Internal Models Validation in Brazil: Analysis of VaR Backtesting Methodologies

Alan Cosme Rodrigues da Silva*
Claudio Henrique da Silveira Barbedo**
Gustavo Silva Araújo***
Myrian Beatriz Eiras das Neves****

Abstract

The purpose of this paper is to analyze backtesting methodologies of VaR, focusing on aspects as suitability to volatile markets and to limited data set. We verify, from regulatory standpoint, tests to complement the Basel traffic light results, using simulated and real data. The results indicate that tests based on failures proportion are not adequate for small samples even for 1,000 observations. The Basel criterion is conservative and has low power, which does not invalidate its application, as the criterion is only one of the procedures adopted in internal models validation process. Thus, it is suggested using tests that capture the shape of returns distribution, as the Kuiper test, in addition to the Basel criterion.

***Keywords*:** Backtest; VaR tests; simulation; market risk.

JEL codes: E37; G21; D81.

Resumo

Este artigo se propõe a analisar diversas metodologias de backtest de VaR, focando em aspectos como adequação a mercados voláteis e limitação de conjunto de dados. Analisa-se, sob a ótica do órgão regulador, testes complementares ao critério proposto pelo Comitê de Basiléia, utilizando carteiras com dados reais e simulados. Os resultados indicam que os testes baseados em freqüência de falhas são inadequados para pequenas amostras e, mesmo para amostra de 1.000 observações, estes testes apresentam fraco desempenho para valores em risco de baixos percentis. Os testes que se baseiam na aderência do modelo de VaR à distribuição dos retornos apresentam melhor desempenho, mesmo em pequenas amostras. O critério quantitativo de Basiléia é conservador e apresenta baixo poder, o que não o invalida visto que sua aplicação representa um dos procedimentos a serem adotados no processo de validação de modelos internos. Sugere-se então a utilização adicional de testes que capturam a forma da distribuição dos retornos como o teste de Kuiper.

***Palavras-chave*:** backtest; valor em risco; Basiléia; VaR; simulação; risco de mercado.

Submitted in December 2004. Revised in January 2006.

*Central Bank of Brazil. Av. Presidente Vargas, 730, 7º andar – Rio de Janeiro – RJ – Brazil – P.O. 20071-001 / Phone: (5521) 3805-5083 – Fax: (5521) 3805-5092. E-mail: alan.cosme@bcb.gov.br

**Central Bank of Brazil. Av. Presidente Vargas, 730, 7º andar – Rio de Janeiro – RJ – Brazil – P.O. 20071-001 / Phone: (5521) 3805-5083 – Fax: (5521) 3805-5092.
E-mail: claudio.barbedo@bcb.gov.br

***Central Bank of Brazil. Av. Presidente Vargas, 730, 7º andar – Rio de Janeiro – RJ – Brazil – P.O. 20071-001 / Phone: (5521) 3805-5083 – Fax: (5521) 3805-5092.
E-mail: gustavo.araujo@bcb.gov.br

****Central Bank of Brazil. Av. Presidente Vargas, 730, 7º andar – Rio de Janeiro – RJ – Brazil – P.O. 20071-001 / Phone: (5521) 3805-5083 – Fax: (5521) 3805-5092.
E-mail: myrian.neves@bcb.gov.br

1. Introduction

The validation of risk models is a critical issue in the acceptance of internal models for market risk management. The models generally used by financial institutions are based on Value at Risk (VaR) concept (Riskmetrics, 1996), whose theoretical and practical framework is well spread out among risk managers. However, the process of empirically accessing risk models accuracy is still a challenge.

For regulators, the validation of internal models evolves qualitative and quantitative features. Although Lopez and Saidenberg (2001) reports the qualitative criteria as potentially of greater importance from a regulatory standpoint, he also remarks the difficulties in using statistical tests to determine the performance of institutions' VaR estimates. Basel Committee[1] proposes a validation model based on the number of exceptions observed during a period, establishing penalties according to the model's performance. The Committee recognizes that this procedure has low power in distinguishing mis-specified from accurate models, besides assuming that the exceptions are independent. However, the possibility of not rejecting mis-specified models is limited by Committee's recommendation of analyzing qualitative aspects, as regular procedures of external audit, periodical monitoring of the risk systems and evaluation the performance of risk model's forecasts.

In literature, there are three groups of methodologies for validating risk measurement models: (i) those based on the frequency of tail losses, considering or not the independence among them; (ii) those based on the size of tail losses and; (iii) the ones based on the adherence of the VaR model to the assets returns distribution. All the tests present advantages and limitations in usage: some are reliable but depends on a very large sample; others are feasible but offer partial information about the model.

The third group of tests is also used in evaluation of others risk measures besides VaR, which focus only in the frequency of extreme values, and thus disregards any loss beyond VaR level ("tail risk"). The tail risk is relevant because it is related to the insolvency of financial institutions, which is among the central issues of financial risk management. To alleviate this problem inherent in VaR, the use of expected shortfall is proposed by Artzner et al. (1997, 1999), that is the conditional expectation of loss given that the loss is beyond the VaR level. The expected shortfall can be used additionally to VaR measures by risk managers.

In this context, this article aims to evaluate the most applied tests available in literature, that are the methods proposed by Kupiec (1995), Christoffersen (1998), Crnkovic and Drachman (1996), Berkowitz (2001), Lopez (1998) and Basel Committee on Banking Supervision (1996). We focus on aspects as suitability to volatile markets such as Brazilian market and to limited data set, verifying, from regulatory standpoint, desirable tests to be used to validate internal models, in addition to Basel traffic light results.

For this, the evaluation analyzes the test performance considering type I and

[1] See Basel Committee on Banking Supervision (1996).

type II errors, using simulated data. Besides, the tests are also performed using real asset returns from Brazilian stock market and spot US dollar quoted in Brazilian real. The purpose is to apply the tests in two widely used VaR models, the historical model and delta-normal model with volatility estimated by the exponentially weighted moving average (EWMA) method.[2]

The results indicate that tests based on the frequency of failures are not adequate to small samples even for 1,000 observations sample; the tests perform weakly for low percentiles of value at risk. Basel criterion is conservative, as a proportion of failures test, what can be adequate under supervisors' position. However, as it does not take into account the returns distribution, it may not reject inaccurate models, what was confirmed by type II error evaluation. In this sense, as a complementation of Basel criterion, it could be implemented an additional test of adherence as the Kuiper test in internal models validation process.

The paper is organized as following: section 2 presents the validation tests, showing the main details discussed in literature; section 3 describes the simulation as well as the methodology to construct VaR models and validation tests; section 4 discusses the results, comparing the performance of all tests. Finally, section 5 presents the conclusion, and proposes issues for future works.

