University of Piraeus
Graduate Department of Banking and Financial management

Petroleum products, petroleum futures and Value at Risk

Christos G. Pliakouras
ph: mxrh/0423

Supervisor: George Skiadopoulos, Ph.D

Piraeus, June 2006

TABLE OF CONTENTS

Chapters	Pages
Abstract	3
1. Introduction	4
2. Value at Risk (VaR).	7
1.1 Variance-Covariance method	8
1.2 Non parametric method	12
1.3 Semi parametric methods	14
3. Backtesting VaR	18
4. Conditional VaR (CVaR)	23
5. Backtesting CVaR	24
6. Data Set	26
6.1 Spot Market	26
6.2 Nearby futures	29
7. Implementation	34
7.1 Calculating and backtesting VaR	34
7.2 Calculating and backtesting ES	36
8. Results for spot market	37
8.1 VaR Results	37
8.2 Remarks on VaR results	48
8.3 CVaR Results	51
8.4 Remarks on CVaR results	53
8.5 Remarks on both evaluation stages	53
9. Results for nearby futures	54
9.1 VaR Results	55
9.2 Remarks on VaR results	63
9.3 CVaR Results	66

9.4 Remarks on CVaR results 68
9.5 Remarks on both evaluation stages 68
10. Conclusion 69
References 71
Appendix A
Appendix B
Appendix C

Abstract

We computed one-day VaR for the petroleum products and their nearby futures at 99% and 95% confidence levels using parametric, semi-parametric and non-parametric approaches. For each VaR calculation we computed the corresponding Expected Shortfall (ES). For backtesting VaR we used Christoffersen's conditional coverage test and for backtesting the ES we used a quantratic score function based on a loss function. In order to decide on the best method for VaR computation, we followed two approaches and compared them to check whether they yield similar results. In the first approach, we selected as the best method the one that passes the conditional coverage test and yields the lowest average VaR value. In the second approach we used the score function based on the loss function. The method that gets accepted from the conditional coverage test and yields the lowest score function value is considered as the best. We found that the two approaches yield similar results in the case of spot market data but not for the nearby futures. In addition, we found that the parametric approaches underachieved in the 99% confidence level, whereas the non-parametric and semi-parametric methods performed quite well in that level. Quite the opposite phenomenon appeared at the 95% confidence level. Finally we found that although no single method can be characterized as the best for all our data sets, the historical simulation (HS) method with 100 days rolling window overachieved in most of the cases.

1 Introduction

Over the past twenty years oil has become the biggest commodity market in the world and it has evolved from a primarily physical activity into a sophisticated financial market with trading horizons now extending to over ten years forward. In the process it has attracted a wide range of participants, which now include investment banks, asset managers for mutual funds, pension funds, insurance companies, hedge funds as well as the traditional oil majors (BP etc) and the physical oil traders.

The agreements that OPEC members reached in the seventies and the reaction that those agreements caused in the oil market agents changed the controlled environment that had characterized the oil market up to then. The new market is relatively free and characterized by high price shifts. An unpredictable, volatile and risky environment has arisen and protection against market risk has become an essential issue.

Oil markets are also characterized by the widespread use of futures and forward contracts. Indeed, derivative contracts such as futures and options are a way to hedge risk for the buyers and sellers of such products. Since their introduction, futures on petroleum products are gaining in importance because they have been designed to serve the oil industry's needs. Petroleum futures are traded across a wide range of maturities: different maturities may be used for different purposes by investors. Furthermore, the petroleum term structure of futures prices evolves stochastically over time. It is typically characterized by high volatility. This attracts speculators and makes the hedging of these contracts a challenging task.

The volatile oil price environment requires risk quantification. Within oil markets, Value at Risk (VaR) and Expected Shortfall (ES) can be used to quantify the maximum price change associated with a likelihood level. This quantification is fundamental when designing risk management strategies. So, our purpose is to implement a number of different approaches in order to determine the best, if any, method of calculating VaR for petroleum products and petroleum futures.

Not many studies have focused on price risk quantification through VaR for petroleum products. The first of them was by P. Giot and S. Laurent (2002). In this study they used daily spot prices for WTI and Brent crude oil from 20/5/87-18/3/2002. For that

data set they computed 95% and 99%VaR with the RiskMetrics™, the skewed student ARCH, and the skewed student APARCH methods. Kupiec's technique was used for backtesting purposes. Their results showed that the best method was the student APARCH, while the RiskMetrics methodology was proven for both data sets adequate for the 95% confidence level, but not for the 99% level of confidence.

The second study was by D.Cabedo and I.Moya. In that study they used daily spot prices for Brent crude oil for the period 01/01/92-31/12/99. They computed 99%VaR using three methods, the GARCH(1,1), the historical simulation and the historical simulation with ARMA forecasts (HSAF). The HSAF method does not use directly the distributions of past returns as HS does, but rather the distribution of forecasting errors, derived from an estimated ARMA model. For backtesting purposes, Kupiec's method was implemented in the time period of 1999. Their results showed that the HSAF method outperformed the HS one and that the GARCH(1,1) approach underestimated VaR.

The third and more sophisticated study was by T.Krehbiel and L.C.Adkins (2003). They used as data for 99%VaR calculation all the petroleum products and natural gas from the NYMEX energy complex (spot and nearby futures) each one for different time horizon. As a backtesting measure they used Kupiec's. They used the EWMA, the AR(1)GARCH(1,1) with normal innovations and the AR(1)GARCH(1,1) with GPD innovations. Their results showed that in the spot prices case only the EVT method was not rejected and that only for WTI, Brent crude and unleaded gasoline. In the nearby futures case all the methods were rejected apart from WTI, heating oil and gasoline, where the EVT method got accepted.

In this dissertation we hope to go one step further from the previous studies. Specifically, as our data set we will use all the petroleum products i.e. WTI crude oil, Brent crude oil, unleaded gasoline and heating oil, along with natural gas and their futures with the shortest maturity. We will use a much wider observation period, both for parameter estimation and backtesting purposes and the same observation period will be used for all data sets so as to make the results comparable. For VaR calculation methods we analyze three parametric, one non-parametric and two semi-parametric. Specifically, we implement the following methods: equally weighted moving average, exponentially weighted moving average, GARCH variance with normal innovations, historical

simulation, filtered historical simulation and the extreme value theory. For each of these methods an ES measure will be computed. In our attempt to check the forecasting power of each VaR method we employed Christoffersen's conditional coverage technique. In contrast to this, earlier research focused mainly on the unconditional coverage of the models. Finally we followed two approaches to guide on VaR model selection process. In the first of these approaches we chose a method as the best when it successfully passed the conditional coverage test and yielded the lowest average VaR value among the other accepted methods. In the second approach we used a score function based on a loss function and selected as the best method the one that yielded the lowest score function value. Following this procedure, we could select a risk model that predicts the VaR number accurately and minimizes, if a VaR violation occurs, the difference between the realized and the expected losses.

So, although our study has the same purpose as other similar studies –to find the best model for VaR computation - the results may differ due to the strictness of our backtesting methods. We should note here that the more difficult a test is –in terms of not yielding easily acceptances- the more robust the results are.

2 Value at Risk

Value at Risk (VaR) measures the potential loss on a portfolio that would result if relatively large adverse price movements were to occur. To quantify potential loss two underlying parameters must be specified: the holding period under consideration (m) and the desired statistical confidence interval (n). The holding period refers to the time frame over which price changes in portfolio value are measured and it is usually calculated in days. It is assumed that the portfolio is held constant over the holding period. The confidence level defines the proportion of trading losses that are covered by the VaR amount. Thus VaR is the loss of our portfolio, which is to be exceeded only $n*100\%$ of the time in the next m trading days. For a more formal definition of VaR:

$$\Pr(r_t < -VaR_t) = n \tag{2.1}$$

This is the definition we are going to use throughout this dissertation.

VaR has a number of significant attractions over other traditional risk measures. Firstly, it enables us to aggregate the risks of sub-positions into an overall measure of portfolio risk. Secondly, it can be applied to any type of portfolio and it enables us to compare the risks of different portfolios. Thirdly, it takes account of all driving risk factors. Fourthly, it is probabilistic and gives a risk manager useful information on the probabilities associated with specified loss amounts. And finally, it is simple. It is expressed in the simplest and most understood unit of measure, the' lost money'. VaR can be implemented in a number of ways. In the next paragraphs we explain most of them.

