicted volatility of the USE4 Country factor computed with EWMA(42) and USE4D, over the period 2007-2010. We see that over the last four months of 2008, the USE4D Model predicted significantly higher volatility for the Country factor than the EWMA(42) forecast.

Conversely, for the last eight months of 2009, the EWMA(42) forecasts were significantly higher. These results are consistent with the greater responsiveness of USE4D relative to EWMA(42).
Figure 10: Predicted volatility of USE4 Country factor for USE4D and EWMA(42) forecasts.

In Figure 11 we plot the volatility of the Earnings Yield factor for GARCH(1,1) and USE4D over the time period 2007-2010. This was a factor to which many quantitative investors had positive exposure over this period.

Prior to the “Quant Meltdown” of August 2007, we see that the GARCH(1,1) forecasts were very stable and about at the same level as USE4D forecasts. During the height of the Quant Meltdown, the GARCH(1,1) forecasts were above 12 percent, versus eight percent for USE4D.

For the next two years, while the Quant Meltdown remained in the estimation window, the GARCH(1,1) forecasts were very jumpy. In August 2009, as the event exits the estimation window, we see that the GARCH(1,1) forecasts suddenly become more stable.
Figure 11: Predicted volatility of USE4 Earnings Yield factor for USE4D and GARCH(1,1) forecasts.

As we have seen, the $Q$-statistic is composed of an over-forecasting and an under-forecasting penalty function. For perfect risk forecasts, the expected value of the under-forecasting penalty function is exactly 1. Assuming normal distribution and perfect risk forecasts, the expected value of the over-forecasting penalty function is 1.27. Plotting these two penalty functions versus time and comparing them to their ideal values can yield important insight into whether the volatility forecasts were too high or too low.

In Figure 12, we plot the mean $Q$-statistic, together with the over-forecasting and under-forecasting penalty functions, averaged across all 73 factors, for EWMA(42) volatility forecasts. The plot covers the financial crisis period of 2008 and 2009, and lines were smoothed using 10-day rolling averages. The $Q$-statistic exhibits several distinct peaks over the course of 2008, with the largest such peak occurring in October 2008. We see that each of these peaks in 2008 was caused by a spike in the under-forecasting penalty function, implying that the EWMA(42) forecast was not responsive enough during this period. We can even use the $Q$-statistic to estimate the magnitude of the bias. In October 2008, the $Q$-statistic was at about 3.9, or about 1.6 above the ideal position. From Figure 4, we see this translates roughly into a 50 percent under-prediction of risk. By contrast, we see that over the last eight months of 2009, the relatively high value of the $Q$-statistic was dominated by the over-forecasting penalty function. In this case, the EWMA(42) model did not adapt quickly enough to reduced volatility levels in the wake of the financial crisis.
Figure 12: Plot of Q-statistic together with over-forecasting and under-forecasting penalty functions for the EWMA(42) forecast. The quantities represent averages across factors. Lines were smoothed using 10-day rolling windows. The EWMA(42) forecast generally under-forecasts for much of 2008 and over-forecasts for most of 2009.

In Figure 13 we plot for the USE4D forecasts the mean Q-statistic, averaged across all factors, together with over-forecasting and under-forecasting penalty functions. From Figure 13, it is clearly evident that the Q-statistic for USE4D was much closer to the ideal value of 2.27 than the corresponding plot in Figure 12 for EWMA(42). Furthermore, we see that the over-forecasting and under-forecasting penalty functions never deviated too far from their ideal positions.
Figure 13: Plot of Q-statistic together with over-forecasting and under-forecasting penalty functions for the USE4D forecast. The quantities represent averages across factors. Lines were smoothed using 10-day rolling windows. Compared to EWMA(42) forecasts, the USE4D Q-statistics are much closer to their ideal values. Furthermore, the over-forecasting and under-forecasting penalties stay relatively close to their ideal values, indicating that the model adapts well to changing volatility levels.

Setting the Correlation Half-Life

Thus far, we have only considered volatility forecasts. In building a factor covariance matrix, however, we must also estimate the off-diagonal covariances. These are given by the product of the factor volatilities and the factor correlations. The correlations, in turn, are estimated using EWMA with a specified half-life parameter.

The correlation half-life is a crucial model parameter. If it is too long, there may be a penalty by using stale data in our correlation estimates. If the correlation half-life is too short, then the covariance matrix may become noisy and ill-conditioned, leading to poor performance and under-estimation of risk of optimized portfolios, as described by Menchero, Wang, and Orr (2011).

One way to determine the optimal correlation half-life is to compute the value that produces the best out-of-sample performance of optimized portfolios. To study this, we construct the minimum-risk fully invested factor portfolio at the start of every day. The Country factor exposure and the sum of industry factor exposures are constrained to equal 1. In Figure 14 we plot the out-of-sample realized volatility over the full sample period as a function of correlation half-life. We see that 200 days appears optimal, as this minimizes the out-of-sample realized volatility. If the correlation half-life is too long, we do see a penalty in out-of-sample volatility by using stale data. Note, however, that this penalty is relatively modest: 8.6 percent at 500 days versus 8.3 percent at 200 days. By contrast, when the correlation half-
life becomes *too short*, there is a steep penalty in out-of-sample volatility. In this case, model estimates are harmed by too much noise in the correlation estimates.

Figure 14: Out-of-sample volatility for minimum-risk fully invested factor portfolio versus correlation half-life. The out-of-sample period consisted of 4328 trading days. The volatility was minimized near a correlation half-life of 200 days.

There are several other possible ways of selecting the correlation half-life parameter. For instance, one may study the half-life parameter that minimizes the out-of-sample volatilities of style optimized portfolios. Another way to set the correlation half-life is to investigate the value that produces the best beta forecasts. Yet another approach is to determine which correlation half-life produces the most accurate risk forecasts for factor-pair portfolios.