2. Tests for Validating Models

We can classify the tests for VaR models validation into three groups: (i) the ones which validate if the number of observed exceptions are consistent with the number of expected exceptions; (ii) the ones which analyze the size of the occurred exceptions and; (iii) the ones which verify the consistency between the risk model and their construction assumptions. The first group is composed by Basel Committee's procedure, the basic frequency of tail losses (or Kupiec) test and the conditional Christoffersen (1998) approach, which is based not only on the frequency of observed exceptions but also on the independence among them. The most reliable test in the second group is the one proposed by Lopez (1998). Finally, remarkable tests in the third group are suggested by Crnkovic and Drachman (1996) and Berkowitz (2001). These tests evaluate the VaR model based on the adherence of its underlying distribution to observed distribution of returns. The following section describes the six tests performed in this paper.

2.1 Tests based on frequency of tail losses

2.1.1 Basel validation

Basel Committee on Banking Supervision (1996) classifies the backtests results of VaR models into three categories. In the first (green zone), the tests results are consistent with an accurate model and the probability of not rejecting an inaccurate model is low. In the extreme (red zone), the tests results are very improbable to come from a suitable model and the probability of rejecting an accurate model

[2] For references about VaR models, see Jorion (2000).

is remote. Between the two cases lies the yellow zone, where the backtest results may be consistent with accurate or inaccurate models and the supervisor incentives the banks to present additional information about theirs models.

The Committee determines 250 days as the minimum sample size to perform the backtest of 1% daily VaR. In this case, the green zone is situated between zero and four exceptions, the yellow zone, between five and nine exceptions and, above ten exceptions, the model is in red zone. According to the backtest results, a multiplication factor – varying between three and four – is applied for capital requirements, which gives a safety margin to regulator.

If a VaR model is reliable then the exceptions do not follow any pattern, which means for instance that exceptions clustering are not expected. The Basel procedure focuses on exceptions frequency neglecting whether they follow unexpected patterns or whether their size is relevant.

2.1.2 Kupiec test

The most widely used test is the frequency of tail losses one proposed by Kupiec (1995), which is based on the frequency of losses that exceed VaR.

Let x be the number of "failures" or exceptions (the number of cases in which loss exceeds forecasted VaR) in a sample of size n. If the VaR model is correct, x follows the binomial distribution with parameter (n, p). Under the null hypothesis, the forecasting model is correct and the observed frequency of tail losses is consistent with the frequency of exceptions predicted by the model. The test is based on likelihood ratio (LR) for null hypothesis, which is given by

$$LR = -2Log\left[(1-p^*)^{n-x}(p^*)^x\right] + 2Log\left[\left(1-\left[\frac{x}{n}\right]\right)^{n-x}\left(\frac{x}{n}\right)^x\right] \quad (1)$$

where $p*$ is the probability of exceptions under null hypothesis, n is the sample size and x is the number of exceptions in the sample. Kupiec calls the LR test the proportion of failures (PF) test. Under null hypothesis, the probability of an exception (p) is equal to the significance level $(p*)$ of the VaR and PF has a chi-square distribution with one degree of freedom. The region, where the null hypotheses cannot be rejected, is determined by the intersection between the PF and the chi-square function. For given sample size and significance level, the test determines inferior and superior limits where the null hypothesis cannot be rejected at the significance level of the test.

However, as the author remarks, the test presents low power in small samples: there is a significant probability of not rejecting the null hypothesis when it is false. Like Basel criterion, this test is exclusively based on the frequency of tail losses and it also neglects potentially important information as the size of tail losses and the temporal dependence in exceptions behavior. This is an important issue given the evidences of conditional volatilities in financial series (Fierli, 2002).

2.2 Christoffersen test

An alternative approach developed by Christoffersen (1998) is to estimate a confidence interval to the number of exceptions based on the available sample and verify whether the observed number of exceptions is consistent with the forecasted, including an independence test. He suggests a procedure to evaluate the precision of predictions in confidence intervals, which tries to capture the VaR estimative conditionality.

The null hypothesis of the unconditional coverage test is that $I_t \sim i.i.d.Bernoulli(p)$, against the alternative that $I_t \sim i.i.d.Bernoulli(\pi)$, where I_t is the hit sequence of VaR$_t$ violations, p is the confidence level and π is equal to the ratio between the number of observations and the size of the sample. The test (uc) is then

$$H_{0,uc} : \pi = p \tag{2}$$

This test implicitly assumes that the hits are independent and because of this, the author tests this hypothesis against an alternative which the hit sequence follows a first order Markov sequence, with switching probability matrix

$$\Pi = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix} \tag{3}$$

where $\pi_{ij}$is the probability of an i on day $t-1$ being followed by a j on day t. The test of independence (ind) is

$$H_{0,ind} : \pi_{01} = \pi_{11}^{'} \tag{4}$$

The two tests are combined in a test of conditional coverage (cc) defined as

$$H_{0,cc} : \pi_{01} = \pi_{11} = p \tag{5}$$

Under the null hypothesis, this statistic test is distributed like a chi-square with 2 degrees of freedom.

Although Christoffersen (2003) criticizes first order Markovian process as a limited alternative compared to others forms of clustering, the present approach is easy to implement and presents the vantage of the advantage of evaluating the dynamic behaviour of exceptions sequence, apart from their frequency estimative. This allows to verify, in case of rejection of a model, if this rejection is due to the incorrect estimative of failures frequency or to the dependence among them.

3. Tests to Evaluate Forecast Model

An alternative robust evaluation involves the utilization of tests that measure the deviation between the empirical returns distribution implicit to VaR model and the model theoretical distribution. Dowd (2002) says that, once these tests are based on more information than frequency of tail losses, it is expected they produce more reliable results.

3.1 Crnkovic and Drachman test

Crnkovic and Drachman (1996) present a test for evaluating risk measurement models based on the forecast of the probability density functions (PDF). The aim is to test the agreement between the PDF forecast and the actual PDF of market variables, i.e., given a forecast of the PDF, one must determine into which percentile the actual PDF fell. The insight is that if the actual returns are random, the percentile of them must be uniformly distributed in the PDF forecast. In this way, for a forecasting method to be considered ideal, all these percentiles must look like a sample from a uniform distribution. To evaluate the independence of the measured percentiles, the authors propose the BDS's statistic, presented in Brock et al. (1996).

The uniform test of the distribution is based on the Kuiper's statistic that is equally sensitive for all values of the sample and therefore put all percentiles on an equal footing. Consider $F(x)$ a cumulative distribution function of realized percentiles, the Kuiper's statistic for the deviation between $F(x)$ and uniform cumulative distribution $G(x)$ is given by

$$K(F(x), G(x)) = \max\{f(x) - g(x)\} - \min\{f(x) - g(x)\} \tag{6}$$

The advantage of the statistic is that, since a significant fraction of the change in value of many portfolios comes from non-linear instruments, the test does not consider only the medium values of the sample, like the Kolmogorov-Smirnov test. The critical values of the test are based on Stephens (1970).

The BDS test can be applied to a series of estimated residuals to check whether the residuals are independent and identically distributed (iid). The idea behind the test is to choose a distance between a pair of points (ϵ). If the observations of the series are iid, then for any pair of points, the probability of the distance between these points being less than or equal to ϵ will be constant and equal to $c_1(\epsilon)$.