2.1 Variance Covariance approach

This approach assumes normality and serial independence for the price changes and the absence of non-linear positions such as options. The dual assumption of normality and serial independence is useful for two reasons. First, normality simplifies VaR calculations because all percentiles are assumed to be known multiples of the standard deviation. Hence, the VaR calculation requires only an estimate of the standard deviation of the portfolio's change in value over the holding period. Second, serial independence means that the size of a price move on one day will not affect estimates of price moves on any other day. Thus, longer horizon standard deviations can be obtained by multiplying daily horizon standard deviations by the square root of the number of days in the longer horizon. So, when these assumptions are made together we can use a single calculation of the portfolio's daily horizon standard deviation to develop VaR measures for any given holding period and any given percentile:

$$VaR^{p}_{\Delta t} = -\sigma_{\Delta t}\Phi^{-1}_{n} \qquad (2.2)$$

Where Φ is the cumulative density function of the standard Normal distribution and n is the confidence level.

Therefore, the problem comes down to estimating standard deviations of portfolio's price changes for a holding period of one day. This can be achieved by several approaches. In this dissertation we are going to use the following models:

Equally Weighted Moving Average (MA), Exponentially Weighted Moving Average (EWMA), GARCH (q, s) variance model.

Equally Weighted Moving Average (MA)

This approach calculates a given portfolio's standard deviation using a fixed amount of historical data. The major difference among MA approaches is the time frame of the fixed amount of data. The calculation of portfolio standard deviations using an equally weighted moving average approach is given by:

$$\sigma_{t+1} = \sqrt{\frac{1}{(m-1)}\sum_{k=1}^{m}(r_{t+1-k} - \mu)^2} \ . \qquad (2.3)$$

Where, σ_t denotes the estimated standard deviation of the portfolio at the beginning of day t. The parameter m specifies the number of days included in the moving average (the observation period), r_t the change in portfolio value in day t and μ the mean change in portfolio value, which is assumed to be zero.

This approach though has its shortcomings: the normality assumption for the price changes is not supported by practice. Price changes seem to follow in reality a distribution that has heavier tails than those of the normal distribution. In addition the choice of m plays a very crucial role in the results because of the equal weights $(1/m)$ putted by the model on the past observations. This means that an extreme return today will bump up variance by $(1/m)$ times the squared return for exactly m periods after which variance immediately will drop back down.

Exponentially Weighted Moving Average (EWMA)

Exponentially weighted moving average approaches emphasize recent observations. In contrast to equally weighted approaches, these approaches attach different weights to the past observations contained in the observation period. Because the weights decline exponentially, the most recent observations receive much more weight than earlier observations. The formula for the portfolio standard deviation under the Exponentially Weighted Moving Average (EWMA) approach is given by:

$$\sigma_t = \sqrt{(1-\lambda)\sum_{s=t-k}^{t-1}\lambda^{t-s-1}(x_s - \mu)^2} \ . \qquad (2.4)$$

The parameter λ, referred to as the 'decay factor', determines the rate at which the weights on past observations decay as they become more distant. In theory, for the weights to sum to one, these approaches should use an infinite number of observations k. In practice, for the values of the decay factor considered here, the sum of the weights will

converge to one as fast as many as the observations are. As with the previous approach, the parameter μ is assumed to be equal to zero. Exponentially Weighted Moving Average (EWMA) approaches clearly aim to capture short- term movements in volatility. The following equation gives a more intuitive understanding of the decay's factor role.

$$\sigma_{t+1} = \sqrt{(1-\lambda)r^2 + \lambda\sigma_t^2} \tag{2.5}$$

As shown, an exponentially weighted moving average on any given day is a combination of two components: i) the weighted average of the previous day which receives a weight of λ and ii) yesterday's squared deviation, which receives a weight of $1-\lambda$. This means that the lower the decay factor, the faster the decay in the influence of a given observation. The model just mentioned is the model that JB Morgan's RiskMetrics™ system for market risk management uses with the decay factor λ set equal to 0.94.

The RiskMetrics™ model has some clear advantages. First, it tracks variance changes in a way that is consistent with observed returns. Recent returns matter more for tomorrow's variance than distant returns as λ is less than one and therefore gets smaller when the lag gets bigger. Second, by setting λ equal to 0.94 no estimation of parameters is necessary. RiskMetrics™ model's main shortcoming is that it provides counterfactual longer horizon forecasts.

The GARCH (*q*, *s*) variance model

This uses the class of models developed by Engle (1982) and Bollerslev (1986). The simplest generalized autoregressive conditional heteroskedasticity model is the GARCH (*q*, *s*) and can be written as:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{s} b_i \sigma_{t-i}^2 \tag{2.6}$$

With $\alpha_i + b_i < 1$. Where q denotes the number of considered lagged variance values and s determines this number for the squared residuals.

In addition, $r_t = \mu + \varepsilon_t$ are the log-returns with $\varepsilon_t \sim IID\ N(0, \sigma_t^2)$

We can notice that the RiskMetrics model can be viewed as a special case of the simple GARCH model if we force $\alpha_1=1-\lambda$ so that $\alpha + b=1$ and further $\alpha_0=0$. However there is an important difference: We can define the unconditional or long run variance σ^2 to be

$$\sigma^2 \equiv E[\sigma_{t+1}^2] = \alpha_0 E[R_t^2] + bE[\sigma_t^2] = \alpha_0 + \alpha_1\sigma^2 + b\sigma^2 \Rightarrow \sigma^2 = \alpha_0/(1-\alpha_1-b) \qquad (2.7)$$

It is now clear that if $\alpha_1 + b = 1$ as in the case in the RiskMetrics model, then the long run variance is not well defined in that model. Thus the RiskMetrics model ignores the fact that the long run variance tends to be relatively stable over time. The GARCH model on the other hand implicitly relies on σ^2. This, can be seen by solving for α_0 in the long run variance equation and substituting in the dynamic variance equation. If we do that we get

$$\sigma_{t+1}^2 = (1-\alpha_1-b)\sigma^2 + \alpha_1 r_t^2 + b\sigma_t^2 = \sigma^2 + \alpha_1(r_t^2-\sigma^2) + b(\sigma_t^2-\sigma^2) \qquad (2.8)$$

Thus tomorrow's variance is a weighted average of the long run variance, today's squared return and today's variance, which is very intuitive. However GARCH modeling has also shortcomings.

The main weakness of this approach is that the assumption of conditional normality does not seem to hold for real data, as shown for instance in Danielsson and de Vries (1997). In addition to that as indicated by Christoffersen and Diebold (2000) volatility forecastability of such models decays quickly with the time horizon of the forecast. As an immediate consequence that volatility forecastability is relevant only for short time horizons.

2.2. Non Parametric method

Historical simulation approach (HS)

This category of VaR approaches is similar to the equally weighted moving average category in that it relies on a specific quantity of past historical observations. Rather than using these observations to calculate the portfolio's standard deviation, however historical simulation approaches use the actual percentiles of the observation period as VaR measures. In other words, an empirical distribution must be derived for the price changes over a period prior to the time of calculation. We estimate the portfolio VaR from the maximum loss in this distribution associated with the required statistic likelihood percentile. More specifically, The VaR number based on the Historical Simulation (HS) method is calculated as:

$$VaR_{t+1} = -Quantile(\{r_t\}_{t=1}^{n}, 100\,p) \tag{2.9}$$

Where r_t is the return on day t and p is the confidence level.

The historical simulation approach has some advantages. The first one is that is very easy to implement. No parameters have to be estimated and therefore no numerical optimization has to be performed. The second advantage is that historical simulation approach is model free. That means that it does not rely on any particular parametric model for variance or a normal distribution for the returns. It lets the past data speak fully about the distribution of tomorrow's returns without imposing any further assumption. This has the obvious advantage that relying on modeling assumptions can be misleading if the model is poor.

The model free nature of HS approach has serious drawbacks however. How large the number of past observations should be? If it is too large, then the most recent observations, which presumably are the most relevant for tomorrow's distribution, will carry very little weight and VaR will tend to look very smooth over time. If otherwise, we choose it small, then the sample may not include enough large losses to enable us to calculate VaR with any precision. For example Hendricks (1996) and Vlaar (2000) argued that historical simulation yields more accurate estimates as the sample size

increases, yet Hoppe (1998) proposed the use of a smaller one, since it accommodates the structural changes of the trading behavior more efficiently. Another disadvantage of the historical simulation approach is that the model does not make the assumptions of normality or serial independence and that has as an effect that we cannot accommodate translations between multiple percentiles and holding periods. In other words, the 95th and 99th percentile VaR measures will not be constant multiples of each other and holding periods other than one day will not be fixed multiples of the one day VaR measures[1].