The results of these investigations were qualitatively similar to the results of Figure 14. Namely, the optimal correlation half-life is close to 200 days. If the correlation half-life is too long, we find a weak penalty, but observe a much larger penalty if the correlation half-life is too short. Based on these studies, we select a 200-day correlation half-life for the USE4D model.
Conclusion

We have investigated the relative accuracy of various volatility forecasts over a one-day prediction horizon. We examined several EWMA forecasts, a GARCH(1,1) model, the USE4S Model (scaled to a one-day horizon), and finally the USE4D Model. We found that the USE4D Model provided the most accurate forecasts among all models considered. Furthermore, the outperformance was consistent across factors as well as persistent across time.
References


Client Service Information is Available 24 Hours a Day

clientservice@msci.com

Notice and Disclaimer

- This document and all of the information contained in it, including without limitation all text, data, graphs, charts (collectively, the "Information") is the property of MSCI Inc. or its subsidiaries (collectively, "MSCI"), or MSCI's licensors, direct or indirect suppliers or any third party involved in making or compiling any Information (collectively, with MSCI, the "Information Providers") and is provided for informational purposes only. The Information may not be reproduced or redisseminated in whole or in part without prior written permission from MSCI.

- The Information may not be used to create derivative works or to verify or correct other data or information. For example (but without limitation), the Information may not be used to create indices, databases, risk models, analytics, software, or in connection with the issuing, offering, sponsoring, managing or marketing of any securities, portfolios, financial products or other investment vehicles utilizing or based on, linked to, tracking or otherwise derived from the Information or any other MSCI data, information, products or services.

- The user of the Information assumes the entire risk of any use it may make or permit to be made of the Information. NONE OF THE INFORMATION PROVIDERS MAKES ANY EXPRESS OR IMPLIED WARRANTIES OR REPRESENTATIONS WITH RESPECT TO THE INFORMATION (OR THE RESULTS TO BE OBTAINED BY THE USE THEREOF), AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, EACH INFORMATION PROVIDER EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF ORIGINALITY, ACCURACY, TIMELINESS, NON-INFRINGEMENT, COMPLETENESS, MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE WITH RESPECT TO ANY OF THE INFORMATION.

- Without limiting any of the foregoing and to the maximum extent permitted by applicable law, in no event shall any Information Provider have any liability regarding any of the Information for any direct, indirect, special, punitive, consequential (including lost profits) or any other damages even if notified of the possibility of such damages. The foregoing shall not exclude or limit any liability that may not by applicable law be excluded or limited, including without limitation (as applicable), any liability for death or personal injury to the extent that such injury results from the negligence or wilful default of itself, its servants, agents or sub-contractors.

- Information containing any historical information, data or analysis should not be taken as an indication or guarantee of any future performance, analysis, forecast or prediction. Past performance does not guarantee future results.

- None of the Information constitutes an offer to sell (or a solicitation of an offer to buy), any security, financial product or other investment vehicle or any trading strategy. You cannot invest in an index.

- MSCI's indirect wholly-owned subsidiary Institutional Shareholder Services, Inc. ("ISS") is a Registered Investment Adviser under the Investment Advisers Act of 1940. Except with respect to any applicable products or services from ISS (including applicable products or services from MSCI ESG Research Information, which are provided by ISS), neither MSCI nor any of its products or services recommends, endorses, approves or otherwise expresses any opinion regarding any issuer, securities, financial products or instruments or trading strategies and neither MSCI nor any of its products or services is intended to constitute investment advice or a recommendation to make (or refrain from making) any kind of investment decision and may not be relied on as such.

- The MSCI ESG Indices use ratings and other data, analysis and information from MSCI ESG Research. MSCI ESG Research is produced by ISS or its subsidiaries. Issuers mentioned or included in any MSCI ESG Research materials may be a client of MSCI, ISS, or another MSCI subsidiary, or the parent of, or affiliated with, a client of MSCI, ISS, or another MSCI subsidiary, including ISS Corporate Services, Inc., which provides tools and services to issuers. MSCI ESG Research materials, including materials utilized in any MSCI ESG Indices or other products, have not been submitted to, nor received approval from, the United States Securities and Exchange Commission or any other regulatory body.

- Any use or access to products, services or information of MSCI requires a license from MSCI. MSCI, Barra, RiskMetrics, ISS, CFRA, FEI, and other MSCI brands and product names are the trademarks, service marks, or registered trademarks or service marks of MSCI or its subsidiaries in the United States and other jurisdictions. The Global Industry Classification Standard (GICS) was developed by and is the exclusive property of MSCI and Standard & Poor’s. “Global Industry Classification Standard (GICS)” is a service mark of MSCI and Standard & Poor’s.

About MSCI

MSCI Inc. is a leading provider of investment decision support tools to investors globally, including asset managers, banks, hedge funds and pension funds. MSCI products and services include indices, portfolio risk and performance analytics, and governance tools.

The company’s flagship product offerings are: the MSCI indices with close to USD 7 trillion estimated to be benchmarked to them on a worldwide basis; Barra multi-asset class factor models, portfolio risk and performance analytics; RiskMetrics multi-asset class market and credit risk analytics; IPD real estate information, indices and analytics; MSCI ESG (environmental, social and governance) Research screening, analysis and ratings; ISS governance research and outsourced proxy voting and reporting services; FEI valuation models and risk management software for the energy and commodities markets; and CFRA forensic accounting risk research, legal/regulatory risk assessment, and due-diligence. MSCI is headquartered in New York, with research and commercial offices around the world.

1As of March 31, 2012, as published by eVestment, Lipper and Bloomberg in September 2012