The set of pairs of points is chosen moving through consecutive observations of the sample in order, i.e., given an observation s, and an observation t of any series, one can construct a set of pairs of the form $\{[x_s, x_t]$;$[x_{s+1}, x_{t+1}]$; $[x_{s+2}, x_{t+2}]$; ...;$[x_{s+m-1}, x_{t+m-1}]$ where m is the number of consecutive points used in the set, which is called dimension. The joint probability of every pair of points in the set satisfying the ϵ condition is $c_m(\epsilon)$. According to the BDS test proceedings, under the assumption of independence, this probability will simply be the product of the individual probabilities for each pair, i.e., if the observations are independent, $c_m(\epsilon) = mxc_1(\epsilon)$. Brock et al. (1996) show that, given the dimension m and the value ϵ, the test statistic is asymptotically distributed as $N(0, 1)$.

The probability of a chosen dimension can be estimated if one goes through all the possible sets of that length that can be drawn from the sample and count the number of sets, which satisfy the condition. The ratio of the number of sets satisfying the condition divided by the total number of sets provides the estimate of the probability. Given a sample of observations of a series X, this condition is stated in mathematical notation, by the authors as:

$$c_{m,n}(\epsilon) = \frac{2}{(n-m+1)(n-m)} \sum_{a=1}^{n-m+1} \sum_{t=a+1}^{n-m+1} \prod_{j=0}^{m-1} I_\epsilon(X_{a+j}, X_{t+j}) \qquad (7)$$

However, as well documented in the literature, the BDS statistic presents a size bias. Kanzler (1999) shows that it is very sensitive to sample size and Belaire-Franch and Belaire-Franch and Bayarri-Contreras (2002) recommend that the test be used to sample higher than 2.500 data. To solve this problem, we used the function of Kanzler that evaluates the significance of BDS statistics under the null hypothesis of iidness, both on small and on large samples.[3]

3.2 Berkowitz test

Based on Crnkovic and Drachman (1996), Berkowitz (2001) developed a new way for evaluating models, stating that the information of entire forecast distributions or only the tail forecast distribution combined with ex-post realizations is enough to construct a powerful test even with sample sizes as small as 100.

Berkowitz advocates an extension of Rosenblatt (1952) transformation which delivers, under the null hypothesis, variables iid which follow $N(0,1)$. This allows for estimation of the Gaussian likelihood and construction of likelihood-based tests, which are convenient, flexible and present good finite sample properties.

The Berkowitz test applied in this work ignores model failures that are limited to the interior of the distribution and the shape of the forecasted tail of the density is compared to the observed tail. Any observations that do not fall in the tail will be intentionally truncated. Let the desired cutoff point be $VaR = \Phi^{-1}(\alpha)$, where $\Phi^{-1}(\cdot)$ is the inverse of cumulative standard normal distribution function, the new variable of interest is $z_t^* = Min(z_t, -VaR)$ to the left tail and $z_t^* = Max(z_t, VaR)$ to the right tail. For example, the log-likelihood function for the left tail is:

$$\begin{aligned} L(\mu, \sigma/z^*) &= \sum_{z_t^* < VaR} \left[-\frac{1}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(z_t^* - \mu\right)^2 \right] \\ &+ \sum_{z_t^* = VaR} \ln\left[1 - \Phi\left(\frac{VaR - \mu}{\sigma}\right)\right] \end{aligned} \qquad (8)$$

The first two terms in equation 8 represent the usual Gaussian likelihood of losses. The third term is a normalization factor arising from the truncation. Tests

[3] The function is available in http://www2.gol.com/users/kanzler/index.htm.

based on this equation should be more powerful than traditional approaches while still allowing users to ignore model failures in the interior of the distribution.

To construct an LR test, note that the null hypothesis again requires that $\mu = 0$ e $\sigma^2 = 1$ and therefore a restricted likelihood $L(0,1)$ can be evaluated and compared to an unrestricted likelihood, $L\left(\hat{\mu}, \hat{\sigma}^2\right)$. Then the tail likelihood ratio will be:

$$LR_{tail} = -2\left(L\left(0,1\right) - L\left(\hat{\mu}, \hat{\sigma}^2\right)\right) \tag{9}$$

Under the null hypothesis, this statistic test is distributed like a chi-square with 2 degrees of freedom.

3.3 Lopez test

An alternative for methodologies based on statistic criteria is the test proposed by Lopez (1999), which allows to rank models but does not give any formal statistical indication of model adequacy. Because it is not a statistical test, it does not present low power of standard tests such as the Kupiec test.

The loss function, defined by the author to any m model, follows the form:

$$mC = \frac{1}{T}\sum_{t=i}^{T} m,tC \tag{10}$$

where $C_{m,t} = \begin{cases} 1 + \left(L_t - VaR_{mt}\right)^2, \; if \quad L_t > VaR_{mt} \\ 0, \; if \quad L_t \leq VaR_{mt} \end{cases}$
and VaR_{mt} is the VaR estimated by model m for period t and L_t is the loss in portfolio market value observed in t.

This function is similar to mean squared error measure used in evaluation of the precision of points forecasts. In this case, the error magnitude only affects the function when the VaR is exceeded; in other words, this function gives a measure of the observed error size when the VaR is exceeded. The best model is the one that joins the exceptions proportion close to expected to low loss function. However, with this loss function, there is no longer a straightforward condition for the benchmark, so it can be estimated by simulation as shown ahead.

In this paper, we implement the loss function method as proposed by Lopez (1998) to real data. For estimating a benchmark for the stocks, we generate 10,000 returns series, with 1,000 observations each, following $N\left(0, \sigma^2\right)$ where the volatility is equal to the standard deviation of the Brazilian market index (IBOVESPA) returns for the first 250 observations of the sample. For US dollar series, the volatility is equal to the standard deviation of ask quotation for the first 250 days of the sample period. For these distributions, the VaR is estimated by the delta-normal methodology, with volatility estimated by the standard deviation of the 250 previous returns.

For each one of the series of returns, it is computed the loss function so we have a distribution of Cm. Once $f(Cm)$ has been generated, we can determine the empirical quantile of the cumulative distribution function, above which the regulators should take a closer look at the assumptions of VaR model. Lopez suggests a threshold quantile of 80 percent, but remarks that this decision should be based both on the severity of the distributional assumptions used and regulators preferences. In this paper, we also use 80% as a threshold.

4. Methodology

4.1 Application in simulated data

For evaluating the internal models developed by the financial institutions, as value at risk (VaR), it is necessary to verify the effectiveness of the statistic tests which have the purpose of testing these models, and select the best ones. For making this verification, asset returns simulations are used. The simulation makes possible to know precisely the process of generation of the asset returns, the DGP (data generating process), and to apply the correct VaR model, because the probability density function used in the DGP is equal to the one used in the model. Thus, the aim of the simulation is to make the test of the test, and for this purpose it is necessary to be sure that the risk model is correct or not, expecting that this fact leads to not reject or reject, respectively, the model.