2.3 Semi parametric approach

Filtered historical simulation approach (FHS)

The presented methods (parametric and non-parametric) face several drawbacks. For example, a risk manager must make an assumption for the underlying distribution in order to calculate the parametric VaR number, while under the framework of the historical simulation technique there is no consistent method of estimating and forecasting the volatility innovation. Hull and White (1998) and Barone-Adesi et al. (1999) combined the two methods in order to lessen the problematic use of the "classical" approaches and introduced the VaR estimate based on the Filtered Historical Simulation:

$$VaR_{t+1}^{p} = -Quantile(\{z_t\}_{t=1}^{n}, 100\,p)\sigma_{t+1} \qquad (2.10)$$

Where z_t is the standardized return of the portfolio on day t, p is the confidence level and σ_t is the portfolio's standard deviation on day t. The Filtered Historical Simulation (FHS), is a semi-parametric method, as it combines the non parametric with the parametric techniques. Specifically, it forecasts the variance through a parametric volatility model and uses the quantile of the standardized returns in order to calculate the VaR number.

[1] For more information on the shortcomings of the HS, see Pritsker (1999).

Barone-Adesi and Giannopoulos (2000) argued that the FHS produces risk forecasts that accommodate the current state of the market and therefore it is better than the Historical Simulation (HS). The combination of a parametric method with the HS might offer an improvement in the calculation of the VaR number, as it accommodates the main characteristics of the empirical distribution (non-zero skewness, fat tails and volatility clustering).

Extreme Value Theory (EVT)

Extreme value theory is a powerful and yet fairly robust framework in which to study the tail behavior of a distribution. The central result in extreme value theory states that the extreme tail of a wide range of distributions can approximately be described by a relatively simple distribution, the so-called generalized Pareto distribution (GPD).

Virtually all results in EVT assume that returns are IID. But this assumption truly does not always hold. If our data exhibit some form of time dependency, which takes the form of clustering, we have to take that in mind. We will estimate the tail of the conditional distribution instead of the unconditional one: we will first estimate the conditional volatility model (e.g., via a GARCH procedure), and then estimate the tail index of conditional standardized data. The time dependency of our data will be picked up by the deterministic part of our model, and then we can treat the random process as independent. So we have to consider the standardized returns:

$$z_{t+1} = r_{t+1} \ / \ \sigma_{t+1} \sim IID\ N(0,1). \qquad (2.11)$$

Fortunately we can in many cases reasonably assume that these standardized returns are IID. Thus we will proceed to apply EVT to the standardized returns and then combine the EVT with a conditional variance model. The Generalized Pareto Distribution (GPD) can describe the behavior of the extremes, which can be summarized by the so-called tail index, which we have to estimate. The estimation technique that has been implemented is based on modeling the exceedances over a threshold u, which is also known as the peaks

over threshold method (POT). The probability that the standardized returns are greater that u is given by:

$$F_u(x) = \Pr\{z - u \le x / z > u\} = \frac{F(x+u) - F(u)}{1 - F(u)} \text{ for } x>u. \quad (2.12)$$

If one let the threshold u to get large, the GPD is the limiting distribution of the exceedances where:

$$G(x:\xi,\beta) = \begin{cases} 1 - (1+\xi\frac{x}{\sigma})^{-1/\xi}, & \text{if } \xi \neq 0 \\ 1 - exp[-x/\sigma], & \text{if } \xi = 0 \end{cases} \quad (2.13)$$

If we consider points x with $x>u$ in the tail of the distribution and let $y=x+u$ then by rearranging equation (2.12), we get:

$$F(y) = 1 - [1 - F(u)][1 - F_u(y-u)]. \quad (2.14)$$

If we let T denote the total sample size and let T_u the number of observations beyond the threshold u, the term $1-F(u)$ can be estimated by the proportion of data point beyond the threshold u, call it T_u/T. $F(u)$ can be estimated by MLE on the standardized observations in excess of the chosen threshold. Assuming $\xi \neq 0$ we then have the distribution:

$$F(y) = 1 - T_u / T(1+\xi(y-u)/\sigma)^{-1/\xi} \quad (2.15)$$

For a positive tail parameter an estimator exists, the so-called Hill estimator. We can write:

$$\Pr(z > y) = 1 - F(y) = L(y)y^{-1/\xi} \approx cy^{-1/\xi}, \text{ for } y > u. \quad (2.16)$$

The approximation builds on the fact that $L(y)$ is a slowly varying function of y for most distributions and thus can be set equal to a constant, c. Given the approximation and using the definition of the conditional distribution we can define the likelihood function for all observations y_i larger than u, as

$$L = \prod_{i=1}^{T_u} f(y_i) / (1 - F(u)) = \prod_{i=1}^{T_u} -\frac{1}{\xi} c y_i^{-1/\xi - 1} \Big/ (cu^{-1/\xi}) \quad \text{for } y_i > u. \quad (2.17)$$

So that the log likelihood function is:

$$\ln L = -\sum_{i=1}^{T_u} -\ln \xi - (1/\ \xi + 1)\ln(y_i) + \frac{1}{\xi}\ln u. \tag{2.18}$$

Taking the derivative with respect to ξ and setting it to zero yields the Hill estimator:

$$\xi = \frac{1}{T_u}\sum_{i=1}^{T_u} \ln(y_i /\ u). \tag{2.19}$$

We can estimate the c parameter by ensuring that the fraction of observations beyond the threshold is accurately captured by the density as in:

$$F(u) = 1 - T_u/T \tag{2.20}$$

From the definition of *F(u)* we can write:

$$1 - cu^{-1/\xi} = 1 - T_u / T. \tag{2.21}$$

Solving for c this equation yields the estimate:

$$c = \frac{T_u}{T}u^{1/\xi}. \tag{2.22}$$

Our estimate of the cumulative density function for excess observations is therefore:

$$\hat{F}(y) = 1 - cy^{-1/\xi} = 1 - \frac{T_u}{T}(y/u)^{-1/\xi}. \tag{2.23}$$

In order to calculate VaR, we define y_i to be a standardized loss, that is:

$$y_i = -\frac{r_i}{\sigma_i}. \tag{2.24}$$

The first step is to estimate *c* and ξ from the losses y_i using the Hill estimator. Next we need to compute the inverse cumulative distribution function, which gives us the quantiles. The third step is to set the estimated cumulative distribution function equal to *p* so that there is only *1- p* probability of getting a standardized loss worse than the quantile, F_{1-p}^{-1}, which is defined by :

$$F(F_{1-p}^{-1}) = 1 - p. \tag{2.25}$$

From the definition of *F(*)* we can solve for the quantile to get:

$$F_{1-p}^{-1} = u\left[\frac{pT}{T_u}\right]^{-\xi}. \tag{2.26}$$

Now VaR from the EVT combined with the variance model is now easily calculated as:

$$VaR_{t+1}^{p} = \sigma_{t+1} u \left[\frac{pT}{T_u} \right]^{-\xi} \tag{2.27}$$

Where p is the confidence level, T is the sample size, T_u is the number of exceedances over the threshold u, σ_t is the portfolio's standard deviation on day t and ξ is the tail index. The above procedure can be characterized as the conditional EVT approach. In the situation where no volatility clustering exists we just model the tails of the unconditional distribution rather than the conditional one. In that case the VaR using the EVT approach is given by the formula:

$$VaR_{t+1}^{p} = u + \frac{\hat{\beta}}{\hat{\xi}} [(\frac{T}{T_u} p)^{-\hat{\xi}} - 1] \tag{2.28}$$

Where as before p is the confidence level, T is the sample size, T_u is the number of exceedances over the threshold u, β is a scale parameter and ξ is the tail index.

The most important in EVT application is the determination of the threshold parameter u. Using the Hill plot (equation 2.19) we can choose the threshold from the area that the plot becomes fairly stable. A more practical way, especially when we are working with a rolling sample, is to set as threshold value such that only a small percentage of the sample exceeds that value.

3. Backtesting VaR

The ultimate test of every VaR model lies in its ability to forecast accurately the maximum loss likely to occur over the given probability level and time horizon. It is important to note that the estimated VaR number must neither overestimate nor underestimate the "true" but unobservable value. In the former case, the financial institution does not use its capital efficiently, while in the latter case it cannot cover future losses. The process of assessing the forecasting performance is called backtesting. The simplest way to back-test the model is to check whether the number of cases that the actual losses exceed the calculated VaR (exceptions) equals the specified confidence level (p). This is the Bank for International Settlements (BIS) guideline[2]. However, since the sample size of daily observations is finite, the actual number of exceptions may differ from the percentage implied by the model's confidence level, even in cases where the model is in fact accurate. Therefore, the accuracy of the model should be examined using various additional tests.