Usually, the tests are evaluated in two distinct forms: the size of the sample and the power of the test. First one tries to observe which is the minimum size of the sample on which the test can be applied. This is made by counting the occurrences of type I error, which means to reject a null hypothesis when it is true. For defining the minimum size of sample, a sequence of possible sample sizes can be used. For verifying the test performance in a sample with n observations, we generate, several times (for example, 10,000 simulations), returns that follow a pre-specified DGP, and the risk measure is calculated based on this known DGP. In this way, it is known that the risk model is a model that cannot be rejected by the test and the null hypothesis is true. After this, the proportion of the times that the null hypothesis is rejected is compared with the significance level of the tests. If the level used in the test is α, we expect the null hypothesis not to be rejected more than α of the time. The minimum size for the application of the test will be the one from which the percentage of rejections is lower than the significance level of the test.

The Basel test indicates a 250 data window for verifying the precision of the VaR-models of financial institutions. Thus, this is the minimum window size evaluated in this paper. Also, we evaluated sample sizes of 500 and 1,000 observations.[4] The aim is, instead of determining the minimum sample size, to evaluate the probability of the type I error for these window sizes.

The VaR used is the delta-normal model with volatility equal to one and with four different probabilities (p), 1%, 2.5%, 5% and 10%. The delta-normal VaR, based in returns, is

[4] 1,000 observations is the minimum size recommended by the Kupiec test (1995).

$$VaR^{1d}_{i,t} = |z_p \times \sigma_{i,t}| \tag{11}$$

where $z_{p\%}$ is the quantile of the standard normal distribution that corresponds to the VaR probability and $\sigma_{i,t}$, the returns volatility of the asset i estimated for date t.

The DGP is generated with the software MATLAB 6.5 and it consists in 10,000 simulations of 250, 500 and 1,000 standard $N(0,1)$ normal returns. We estimate the 1-day VaR and then apply the tests: Christoffersen, Kupiec, Tail Berkowitz, CD, and Basel, each one with significance level of 5%, and count, for the 10,000 simulations, the number of the null hypothesis rejections.

The second way for analyzing the VaR tests is to verify their power, which is the probability to reject a null hypothesis when it is false. This probability is equal to $(1-\beta)$ where β is the probability of the occurrence of type II error, which means not to reject a null hypothesis when it is false. If the power of the test is low, then it is probable that it mis-classifies an inaccurate VaR model as well-specified.

For evaluating the power of the test it is necessary that the generation process of the asset returns (DGP) to be different of the probability distribution of the returns assumed by the model that estimates the risk measure. This way, the same simulation procedure is carried through. We generate returns with two different distributions: first, a t-Student, with variance of 6 and 1.5 degrees of freedom, that presents fatter tails than the normal standard distribution. The second distribution is generated with returns that follow a first-order auto-regressive process $(AR1)$. We choose this distribution to compare the tests performance in series of returns with auto-correlation. For both the distributions, the VaR is estimated by the delta-normal methodology, remaining the assumption that the returns are normally distributed, expecting that the tests reject their respective null hypotheses because they are false. The volatility used in the VaR model is estimated by the standard deviation of the 250 previous returns. The statistical tests are applied to each one of the series of returns and we count the number of the null hypothesis rejections. The tests with better performance are the ones that present the higher proportions of rejection of null hypotheses.

4.2 Application in real data

The statistical tests also are applied to the Brazilian market. The purpose is to apply the tests in two widely used VaR models, the historical and delta-normal with volatility estimated by the exponentially weighted moving average (EWMA) method, previously knowing, in accordance to the simulations results (subsection 3.1), how each test performs. We used long and short positions of stocks traded in the São Paulo Stock Exchange (BOVESPA) and US dollar quoted in Brazilian reais, from 02/01/1999 to 02/27/2004. The ten most traded stocks of different economy sectors are selected, as presented in Table 1. The stock quotations used are dividend-adjusted closing prices, and, for the US dollar series, the ask quotation is obtained from Brazilian Central Bank Database.

Table 1
Selected stocks of main Brazilian economic sectors

Sector	Company	Action
Food	Sadia	SDIA4
Banking	Bradesco	BBDC4
Beverages	Ambev	AMBV4
Media	Net	PLIM4
Mining	Vale do Rio Doce	VALE5
Oil and Gas	Petrobras	PETR4
Siderurgy and Metallurgy	Cia Siderúrgica Nacional	CSNA3
Telecommunications	Telemar	TNLP4
Transports	Embraer	EMBR4
Retail	Lojas Americanas	LAME4

We obtain 1,250 observations. 250 are used for estimating the model and 1,000 for backtest. We choose this sample size for backtesting VaR models because the tests of Kupiec, Christoffersen and Crnkovic and Drachman have low power for reduced samples. Then, the values of VaR of the percentiles of 1% for left tail (long position) and 99% for right tail (short position) for the holding period of one day are calculated.

The delta-normal VaR model is obtained by the same equation used in the simulations, with the difference that the volatility used here is conditional, estimated by the exponential weighting method. This way, $\sigma_{i,t}$ is substituted in the equation by $h_{i,t}$ which means the daily conditional volatility of the logarithmic returns of the asset i, estimated for date t. The equation for $h_{i,t}$ is

$$h_{i,t} = \sqrt{\lambda h_{i,t-1}^2 + (1-\lambda) r_{i,t-1}^2} \tag{12}$$

where r_t is the logarithmic return of the asset for period t and λ is the decay factor, with the constraint that $0 < \lambda < 1$. The mainly hypothesis of the model is that the asset prices are log-normal. The decay factor is estimated by the root mean squared error (RMSE) criterion, using a 250 working days window.

The historical model consists in a non-parametric model that uses one given quantile of arithmetical returns series as a VaR estimate. It uses a moving window of n days to determine the observations that compose the series of returns. The hypothesis of the model is that the distribution of probability for the studied return is the historical distribution. In this work, a 250 days moving window is used, generating 1,000 observations for backtesting.

4.3 Tests implementation

The Basel criterion indicates that the maximum number of errors in each region must be calculated with the binomial probability distribution associated to a confidence level. The green zone is characterized by the binomial cumulative distribution of the errors inferior to 95%. For example, for 1,000 observations with the confidence level of 99%, the green zone is limited by fourteen failures. The number of errors in the yellow zone starts at the point that the binomial cumulative probability equals or exceeds 95%. For 1,000 observations, for the same confidence level, fifteen failures are obtained with 95.21%, so the yellow zone starts from this number of failures and finishes at the point where the probability is inferior to 99.99% (23 failures). From this point, the red zone is initiated.

For implementing the independence test of Crnkovic and Drachman (1996), Brock et al. (1993) recommend using ϵ between 0.5 and 2 times the unconditional standard deviation of the series, and the dimensional parameter m, between 2 and 10. In this paper, the parameters of 0.433 for ϵ (1.5 times the standard deviation of the uniform distribution) and 2 for m are used, as described in Crnkovic and Drachman (1996). As the statistics of the BDS test has normal asymptotic distribution $(0, 1)$, the critical values used are from the two-tailed normal distribution. All the tests had been implemented using the software MATLAB 6.5.

5. Results

5.1 Type I error analysis – simulated data

Table 2 presents the results of all tests applied to 10,000 returns simulated according to a normal $(0, 1)$, with a delta normal VaR model constructed on a normal $(0, 1)$ hypothesis, considering 1%, 2.5%, 5% and 10% probabilities, for 250, 500 and 1,000 sample sizes. In this case, it is previously known that the VaR model is consistent to the returns distribution and it is expected the tests do not reject the VaR model. However, the tests are expected to produce rejections that represent type I error because of hypothesis tests implicit to them. In each simulation, there can be one of two answers: reject or not reject by 5% significance level. For a set of n simulations, we have a distribution of results, which can approximate as a binomial distribution. In this case, considering n=10,000, the 5% percentil is equal to 535 rejections. So, we consider satisfactory all the results below 535 rejections or 5.35% of total simulations.

Kupiec test presents results with high values of type I error for 1% VaR for all sample sizes. For 2.5% VaR, type I error is low only for 1,000 observations sample size. For others probabilities, the results are not reliable only for 250 observations.

For Christoffersen test, we verify only two situations that the test presents a rejection percentage above 5.35%. For a best test analysis, Table 2 presents the results separated in unconditional Christoffersen, which evaluates proportion of failures, and independence test. In proportion of failures test, the number of rejections is superior to the threshold for all probabilities in 250 sample size, which confirms that the test is not useful for small samples. Besides, in all sample sizes to test 1% VaR, the number of rejections is higher than expected, reflecting how critics is the size problem for extremes values. The results of independence test are similar to joint test, with two cases of superior percentage of failures. In this case, it is important to observe the good results for joint Christoffersen test are much influenced by low type I error of independence test.

Table 2
Results of Kupiec, Christoffersen, Berkowitz Tail and Crnkovic & Drachman tests applied to delta normal (0.1) VaR to simulated returns $R_t \sim N(0.1)$, where p represents the probability of the VaR model. Sample sizes of 250, 500 and 1,000 observations are evaluated and the results represent percentage of model rejections over 10,000 simulations

			Christoffersen			Tail	Crnkovic & Drachman		
p	Observations	Kupiec	Unconditional	Independence	Conjoint	Berkowitz	BDS	Kuiper	Conjoint
		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
1%	1	5.49	5.49	1.74	2.63	5.35	6.07	4.94	10.06
	500	6.96	6.96	1.52	1.74	5.79	9.18	4.96	13.59
	250	9.70	9.70	1.24	1.21	6.00	14.73	4.99	18.78
2,50%	1	4.57	4.57	1.94	4.06	5.34	6.07	4.94	10.06
	500	5.66	5.66	1.36	3.51	5.09	9.18	4.96	13.59
	250	7.50	7.50	1.41	3.25	5.60	14.73	4.99	18.78
5%	1	5.12	5.12	8.33	5.62	5.35	6.07	4.94	10.06
	500	5.17	5.17	3.19	3.79	4.88	9.18	4.96	13.59
	250	5.79	5.79	1.86	3.98	5.20	14.73	4.99	18.78
10%	1	4.52	5.07	5.26	5.21	5.01	6.07	4.94	10.06
	500	5.10	5.10	5.33	4.90	4.78	9.18	4.96	13.59
	250	5.58	5.58	7.90	5.50	4.81	14.73	4.99	18.78

The tail Berkowitz test presents three occurrences above the threshold for 1% and 2.5% VaR models, which indicates this test is difficult to be applied in tail extreme values for few observations distributions.

Crnkovic & Drachman (CD) test represents Kuiper and BDS conjoint analysis. The test does not present accurate results as the number of rejections is above the threshold for all sample sizes. To a better analysis, we separate the test results into Kuiper and BDS. The results of the Kuiper test are satisfactory for all the sample sizes and significance levels, with numbers of rejections lower than the threshold. The results of the independence test show that, for small samples, the test presents low power, even using the auxiliary function to evaluate the significance level, and the test tends to produce better results with big samples.

Table 3 presents the simulations classified according to the Basel criterion. This criterion considers a maximum number of failures to which there is a cumulative probability of 95% following a binomial distribution whose success probability is equal to the VaR significance level. In this way, it is expected 95% of simulations to fall into green zone. However, the results show that simulations percentage classified into green zone is lower than expected, exceeding the expected percentage in yellow zone. Nevertheless, a Basel criterion is conservative as, for supervisors, the question is rejecting models with excessive exceptions no matter if reliable models are also rejected.

Table 3
Results of Basel criterion applied to the VaR delta-normal (0.1) for simulated returns $R_t \sim N(0.1)$, where p represents the probability of the VaR model. Sample sizes of 250, 500 and 1,000 observations are evaluated and the results represent the percentage of model classification in each zone, considering 10,000 simulations

p	Observations	Green zone (%)	Yellow zone (%)	red zone (%)
1%	1	91.90	8.09	0.01
	500	93.51	6.47	0.02
	250	88.94	11.02	0.04
2,50%	1	92.81	7.19	0.00
	500	91.62	8.38	0.00
	250	94.71	5.28	0.01
5%	1	94.55	5.42	0.03
	500	93.31	6.69	0.00
	250	92.18	7.81	0.01
10%	1	94.61	5.36	0.03
	500	93.85	6.14	0.01
	250	93.75	6.25	0.00

5.2 Power test analysis – simulated data

Tables 4 and 5 evaluate the power of the tests for a returns generation process following a t Student distribution and an AR process respectively. The tests present higher power as a higher rejection percentage indicates better capacity to detect a false null hypothesis.

In Table 4, tail Berkowitz and Kuiper tests presents best test power results for all sample sizes and confidence levels, reaching higher than 99% rejection percentage.

We verify that the power of failure proportion test and Christoffersen independence test present low values. This demonstrates that when the returns generation process comes close to VaR model underlying distribution, these tests are not capable to make a reliable discrimination.

The results for AR1 simulated returns, presented at Table 5, are satisfactory, as most tests present rejection rate higher than 90%. This happens because data generation process is very different from VaR model underlying distribution. For Christoffersen independence test, low power can be justified by the limitation of Markovian process to adequately model returns clustering.

For conjoint Christoffersen test, instead of what happens to type I error, now the results are influenced by unconditional Christoffersen test what delivers high power tests as results.