[2] According to the Basle Committee, each bank must meet a capital requirement expressed as the maximum of the bank's previous day's VAR number and an average of the daily VAR measures on each of the preceding 60 trading days, adjusted by a multiplication factor. The multiplication factor is to be set within a range of three to four depending on the supervisor's assessment of the bank's risk management practices and the results of a simple back-test that counts the number of exceptions. For a 250-day backtesting period and a 99% confidence level, a model that provides up to four exceptions falls in the 'green zone', receiving a multiplication factor of three. A model with five to nine exceptions falls in the 'yellow zone', with the multiplication factor increasing gradually up to 3.85. For 10 or more exceptions, the model falls in the 'red zone', where the multiplication factor is set equal to four.

Unconditional coverage

The most common backtest is the test for unconditional coverage proposed by Kupiec (1995).

If we recall that a VaR_{t+1}^{p} measure promises that the actual return will only be worse than the forecast $p100\%$ of the time and observe a time series of past VaR forecasts and past returns, we can define the "hit sequence" of VaR violations as:

$$I_{t+1} = \begin{cases} 1 & if \quad r_{t+1} < -VaR_{t+1}^{p} \\ 0 & if \quad r_{t+1} > -VaR_{t+1}^{p} \end{cases} \tag{3.1}$$

the "hit sequence" returns a *1* on day $t+1$ if the loss on that day was larger than the VaR number predicted in advance for that day and a 0 if it was not violated. When backtesting the risk model we construct a sequence $\{I_{t+1}\}_{1}^{T}$ across T days indicating when the past violations occurred. The simplest method in determining the adequacy of a VaR measure is to test the null hypothesis that the proportion of violations, call it π, is equal to the expected one, call it p. To test this we write the likelihood of the IID Bernoulli hit sequence:

$$L(\pi) = \prod_{t=1}^{T} (1-\pi)^{1-I_{t+1}} \pi^{I_{t+1}} = (1-\pi)^{T_0} \pi^{T_1} \tag{3.2}$$

Where T_0 and T_1 is the number of *0*s and *1*s in the sample.

We can estimate π from $\hat{\pi} = \frac{T_1}{T}$ and with this we have:

$$L(\hat{\pi}) = (1-\frac{T_1}{T})^{T_0} (\frac{T_1}{T})^{T_1} \tag{3.3}$$

Under the null that $\pi=p$ we have the likelihood:

$$L(p) = \prod_{t=1}^{T} (1-p)^{1-I_{t+1}} p^{I_{t+1}} = (1-p)^{T_0} p^{T_1} \tag{3.4}$$

We can check the unconditional coverage hypothesis using the likelihood ratio test:

$$LR_{uc} = -2\ln[L(p)/L(\hat{\pi})] \tag{3.5}$$

or

$$LR_{uc} = -2\ln[(1-p)^{T_0} p^{T_1} / \{(1-\frac{T_1}{T})^{T_0}(\frac{T_1}{T})^{T_1}\}] \sim X_1^2. \tag{3.6}$$

Choosing a significance level for the test we will have a critical value from the X_1^2 distribution. If the LR_{uc} test value is larger than the critical value then we will reject the VaR model at that significance level. The choice of significance level comes down to an assessment of the cost of making two types of mistakes: we could reject a correct model (Type I error) or we could fail to reject an incorrect model (Type II error). Increasing the significance level implies larger Type I errors but smaller Type II errors and vice versa. In risk management the Type II errors may be very costly so that a high significance level such as 10% may be the appropriate one. This test can reject a model for both high and low failures, but as stated by Kupiec, its power is generally poor.

Joint hypothesis testing

Christoffersen (1998) developed a conditional coverage test, which jointly investigates the assumption of unconditional coverage and independence of failures.
He does that in a likelihood ratio (LR) testing framework firstly by specifying an LR test of correct unconditional coverage, secondly by an LR test of independence and finally an LR test that combines the previous two to form a complete test of the conditional coverage.

Independence Testing

If the VaR violations in a sample happen around the same time, we would not be happy with a VaR with just a correct unconditional coverage. We would have to establish a test that will be able to reject a VaR with clustered violations. Christoffersen assumes that the hit sequence is dependent over time and that it can be described as a so-called first order Markov sequence with transition probability matrix:

$$\Pi_1 = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix} \tag{3.7}$$

These transition probabilities simply mean that conditional on today being a nonviolation the probability of tomorrow being a violation is π_{01}. The probability of tomorrow being a violation given today is also a violation is π_{11}. The first order Markov property refers to the assumption that only today's outcome matters for tomorrow's outcome. The probability of a nonviolation following a nonviolation is $1-\pi_{01}$ and the probability of a nonviolation following a violation is $1-\pi_{11}$. If we observe a sample of T observations then we can write the likelihood function of the first order Markov process as:

$$L(\Pi_1) = (1-\pi_{01})^{T_{00}} \pi_{01}^{T_{01}} (1-\pi_{11})^{T_{10}} \pi_{11}^{T_{11}} \tag{3.8}$$

Where T_{ij} ,ij=0,1 is the number of observations with a j following an i. Solving for the maximum likelihood estimates we get the matrix of the estimated transition probabilities:

$$\hat{\Pi}_1 = \begin{bmatrix} \hat{\pi}_{00} & \hat{\pi}_{01} \\ \hat{\pi}_{10} & \hat{\pi}_{11} \end{bmatrix} = \begin{bmatrix} \dfrac{T_{00}}{T_{00}+T_{01}} & \dfrac{T_{01}}{T_{00}+T_{01}} \\ \dfrac{T_{10}}{T_{10}+T_{11}} & \dfrac{T_{11}}{T_{10}+T_{11}} \end{bmatrix} \tag{3.9}$$

We can test the independence hypothesis that $\pi_{01} = \pi_{11}$ using a likelihood ratio test:

$$LR_{ind} = -2\ln[L(\hat{\pi}) / L(\hat{\Pi}_1)] \sim X_1^2 \tag{3.10}$$

Where $L(\hat{\pi})$ is the likelihood for the alternative hypothesis from the LR_{cc} test.

Conditional coverage test

Ultimately we can test simultaneously if the VaR violations are independent and the average number of violations is correct. We can test this using the conditional coverage test. Under the null hypothesis that the failure process is independent and the expected proportion of violations is equal to p, the appropriate likelihood ratio is:

$$LR_{CC} = -2\ln[L(p)/\ L(\hat{\Pi}_1)] \sim X_1^2 \qquad (3.11)$$

Which corresponds to testing that $\pi_{01} = \pi_{11} = p$. We can notice that the joint test of conditional coverage can be calculated by simply summing the two individual tests for unconditional coverage and independence, i.e.

$$LR_{cc} = LR_{uc} + LR_{ind} \qquad (3.12)$$

4. Conditional VaR (CVaR)

A key shortcoming of VaR is that it only tells us the most we can lose if a tail event does not occur; if a tail event does occur, we can expect to lose more than the VaR, but the VaR itself gives us no indication of how much that might be. The magnitude however should be of a serious concern for risk management purposes. Large VaR exceedances are much more likely to cause financial distress, than are small exceedances, and we should therefore consider a risk measure that accounts for the magnitude of large losses as well as their probability of occurring. Artzner et al. (1997) proposed as that risk measure the conditional VaR or expected shortfall (for continuous distributions is the same). Expected shortfall (ES) is defined as:

$$ES_{t+1}^{p} = -E_t[r_{t+1} \mid r_{t+1} < -VaR_{t+1}^{p}]. \qquad (4.1)$$

The ES can be considered a better risk measure than VaR for a number of reasons: The expected shortfall tells us the expected value of tomorrow's return, conditional on it being worse than the VaR, it tells us how bad, bad things might be while VaR tells us nothing other than to expect a loss higher than the VaR itself. Generally, ES gives a more complete description of the losses possible in the tails of the distribution. Artzner et al. (1999) defined four coherence properties; monotonicity, translation invariance, homogeneity and subadditivity, desirable of a risk statistic used for capital adequacy purposes. For risk measurement applications from non-normal distributions the VaR statistic does not exhibit the subadditivity property. Expected shortfall satisfies all four coherence properties.

But how can we compute ES? It really depends on the distribution assumption we make for the returns. For the Normal distribution we have: The expected value of a normal variable with zero mean return truncated at the VaR is

$$ES_{t+1}^{p} = -E_t[r_{t+1} / r_{t+1} < -VaR_{t+1}^{p}] = \sigma_{t+1} \frac{\varphi(-VaR_{t+1}^{p} \ / \ \sigma_{t+1})}{\Phi(-VaR_{t+1}^{p} \ / \ \sigma_{t+1})} \qquad (4.2)$$

Where $\varphi(*)$ denotes the density function and $\Phi(*)$ the cumulative density function of the standard Normal distribution. In the Normal case we know that $VaR_{t+1}^{p} = -\sigma_{t+1}\Phi_{p}^{-1}$ thus:

$$ES_{t+1}^{p} = \sigma_{t+1} \frac{\phi(\Phi_{p}^{-1})}{p} \tag{4.3}$$

So, the expected shortfall for Normal distributions can be computed as:

$$ES_{t+1}^{p} = -VaR_{t+1}^{p} \frac{\phi(\Phi_{p}^{-1})}{p\Phi_{p}^{-1}} \tag{4.4}$$

Similarly, for fat-tailed distributions we have:

$$ES_{t+1}^{p} = VaR_{t+1}^{p} \frac{1}{1-\xi}, \tag{4.5}$$

Where ξ tail estimator as we have seen in the EVT analysis[3].

5. Backtesting the expected shortfall

The statistical adequacy of the VaR forecasts is obtained by previous backtesting tests: the unconditional coverage (equation 3.6), the independence test (equation 3.10) and the conditional coverage test (equation 3.12). If a model is not rejected, it forecasts VaR accurately. However, in most cases, more than one model seems to be adequate and hence, the risk manager cannot select a unique risk management technique.

To overcome this shortcoming of the backtesting measures, Lopez (1999) proposed a forecast evaluation framework based on loss function. The loss function enables us to rank the models and specify a utility function that accommodates the specific concerns of a risk manager. He suggested the following loss function:

$$C_t = \begin{cases} 1 + (L_t - VaR)^2 & \text{if } L_t > VaR_t \\ 0 & \text{if } L_t \leq VaR_t \end{cases} \tag{5.1}$$

[3] Equation 4.5 is valid for small values of p.

where L_t denotes the actual loss. However, this loss function it is not much intuitive, because squared returns have no ready interpretation.

So, Dowd (2000) proposed a different loss function more intuitive. He suggested as a loss function the tail loss itself[4]:

$$C_t = \begin{cases} L_t & \text{if } L_t > VaR_t \\ 0 & \text{if } L_t \leq VaR_t \end{cases} \tag{5.2}$$

The expected value of the tail loss is of the ES, so we can use the ES as a benchmark and use a quadratic score function such as:

$$QS = \frac{2}{n}\sum_{t=1}^{n} (C_t - ES_t)^2 . \tag{5.3}$$

The lowest the value of the score function the better the model.

This approach penalizes deviations of the tail losses from their expected value. Moreover because it is quadratic, it gives very high tail losses much greater weight than more 'normal' tail losses and therefore comes down hard on large losses. The lowest the value of the score function, the better the model.

[4] It should be noted that as with any application of functions, this approach is vulnerable to misspecification of the loss function.

6. Data Set

6.1 Spot prices

We used daily spot prices obtained from DataStream for WTI Crude oil (CL), Brent Crude oil (CO), Unleaded Gasoline (HU), Heating Oil (HO) and Natural Gas (NG)[5]. All of the above are traded on NYMEX except Brent crude oil, which is traded in IPE.

Our data were from a period from 01/11/1993 until 31/08/2005. We split that data into two sub periods. The first is from 01/11/1993 until 31/12/1997 and we call it the in-sample period since it is used for parameter estimation (where that is necessary). The second sub period, which is called the out of sample period, is from 01/01/1998 until 31/08/2005 and it is used for backtesting. More specifically, we worked with 3087 observations and generated 2000 out-of sample forecasts. We transformed these prices into log returns, which we then used in our calculations.

As a first step we run the log returns through some statistical tests (Table 1a). All logarithmic price changes series are decidedly non-normal. The Bera-Jarque test's null hypothesis of normality is rejected at standard confidence levels for all series. The unconditional distributions of the daily logarithmic price changes are highly kurtotic and skewed. The petroleum products and natural gas exhibit negative skewness; the unconditional distribution has a long left tail relative to a symmetric distribution. In addition, serially dependence for the log returns was present, except for unleaded gasoline. The squared returns of all the petroleum products and natural gas exhibit significant correlation and persistence in the second order moments. Autocorrelation p-values for lags 10, 15, and 20 for the log price change series and squared log price change series are presented in the lower panel of Table 1a.

---

[5] CL - West Texas Intermediate crude oil at Cushing, Oklahoma
CO - Brent crude oil (IPE)
HO - Heating oil at New York harbor
HU - Unleaded Gasoline at New York harbor
NG - Natural Gas at Henry Hub Louisiana

Table 1a.

	CL	CO	HO	HU	NG
Mean	0.00044	0.00046	0.00044	0.00057	0.00056
Median	0.00000	0.00069	0.00000	0.00000	0.00000
Maximum	0.15873	0.12556	0.24115	0.23002	0.87547
Minimum	-0.17217	-0.11353	-0.33438	-0.16486	-1.27300
St. Deviation	0.02368	0.01771	0.02675	0.02751	0.06136
Skewness	-0.32527	-0.30361	-0.46704	-0.20661	-1.46454
Kurtosis	7.19516	6.20824	18.1152	6.62672	110.616
Jarque-Bera (probability)	2318.1** (0.000)	1371.3** (0.000)	29499.1** (0.000)	1713.7** (0.000)	1490742** (0.000)
Autocorrelation (returns)					
p-value(Lag 10)	0.022	0.000	0.001	0.403	0.001
p-value(lag 15)	0.015	0.000	0.000	0.377	0.000
p-value(lag 20)	0.010	0.000	0.000	0.607	0.000
Lung-Box (20)	45.47*	133.0*	54.34*	17.70	352.90*
Autocorrelation (returns^2)					
p-value(Lag 10)	0.000	0.000	0.000	0.000	0.000
p-value(lag 15)	0.000	0.000	0.000	0.000	0.000
p-value(lag 20)	0.000	0.000	0.000	0.000	0.000
Lung-Box (20)	122.22*	229.58*	1344.90*	111.55*	1109.60*

Table 1a: Descriptive statistics for daily spot log returns for the energy complex. Statistics with an asterisk denotes rejection of the null hypothesis (H_0: absence of correlation) at the 5% significance level. Statistics with a double asterisk denotes rejection of the normality assumption.

	CL	CO	HO	HU	NG
Mean	25.827	24.226	70.942	73.245	3.497
Median	23.190	22.260	63.190	65.830	2.710
Maximum	69.820	66.930	205.09	280.45	19.00
Minimum	10.730	9.7900	28.420	29.070	1.035
St. Deviation	10.560	10.230	30.128	29.133	1.892
Skewness	1.374	1.496	1.476	1.381	1.305
Kurtosis	4.861	5.445	5.088	5.604	5.305

Table 1b: WTI crude oil and Brent crude oil are measured in $/Barrel. Heating oil and unleaded gasoline in cents/gallon and natural gas in $/MMBTU

6.2 Nearby Futures

The NYMEX light sweet (low sulfur) crude oil futures contract is the world's most heavily traded commodity futures contract. It has been trading since 1983. Each futures contract is written on 1,000 barrels of crude oil. On any given day, there are contracts trading for the next 30 consecutive months as well as contracts for delivery in 36, 48, 60, 72, and 84 months (35 futures contracts in total). The delivery period is a full month, meaning that deliveries must be initiated on or after the first calendar day and completed on or before the last calendar day of the delivery month. Trading terminates at the close of the third business day prior to the 25th calendar day of the month preceding the delivery month. Settlement is done with physical delivery, even though most of the contracts are closed before expiration. The underlying asset can be thought to be the West Texas Intermediate (WTI) that serves as the reference for most crude oil transactions. However, a number of other grades of crude are also deliverable[6]. The delivery point is Cushing, Oklahoma.

The IPE in London is the second most liquid crude oil market in the world. The Brent Crude futures contract has been trading on the IPE since 1988. It is part of the Brent blend complex (that also consists of the physical and forward Brent) that is used as a basis for pricing the two thirds of the world's traded crude oil. Each futures contract is 1,000 barrels of Brent crude oil. There are contracts trading for the next twelve consecutive months, then quarterly out to a maximum 24 months, and then half-yearly out to a maximum 36 months (eighteen futures contracts in total). Trading terminates at the close of the business day immediately preceding the 15th day prior to the first day of the delivery month. Settlement is done with physical delivery or alternatively there is the option to settle in cash against the IPE Brent Index price of the day following the last trading day of the futures contract. The underlying asset is the pipeline-exported Brent

[6] Deliverable US crudes are crudes with sulfur content of 0.42% by weight (or less) and an American Petroleum Institute (API) gravity between 37° and 42°. Deliverable streams are the WTI, Low Sweet Mix, New Mexico Sweet, North Texas Sweet, Oklahoma Sweet, and South Texas Sweet. Deliverable non-US crudes are crudes with an API gravity between 34° and 42°. Deliverable streams are the UK's Brent and Forties and Norway's Oseberg Blend at a discount of $0.30 per barrel, Nigeria's Bonny Light and Colombia's

blend supplied at the Sullom Voe terminal in the North Sea. The prices of the NYMEX and IPE contracts are quoted in US dollars and cents per barrel and are used as benchmarks for pricing crude oil and its refined products on an international basis.