Table 4
Results of Kupiec, Christoffersen, Berkowitz Tail and Crnkovic & Drachman tests applied to delta normal (0.1) VaR for simulated returns $R_t \sim t(\nu = 6, \sigma = 1.5)$, where p represents the probability of the VaR model. Sample sizes of 250, 500 and 1,000 observations are evaluated and the results represent percentage of model's rejections over 10,000 simulations

			Christoffersen			Tail	Crnkovic & Drachman		
p	Observations	Kupiec	Unconditional	Independence	Conjoint	Berkowitz	BDS	Kuiper	Conjoint
		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
1%	1	35.73	35.73	1.92	21.06	100.00	14.05	100.00	100.00
	500	22.88	22.88	1.28	8.53	100.00	22.57	100.00	100.00
	250	9.71	9.71	1.83	5.57	99.34	31.54	99.12	98.80
2,50%	1	5.00	5.00	1.99	4.64	100.00	14.05	100.00	100.00
	500	6.08	6.08	1.67	3.73	100.00	22.57	100.00	100.00
	250	6.62	6.62	1.40	3.36	99.63	31.54	99.12	98.80
5%	1	6.24	6.24	8.39	5.66	100.00	14.05	100.00	100.00
	500	6.18	6.18	2.60	3.76	100.00	22.57	100.00	100.00
	250	6.34	6.34	1.54	4.21	99.72	31.54	99.12	98.80
10%	1	32.71	32.71	5.46	26.19	100.00	14.05	100.00	100.00
	500	19.90	19.90	6.44	14.89	100.00	22.57	100.00	100.00
	250	13.81	13.81	6.70	11.36	99.74	31.54	99.12	98.80

Table 5
Results of Kupiec, Christoffersen, Berkowitz Tail and Crnkovic & Drachman tests applied to delta normal (0.1) VaR for simulated returns $R_t = R_{t-1} + \epsilon_t R_t$ where $\epsilon_t \sim N(0, 1)$, where p represents the probability of the VaR model. Sample sizes of 250, 500 and 1,000 observations are evaluated and the results represent percentage of model's rejections over 10,000 simulations

			Christoffersen			Tail	Crnkovic & Drachman		
p	Observations	Kupiec	Unconditional	Independence	Conjoint	Berkowitz	BDS	Kuiper	Conjoint
		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
1%	1	96.97	96.97	62.37	99.97	99.42	92.21	100.00	100.00
	500	96.38	96.38	51.60	98.69	98.03	89.59	100.00	100.00
	250	96.03	96.03	40.96	47.41	67.20	86.19	100.00	100.00
2,50%	1	96.82	96.82	64.00	100.00	99.96	92.21	100.00	100.00
	500	96.48	96.48	53.02	100.00	99.49	89.59	100.00	100.00
	250	95.97	95.97	42.27	99.41	97.88	86.19	100.00	100.00
5%	1	96.93	96.93	65.13	100.00	100.00	92.21	100.00	100.00
	500	96.39	96.39	54.33	100.00	99.96	89.59	100.00	100.00
	250	95.69	95.69	43.36	99.99	99.07	86.19	100.00	100.00
10%	1	97.08	97.16	66.23	100.00	100.00	92.21	100.00	100.00
	500	96.39	96.39	55.48	100.00	100.00	89.59	100.00	100.00
	250	95.70	95.70	44.16	100.00	99.84	86.19	100.00	100.00

Table 6 presents the Basel criterion results for the data generated by the t Student distribution. We observe a relevant number of simulations classified at green zone. This data generation process presents fatter tails than the normal distribution underlying to model's VaR as more extreme the percentile is. This explains the reason why, for 1% and 2.5% VaR, there are more simulations at yellow and red zone, while for 5% and 10%, the number of failures is close to the number expected according to Basel criterion, increasing the number at green zone. For results presented at Table 6 related to AR1 returns, there are less simulations at green zone, although the percentage is still high (above 37%), which occurs as the criterion only takes care of the exception's number, apart from the difference between forecasted and observed returns distributions.

Table 6
Results of Basel criterion applied to a VaR delta-normal for simulated returns where $R_t \sim t(\nu = 6, \sigma = 1.5)$ and $R_t = R_{t-1} + \epsilon_t$ where $\epsilon_t \sim N(0,1)$, where p represents the probability of the VaR model. Sample sizes of 250, 500 and 1,000 observations are evaluated and the results represent the percentage of model classification in each zone, considering 10,000 simulations

		$t(6)$ Student			AR(1)		
p	Observations	Green zone (%)	Yellow zone (%)	Red zone (%)	Green zone (%)	Yellow zone (%)	Red zone (%)
1%	1,000	41.80	57.13	1.07	37.76	1.59	60.66
	500	64.12	35.26	0.63	46.23	1.61	52.16
	250	66.82	32.86	0.32	53.68	2.12	44.20
2.5%	1,000	85.29	14.71	0.01	37.48	1.49	61.04
	500	85.87	14.12	0.01	44.73	1.80	53.47
	250	92.16	7.82	0.02	51.90	1.77	46.34
5%	1,000	99.06	0.94	0.00	38.07	1.31	60.63
	500	97.51	2.49	0.00	44.26	1.99	53.76
	250	95.60	4.41	0.00	50.17	2.24	47.60
10%	1,000	99.98	0.02	0.00	39.29	1.58	59.13
	500	99.74	0.26	0.00	44.51	2.09	53.40
	250	99.15	0.85	0.00	49.46	2.27	48.27

Finally, we can verify the importance of tests that compare observed and forecasted distributions to complement Basel criterion in risk models validation.

5.3 Results – real data

First of all, the normality of stocks and dollar returns is tested based on Jacque-Bera statistic. We intend to verify, by kurtosis und asymmetry measures, financial series stylized facts suggested by literature and their importance to the tests. The normality hypothesis is rejected to all studied assets and there is kurtosis excess related to normal distribution.

Table 7
Descriptive statistic of selected stocks for VaR modelling, from 02/01/1999 to 02/27/2004

	AMBV4	BBCD4	CSNA3	EMBR4	LAME4	PETR4	PLIM4	SDIA4	TNLP4	VALE5	DOLAR
Mean	0.0013	0.0005	0.0022	0.0013	0.0015	0.0008	-0.0042	0.0014	0.0001	0.0013	0.0005
Standard-deviation	0.0229	0.0254	0.0284	0.0297	0.0322	0.0223	0.0527	0.0239	0.0270	0.0202	0.0106
Asymmetry	0.0151	-0.0532	0.1194	-0.5467	0.4738	0.0112	0.1160	0.2263	0.0515	0.1260	-0.7300
Kurtosis	56.131	42.252	41.177	74.956	60.023	46.118	59.221	48.314	38.753	42.180	119.765
Jarque-Bera statistic	284.56	63.02	54.43	891.90	413.00	108.26	358.03	148.28	32.36	64.46	3446.23
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 8 presents results to 1% and 99% VaR. As unconditional Christoffersen test is similar to Kupiec's test, presenting the same results, it is shown only Kupiec and conjoint Christoffersen in table.

Because of asymmetry in distributions, the tests results are different for short and long positions. VaR calculated from historical simulation performs well at Kupiec test, especially for short position, which indicates approval of the model. For the delta-normal model, Kupiec test reject almost all assets, which is expected, as the returns series are not normal. These results must be seen with caution as, for this sample size, the test presents, in simulations, type I error higher than the accepted limit.

When the independence test is added to the proportion of failures test, the historical VaR is rejected for some series; the independence test captures clusters of exceptions that the tests purely based in frequency of failures do not capture. This occurs, for example, to BBDC4 and TNLP4 series. On the other hand, for the delta-normal model, we observe that the conjoint test does not reject models rejected only by the proportion of failures test (CSNA3 and PLIM4).

The Basel criterion classifies more models into the green zone, because it is more conservative than the proportion of failures tests. However, due to the low power of the test, when it classifies the models in the green zone, these results are not reliable, what can be evidenced by the adherence tests. For example, the delta-normal models for the series of LAME4, PLIM4 and PETR4 are classified into the green zone and are rejected by higher power tests as the Berkowitz and Kuiper.