Gasoline and heating oil (also known as No. 2 fuel oil) are two most important refined products, accounting for approximately 40% and 25% of the yield of a crude oil barrel respectively. Both heating oil and gasoline futures trade in NYMEX in contracts of 42,000 US gallons (equivalent to 1,000 barrels). Prices are quoted in US dollars and cents per gallon. There exist contracts for the next 18 consecutive months for heating oil and the next 12 consecutive months for gasoline. The delivery period begins on the day after the fifth business day of the delivery month and ends on the last business day of the delivery month. Trading terminates at the close of business on the last business day of the month preceding the delivery month. Settlement is done with physical delivery. The grade and quality of the deliverable heating oil and gasoline conform to industry standards for fungible No. 2 heating oil, and for Phase II Complex Model Reformulated Gasoline in accordance with Colonial Pipeline Co. specifications for fungible A grade, 87 octane index gasoline, respectively. Trading in the three NYMEX petroleum futures contracts is conducted by open outcry from 10:05am until 2:30pm New York time. Trading in the IPE contract is conducted by open outcry from 10:02am until 7:30pm London time (5:02am until 2:30pm New York time).

NYMEX energy complex futures contracts have differing specifications concerning the maximum allowable price change. These contract specifications influence the magnitude of observed price changes. The Brent crude oil contract has no maximum price change limits, West Texas Intermediate crude, heating oil and unleaded gasoline contracts specify conditional maximum price change limits. In futures markets, where maximum price change limits become binding it is expected that with sufficient volatility of spot market price changes, the range of logarithmic changes of the nearby futures contract price will be less than the range of the logarithmic changes of the spot market price.

We obtained from DataStream daily prices of the nearby futures for WTI crude oil, heating oil, unleaded gasoline traded in NYMEX and Brent crude nearby futures contract traded in IPE. We used the nearby futures i.e. the futures with the shortest maturity because they are far more liquid than futures contracts with longer maturities.

Our data were from a period from 01/11/1993 until 31/08/2005. We split that data into two sub periods. The first is from 01/11/1993 until 31/12/1997 and we call it the in-sample period since it is used for parameter estimation (where that is necessary). The second sub period, which is called the out of sample period, is from 01/01/1998 until 31/08/2005 and it is used for backtesting. More specifically, we worked with 3087 observations and generated 2000 out-of sample forecasts. We transformed these prices into log returns, which we then used in our calculations.

The nearby futures price series for each commodity is constructed from the settlement price of the futures contract nearest delivery. The nearby series rolls to the settlement prices of the next contract on the nearby contract's last day of trading.

As a first step we run the log returns through some statistical tests (Table 2a). As mentioned before, using the maximum and minimum reported in Table 2a to compare the range of daily logarithmic price changes in spot and futures market for a specific commodity reveals that the range of logarithmic price changes in the nearby futures where maximum allowable price change exists, is less than the range of price change in the spot market.

All logarithmic price changes series are decidedly non-normal. The Bera-Jarque test's null hypothesis of normality is rejected at standard confidence levels for all series. The unconditional distributions of the daily logarithmic price changes are highly kurtotic and skewed. In addition, no significant serially dependence was present; Using the Ljung-Box test we can see that the null (H_0: no significant autocorrelation exists) is not rejected at the 0.05 significance level for all returns series, which indicates the absence of significant autocorrelation for these series The squared returns though, exhibit significant correlation except for gasoline return series as can also been seen from the corresponding p-values. Autocorrelation p-values for lags 10, 15, and 20 for the log price change series and squared log price change series are presented in the lower panel of Table 2a.

Table 2a

	CL	CO	HO	HU
Mean	0.000449	0.000460	0.000445	0.000503
Median	0.000000	0.000000	0.000000	0.000000
Maximum	0.122260	0.661070	0.103980	0.121210
Minimum	-0.165450	-0.144370	-0.138120	-0.221100
St. Deviation	0.021610	0.023950	0.022269	0.023496
Skewness	-0.417965	6.609266	-0.331327	-0.654482
Kurtosis	6.349067	191.2435	5.510824	9.106032
Jarque-Bera (probability)	1532.57* (0.000)	4580.38* (0.000)	867.36* (0.000)	5015.99* (0.000)
Autocorrelation (returns)				
p-value(Lag 10)	0.209	0.667	0.851	0.723
p-value(lag 15)	0.060	0.841	0.420	0.568
p-value(lag 20)	0.014	0.848	0.287	0.695
Lung-Box (20)	36.375	13.650	23.036	16.337
Autocorrelation (returns^2)				
p-value(Lag 10)	0.000	0.001	0.001	0.526
p-value(lag 15)	0.001	0.001	0.001	0.858
p-value(lag 20)	0.001	0.001	0.001	0.930
Lung-Box (20)	127.69*	125.88*	222.81*	11.561

Table 2a: Descriptive statistics for daily nearby futures log returns for the energy complex. Statistics with an asterisk denotes rejection of the null hypothesis (H_0: absence of correlation) at the 5% significance level. Statistics with a double asterisk denotes rejection of the normality assumption.

Table 2b. Descriptive statistics for daily nearby futures prices for the energy complex

	CL	CO	HO	HU
Mean	25.754	23.854	70.484	75.894
Median	22.905	21.135	62.485	66.250
Maximum	69.810	67.570	208.50	225.530
Minimum	10.720	9.6400	29.520	32.660
St. Deviation	10.569	10.515	29.972	29.750
Skewness	1.424	1.449	1.576	1.378
Kurtosis	5.009	5.267	5.477	4.990

Table 2b: WTI crude oil and Brent crude oil are measured in $/Barrel. Heating oil and unleaded gasoline in cents/gallon

7. Implementation[7]

7.1 Calculating and backtesting VaR

The method we followed was the same for both the spot market and the nearby futures[8]. We computed one-day VaR forecasts for each of the methods discussed above, for 99% and 95% confidence levels. In the variance – covariance approach, the volatility coefficients are estimated by the three different models discussed earlier, a MA, an EWMA and a GARCH (1,1)[9] model with normal innovations (equations 2.3, 2.5, 2.6 respectively). In the MA approach we used a rolling sample of 74 observations because this model yields the lowest root percentage mean square error (Figlewski 1997). In the EWMA model, the decay factor is set equal to 0.94 following RiskMetrics™ methodology (JP Morgan 1996). In the GARCH variance model as far as the order of the conditional volatility specifications, we decided to use $m=s=1$[10]. The specific model proved adequate as it captured the volatility-clustering phenomenon with success. This can be verified by the ACF of the squared standardized innovations (see Appendix C) and also by the Ljung-Box-Pierce test for the standardized innovations.

In the HS approach, VaR is calculated using one-day rolling window of 100 and 250 observations (HS-100 and HS-250) using equation (2.9). In the FHS approach, VaR is calculated as before (FHS-100 and FHS-250), but now our observations are the standardized returns- equation (2.24)- rather than simply the returns as it were in the HS approach. The equation we used for VaR computation via FHS is equation (2.10) and the returns are standardized by the volatility obtained from the corresponding GARCH(1,1) model for each data set .

---

[7] All computations needed in this dissertation were implemented in MATLAB™. The codes are available upon request.

[8] The exception is gasoline nearby futures, where no correlation structure for the squared returns was found. So, no model for describing the conditional volatility was needed. As a result we didn't use a GARCH nor a FHS method. In the EVT approach we the unconditional EVT procedure (equation 2.28) rather than the conditional one, which was needed in the other cases where the volatility clustering was evident.

[9] In the case of spot market data for WTI crude oil, Brent crude oil, Heating oil and Natural gas we specifically used an AR(1)GARCH(1,1) because we found significant AR coefficients.

[10] Although that in the majority of empirical volatility forecasting studies, the order of one lag has proven to work efficiently, we also checked orders bigger than one lag but the test we imposed on them showed the superiority of the order of one lag.

In the EVT approach we used formula (2.27) to calculate VaR and we used a rolling sample of 1000 observations as proposed by Gencay and Selcuk (2004). As observations we used the standardized losses. The threshold value was set such, that when calculating VaR at 95% confidence level only 5% of the sample observations exceed that and when calculating 99% VaR only 1% of sample observations exceed that. This method of threshold calculation was preferred as more practical instead of estimating the threshold at each step[11]. After that, we calculate the number of exceedances over the threshold, which we use in the estimation of the tail index. Next, by using the rolling window of 1000 observations we estimated at each step the tail index with maximum likelihood estimation.

All the methods above produced series of VaR estimates for 99% and 95% confidence level. From that series we kept the last 2000 VaR forecasts because this number is in line with our out-of sample period in which the VaR estimates will be backtested.