For the Crnkovic and Drachman test, it is observed that there are rejections for most of the assets, what occurs due to the BDS test performance. However, the Kuiper test presents a low type I error and a high power, what can be verified in the simulations. So its results, presented in Table 8, can be used to select the best models.

As Kuiper test is based on the adherence of the model to the distribution of returns, it is verified that this test rejects the delta-normal model for almost all the assets, due to the kurtosis and asymmetry presence in the financial series. For the historical model, the test does not reject the model for almost all the series, what indicates that the underlying hypothesis to the model is well adjusted.

For the Berkowitz test, the results also indicate that the historical model is better adjusted than the delta-normal one. In this case, the results are identical to the Kupiec test, not rejecting the majority of the models. However, considering that this test only evaluates the tail of the distribution and has a high power, its results are trustworthier than the Kupiec test. For the delta-normal model, the Berkowitz test rejects all series for the short position, except to TNLP4. For long position, the test added more information than Kuiper test. This occurs because these tests are complementaries: while the last one is based on the forecast of all the distribution of returns, the Berkowitz test is based only on the tail of the distribution.

In general, we verify more rejections for the delta-normal model, because the underlying distribution of this model is normal and the returns of the assets do not present any sign of normality as shown in Table 7. This fact does not allow the adherence of the theoretical distribution to the empirical one.

Table 8
Results of Kupiec, Christoffersen, Berkowitz Tail and Crnkovic & Drachman tests applied to delta normal and historical simulation VaR for the selected assets, from 02/01/1999 to 02/27/2004, for 1% and 99% quantiles, where R and NR represents reject and not reject the model respectively

p = 1%	Historical Simulation							Delta-Normal						
		Chris	Tail						Chris	Tail				
Asset	Kupiec	Conj.	Berkowitz	Kuiper	BDS	CD	Basel	Kupiec	Conj.	Berkowitz	Kuiper	BDS	CD	Basel
AMBV4	NR	NR	NR	NR	R	R	Green	NR	NR	R	R	R	R	Yellow
BBDC4	NR	R	NR	NR	R	R	Green	R	R	R	NR	NR	NR	Yellow
CSNA3	NR	NR	NR	NR	R	R	Yellow	NR	NR	NR	NR	R	R	Yellow
EMBR4	NR	NR	NR	NR	R	R	Green	R	R	R	R	R	R	Yellow
LAME4	NR	NR	NR	R	R	R	Yellow	NR	NR	NR	R	R	R	Green
PETR4	NR	NR	NR	NR	NR	R	Green	R	R	R	R	NR	R	Yellow
PLIM4	R	R	R	R	R	R	Yellow	NR	NR	R	R	NR	R	Green
SDIA4	NR	NR	NR	R	R	R	Yellow	NR	NR	R	R	R	R	Green
TNLP4	NR	R	NR	NR	NR	NR	Green	NR	NR	NR	NR	NR	NR	Yellow
VALE5	NR	NR	NR	NR	R	R	Green	NR	NR	R	NR	R	R	Green
U.S. dollar	NR	R	R	NR	NR	R	Green	R	R	R	R	R	R	Yellow
p = 99%	Historical Simulation							Delta-Normal						
		Chris	Tail						Chris	Tail				
Asset	Kupiec	Conj.	Berkowitz	Kuiper	BDS	CD	Basel	Kupiec	Conj.	Berkowitz	Kuiper	BDS	CD	Basel
AMBV4	NR	NR	NR	NR	R	R	Green	R	R	R	R	R	R	Yellow
BBDC4	NR	NR	NR	NR	R	R	Green	NR	NR	R	NR	NR	NR	Yellow
CSNA3	NR	NR	NR	NR	R	R	Yellow	R	NR	R	NR	R	R	Yellow
EMBR4	NR	NR	NR	NR	R	R	Green	R	R	R	R	R	R	Yellow
LAME4	NR	R	NR	R	R	R	Green	R	R	R	R	R	R	Red
PETR4	NR	NR	NR	NR	NR	R	Green	NR	NR	R	R	NR	R	Green
PLIM4	NR	NR	NR	R	NR	R	Green	R	NR	R	R	NR	R	Yellow
SDIA4	NR	R	NR	R	R	R	Green	R	R	R	R	R	R	Yellow
TNLP4	NR	NR	NR	NR	NR	NR	Green	NR	NR	NR	NR	NR	NR	Green
VALE5	NR	NR	NR	NR	R	R	Green	R	R	R	NR	R	R	Yellow
U.S. dollar	NR	R	NR	NR	R	R	Yellow	R	R	R	R	R	R	Yellow

We implement the loss function method as proposed by Lopez, estimating a threshold of 1.0003 for the stocks and 1.0001 for the US dollar. This threshold is the maximum value allowed of the observed error when the VaR is exceeded. We observe that the delta-normal model presents a better performance for the 2.5%, 5%, 95% and 97.5% percentiles, because this model is more sensitive to volatility changes. However, for the 1% and 99% percentiles, it is not possible to define which model is the best one as the historical model presents good performance for extreme values.

Table 9
Results of Lopez test applied to a delta-normal and Historical VaR for each asset, where quantile represents the probability of the VaR model. The Models are evaluated according to the Benchmark of the Test, estimated by 10,000 simulations

		Quantile					
Model	Asset	1%	2.50%	5%	95%	97.5%	99%
Historical	AMBV4	R	R	R	R	R	R
Delta-Normal	AMBV4	R	R	R	R	R	NR
Historical	BBDC4	R	R	R	R	NR	NR
Delta-Normal	BBDC4	R	R	R	NR	NR	NR
Historical	CSNA3	NR	R	R	R	R	R
Delta-Normal	CSNA3	NR	NR	R	R	R	R
Historical	EMBR4	R	R	R	R	R	R
Delta-Normal	EMBR4	R	R	R	R	R	R
Historical	LAME4	R	R	R	R	R	R
Delta-Normal	LAME4	R	R	R	R	R	R
Historical	PETR4	NR	R	R	R	R	R
Delta-Normal	PETR4	R	R	R	NR	NR	NR
Historical	PLIM4	R	R	R	R	R	R
Delta-Normal	PLIM4	R	R	R	R	R	R
Historical	SDIA4	NR	R	R	R	R	R
Delta-Normal	SDIA4	NR	R	R	R	R	R
Historical	TNLP4	NR	NR	R	R	R	R
Delta-Normal	TNLP4	NR	NR	NR	NR	NR	R
Historical	VALE5	NR	NR	NR	NR	NR	NR
Delta-Normal	VALE5	R	NR	NR	NR	NR	NR
Historical	U.S. Dollar	R	R	R	R	R	NR
Delta-Normal	U.S. Dollar	NR	NR	NR	NR	NR	NR

6. Is the Basel Criterion Adequate?

From practical standpoint, the Basel criterion focuses exclusively on the number of failures of the model. Thus, it is possible imagining a model in which the underlying assumptions do not adhere to the empirical return distributions and even though the model is not rejected. For evaluating the performance of Basel criterion, we execute a simulation exercise where the manager, although with an aggressive style, keeps the model's performance into the Basel limits.

In this way, we generate 1,000 uniformly distributed random returns ranging from –50% to 50%. The VaR model is constructed based on the assumption that the returns follow a stochastic continuous-time motion with a random jump process, where

$$dS_t = a_t dt + \sigma_t dW_t + dJ_t \tag{13}$$

for $t \geq 0$, where $\sigma = 0,616$, $a = 0,000635$, dW_t is a standard Wiener process and dJt represents the random process that follows a Poisson distribution. The jump frequency parameter and the jump magnitude are randomly generated. The percentile VaR is applied to the simulated returns series and it is expected no more than four exceptions in each 250 days, so that, the model should be classified into Basel green zone.

In order to evaluate if the Basel criterion is sufficient, we use the Kuiper test that is based on the entire distribution. From the cumulative returns distribution and the cumulative distribution of the VaR model, we calculate the Kuiper statistic. The process is repeated 1,000 times and we verify the distance of Kuiper measure for each return distribution versus VaR distribution pair. The measure mean for the return series that were classified as green zone by Basel criterion is presented on Table 10.

Table 10
Percentage of classification into Basel green zone of an inaccurate VaR model for each 250 days and the mean of Kuiper statistic, considering 1,000 simulations

Basel Green Zone	1st period	95.6%
	2nd period	96.7%
	3rd period	96.6%
	4th period	94.9%
Kuiper	Mean of Statistic	0.8903

Considering that the critical value of the Kuiper measure for 1,000 observations interval is 0.054971, which means that this is the maximum value for a model to be deemed "accurate", we can verify that the constructed VaR model follows the Basel proposal in terms of number of failures but it is a model in which the underlying distribution does not adhere the simulated returns distribution. So, tests that evaluate the models underlying assumptions could be used in addition to the Basel diagnosis in risk models validation, at least in high volatile markets, whose returns can change dramatically.

7. Conclusion

The aim of this paper is to analyze tests for evaluating the accuracy of risk models. The studied tests are the ones proposed by Kupiec (1995), Christoffersen (1998), Crnkovic and Drachman (1996), Berkowitz (2001), Lopez (1998) and the Basel Committee. We focus on aspects as suitability to volatile markets such as Brazilian market and to limited data, verifying, from regulatory standpoint, desirable tests to be used to validate internal models. We analyze the performance of the tests based on the type I error and type II error. For this purpose, series of returns for three distinct windows are simulated (250, 500 and 1,000 observations) using normal standard, t-student and first-order autoregressive distributions. In addition, the tests also are applied to the historical and delta-normal 1%-VaR on long and short positions of ten stocks traded in the São Paulo Stock Exchange (BOVESPA) and US dollar quoted in Brazilian real, from 02/01/1999 to 02/27/2004.

The results of the simulations indicate that the proportions of failures tests (Kupiec and unconditional Christoffersen) are not appropriate for small samples and, even for sample of 1,000 observations, these tests present weak performance for values at risk with low percentiles. When we include the independence test, we notice that the conditional test of Christoffersen presents better performance both in relation to the size and to the power of the test. The results of the conditional test suggest that the independence test influences the size while the proportion of failures influences the power of the test.

In the case of the Crnkovic and Drachman test, we verify that the test is not adequate for the sample sizes used in this paper because the BDS test is only appropriated to samples with a higher number of observations. However, the Kuiper test is adequate to capture the shape of the distribution of returns, even for small samples. The tail Berkowitz test presents good performance for both size and power of the test, except for the sample of 250 observations on low percentiles VaR. This happens because of the difficulty for modeling the tail shape of distributions with few observations.

The Basel criterion, as a proportion of failures test, is conservative, what can be desirable in supervisor's point of view. However, by not taking into account the shape of the distribution of returns, it has low power in distinguiching mis-specified from accurate models. It is important to remind that in the internal model validation process, the statistic procedure analyzed here represents one of the issues that have to be considered by banking supervision. The risk of not rejecting mis-specified models using Basel criterion is diminished by recommendation of analyzing qualitative aspects, as regular procedures of external audit, as well the evaluation of any modification in the risk measurement process..

In case of banking supervision desires to evaluate directly the statistics behind internal models, it is suggested, as a complement to Basel criterion, the use of a test of adherence, as the Kuiper test, especially in high volatile markets.

References

Artzner, P., Delbaen, F., Eber, J., & Heath, D. (1997). Thinking coherently. *Risk*, 10(11).

Artzner, P., Delbaen, F., Eber, J., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

Basel Committee on Banking Supervision (1996). Supervisory framework for the use of backtesting in conjunction with the internal models approach to market risk capital requirements. Basel Committee on Banking Supervision, Basel (January).

Belaire-Franch, J. & Bayarri-Contreras, D. (2002). The BDS test: A practioner's guide. In http://aeser.anaeco.uv.es/pdf/dt/dt02-01.pdf . Accessed on August 2004.

Berkowitz, J. (2001). Testing density forecasts with applications to risk management. *Journal of Business & Economic Statistics*, 19(4).

Brock, W., Dechert, D., Sheinkman, J., & Lebaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, 15(3):197–235.

Brock, W., Hsieh, D., & Lebaron, B. (1993). *Non-Linear Dynamics, Chaos and Instability*. Massachusetts Institute of Technology Press, Cambridge, Massachusetts.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39:841–862.

Christoffersen, P. F. (2003). Backtesting value-at-risk: A duration-based approach. In http://ssrn.com/abstract=418762. Accessed in August 2004.

Crnkovic, C. & Drachman, J. (1996). Quality control. *Risk*, 9(9):139–143.

Dowd, K. (2002). *Measuring Market Risk*. John Wyley & Sons, England.

Fierli, F. (2002). Applying and testing VaR estimation methods for non-linear portfolios. University of Southern Switzerland. Working Paper.

Jorion, P. (2000). *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, University of California at Irvine.

Kanzler, L. (1999). Very fast and correctly sized estimation of the BDS statistic. Oxford University, Department of Economics, Working Paper. Available on http://www2.gol.com/users/kanzler accessed in August 2004.

Kupiec, P. (1995). Techniques for berifying the accuracy of risk measurement models. *Journal of Derivatives*, 2:73–84.

Lopez, J. A. (1998). Methods for evaluating value-at-risk estimates. Research and Market Analysis Group, Federal Reserve Bank of New York. mimeo.

Lopez, J. A. (1999). Regulatory evaluation of value-at-risk models. *Journal of Risk*, 1:37–64.

Lopez, J. A. & Saidenberg, M. R. (2001). The development of internal models approaches to bank regulation & supervision: Lessons from the market risk amendment. In: http://www.frbsf.org. Accessed on August 2004.

Riskmetrics (1996). *Technical Document*. J. P. Morgan, 4th edition.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. Annals of Mathematical Statistics, 23.

Stephens, M. (1970). Use of Kolmogorov-Smirnov, cramer-von mises and related statistics without extensive tables. *Journal of the Royal Statistical Society*, 32.