For backtesting VaR models, we use the out-of sample estimations of VaR and we check them against the realized losses. We put all the models' results through the Christoffersen tests and only those that pass the conditional coverage test (equation 3.12) are accepted. We will call this the first backtesting stage. It should be mentioned here that the significance level, which the test relies on, is set equal to 10%. We try to achieve three objectives by using a high cut-off point: first to ensure that the "successful" models will accurate estimate the expected coverage rate. Second, the VaR violations will not be clustered and third to reject easily an incorrect model. Specifically, this level is chosen so as to avoid accepting an incorrect model, at the cost of perhaps rejecting a correct model. From that it follows naturally that our test's results are quite conservative and that fact should be taken in mind.

[11] A different way of setting the threshold value would be to construct the Hill plot and define u from the area where the plots become stable. We instead used those plots as a verification of our initial threshold value choice. The specific plots are not presented since they did not yield different results from our initial choice of threshold determination but are available upon request.

Finally in our effort to find the best method for VaR calculation, we compute the average VaR value for all the methods and characterize as the best, the one that gets accepted from the conditional coverage test and yields the lowest average VaR value[12].

7.2 Calculating and backtesting ES

For the variance covariance approach, ES is calculated using formula (4.4) and in the EVT approach using the formula (4.5). In the HS (FHS) approach the procedure is quite different. Using the same rolling window of observations as in the VaR calculation, we find the losses (standardized losses) that exceed the estimated VaR at each step, and we take their average value. This is the estimated ES. Again we keep the last 2000 values from the estimated ES series, in order to compare them with the VaR results obtained earlier and with the realized losses of the out-of sample period.

The models that have not been rejected from the VaR backtesting (first backtesting stage) are backtested again using the loss function (equation 5.2) and the score function (equation 5.3). The model that performs the lowest value from the score function is the most appropriate. We will call this the second evaluation stage.

[12] This is the first of the approaches we are going to use in order to determine the best method for VaR computation. We will call this as the first evaluation stage.

8. Results for spot market

8.1 VaR Results

The following Tables from 3 to 7, present the results obtained for all our data sets and for all the methods employed for VaR calculation. We present the average VaR for each method and for each confidence level. In addition the number of exceptions is presented and by number of exceptions we mean the number that shows how many times the actual losses exceeded our VaR prediction. Finally the results of the likelihood ratio tests we used for backtesting are also shown. In addition, after each table, we present for all the methods employed, the figures of the calculated out-of sample 99% and 95% VaRs, respectively, against the actual losses (negative log-returns).

Table 3. VaR backtesting results for WTI crude oil spot

	MA	EWMA	AR(1)GARCH(1,1)	HS-100	HS-250	FHS-100	FHS-250	EVT
99% VaR								
Average VaR	0.0575	0.0572	0.0676	**0.0625**	**0.0671**	**0.0688**	**0.0719**	**0.0738**
No.of exceptions	36	45	33	16	23	23	25	26
LR_{uc}	10.467*	23.326*	5.244*	0.863	0.436	0.456	1.174	1.667
LR_{ind}	0.170	0.792	0.444	0.258	1.236	1.288	0.989	0.789
LR_{cc}	10.637**	24.118**	5.688**	**1.121**	**1.673**	**1.599**	**2.164**	**2.546**
95% VaR								
Average VaR	0.0407	**0.0405**	**0.0389**	0.0386	0.0394	**0.0403**	**0.0412**	**0.0419**
No.of exceptions	103	103	113	111	116	105	106	97
LR_{uc}	0.097	0.097	1.724	1.243	2.585	0.264	0.378	0.092
LR_{ind}	5.343*	1.334	1.079	4.998*	8.823*	0.055	0.028	0.363
LR_{cc}	5.440**	**1.431**	**2.803**	6.241**	11.409**	**0.319**	**0.406**	**0.456**

Table 3: This table shows the components of Christoffersen's test applied to our family of methods for 99% and 95% VaR estimation. The critical value for the LR_{uc} and LR_{ind} tests is 2.706 and for the LR_{cc} test it is 4.605. The abbreviations LR_{uc}, LR_{ind}, LR_{cc} are for the likelihood ratio tests of unconditional coverage, independence and conditional coverage respectively (equations 3.6, 3.10, 3.12). Our out-of sample size is 2000. The average VaR values are in terms of log returns. A single asterisk denotes rejection of the method for the particular test and the double asterisk denotes rejection of the method for the final test of conditional coverage at the 10% significance level.

From Table 3 we see that the models that pass the test for the 99% confidence level are the EVT, HS-100, HS-250, FHS-100 and FHS-250. All of them fall in the 'green' BIS zone, but the least conservative, as it can be seen from its average value, is the HS-100 model. For the 95% confidence level most of the models (except the HS and the MA) pass the test. In fact the one with the lowest average is the one with the most exceptions and we mean the GARCH (1,1) model. However, as we have mentioned before, the absolute number of exceedances is only one factor among others that affects the final outcome of the backtest. So the GARCH (1,1) model is considered to be the best among the others accepted models for calculating 95% VaR.

Figure 1. 99% and 95% VaR for WTI crude oil spot against actual losses.
In the y-axis are the losses, which correspond, to negative log returns and in the x-axis is the out of sample period. The panels from A to D present the 99% VaR plots and the panels from E to H present the 95% VaR plots.

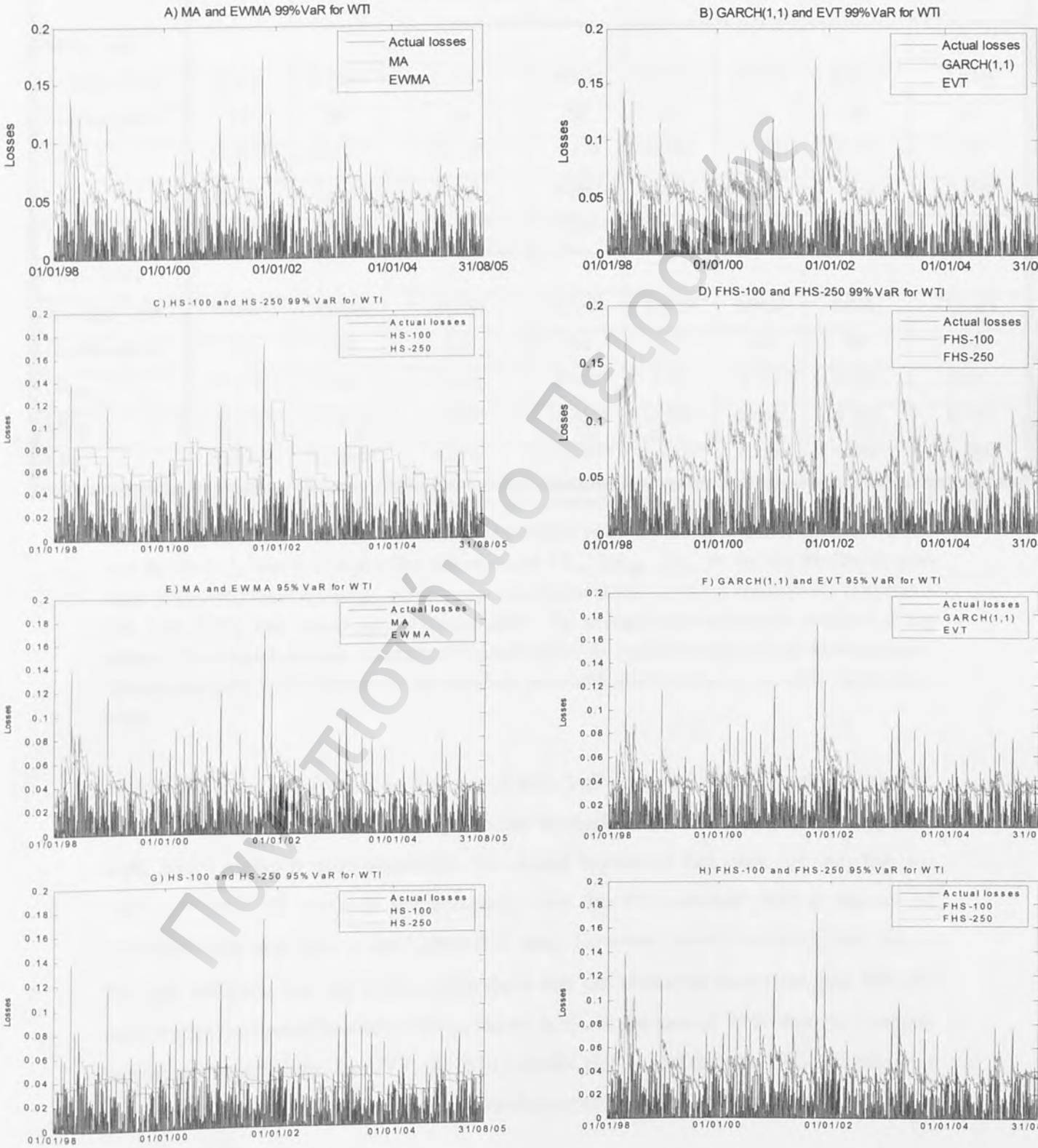


Table 4. VaR backtesting results for Brent crude oil spot

	MA	EWMA	AR(1)GARCH(1,1)	HS-100	HS-250	FHS-100	FHS-250	EVT
99% VaR								
Average VaR	0.0433	0.0430	0.0551	0.0456	0.0472	**0.0689**	**0.0719**	**0.0738**
No.of exceptions	39	40	50	25	29	22	24	26
LR_{uc}	14.292*	15.674*	32.116*	1.174	3.600*	0.197	0.763	1.667
LR_{ind}	1.469	1.338	0.402	4.286*	3.256*	1.374	1.108	0.879
LR_{cc}	15.762**	17.013**	32.519**	5.460**	6.857**	**1.572**	**1.871**	**2.546**
95% VaR								
Average VaR	0.0306	0.0304	**0.0389**	0.0293	0.0294	**0.0403**	**0.0411**	**0.0419**
No.of exceptions	106	108	113	105	113	104	107	97
LR_{uc}	0.378	0.665	1.724	0.264	1.724	0.170	0.512	0.092
LR_{ind}	10.272*	9.5538*	1.079	15.465*	11.998*	0.069	0.296	0.363
LR_{cc}	10.650**	10.219**	**2.803**	15.730**	13.722**	**0.239**	**0.809**	**0.456**

Table 4: This table shows the components of Christoffersen's test applied to our family of methods for 99% and 95% VaR estimation. The critical value for the LR_{uc} and LR_{ind} tests is 2.706 and for the LR_{cc} test it is 4.605.The abbreviations $LR_{uc,}$ LR_{ind}, LR_{cc} are for the likelihood ratio tests of unconditional coverage, independence and conditional coverage respectively (equations 3.6, 3.10, 3.12). Our out-of sample size is 2000. The average VaR values are in terms of log returns. The asterisk denotes rejection of the method for the particular test and the double asterisk denotes rejection of the method for the final test of conditional coverage at the 10% significance level.

From Table 4 we can see that in the case of 99% VaR only the two FHS methods and the EVT method are accepted, but the FHS-100 method yields the least average value for VaR, which makes it more preferable. We should mentioned that, although only the two FHS and the EVT methods are accepted, both the HS methods yield a number of exceptions that also falls in the 'green' BIS zone. However the HS methods, both fail to pass the independence test as they give quite few but clustered exceptions (the HS-250 fails to pass the unconditional coverage test as well). In the case of 95% VaR the methods that are not rejected are the EVT, the AR(1)GARCH(1,1) and the two FHS. It should be mentioned that only the EVT method yields fewer exceptions than the expected.

Figure 2. 99% and 95% VaR for Brent crude oil spot against actual losses.
In the y-axis are the negative log returns, which correspond, to losses and in the x-axis is the out of sample period. The panels from A to D present the 99% VaR plots and the panels from E to H present the 95% VaR plots.

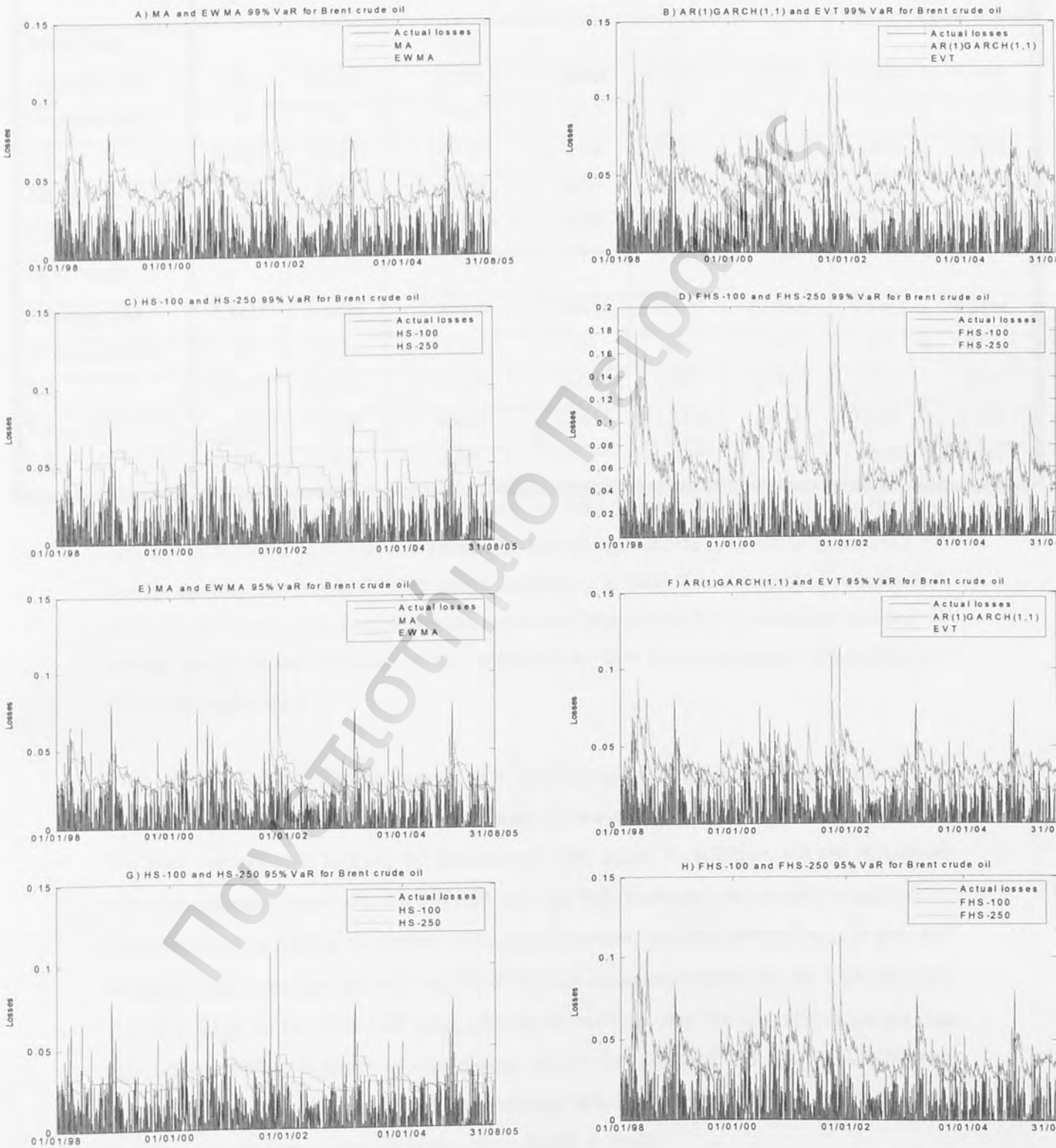


Table 5. VaR backtesting results for heating oil spot

	MA	EWMA	AR(1)GARCH(1,1)	HS-100	HS-250	FHS-100	FHS-250	EVT
99% VaR								
Average VaR	0.0639	0.0630	0.0581	**0.0661**	0.0684	**0.0664**	**0.0686**	**0.0685**
No.of exceptions	32	36	43	23	28	18	26	23
LR_{uc}	6.165*	10.466*	20.121*	0.436	2.882	0.207	1.667	0.4366
LR_{ind}	2.616	1.908	0.990	0.535	12.356*	0.327	0.685	0.535
LR_{cc}	8.782**	12.374**	21.112**	**0.972**	15.239**	**0.534**	**2.352**	**0.972**
95% VaR								
Average VaR	**0.0452**	**0.0446**	**0.0410**	0.0416	0.0419	**0.0442**	**0.0443**	**0.0441**
No.of exceptions	104	102	118	104	103	94	98	103
LR_{uc}	0.170	0.044	3.251*	0.170	0.097	0.380	0.040	0.970
LR_{ind}	**1.219**	**1.455**	**0.621**	8.843*	11.411*	0.564	0.985	1.334
LR_{cc}	**1.389**	**1.498**	**3.872**	9.014**	11.508**	**0.944**	**1.025**	**1.431**

Table 5: This table shows the components of Christoffersen's test applied to our family of methods for 99% and 95% VaR estimation. The critical value for the LR_{uc} and LR_{ind} tests is 2.706 and for the LR_{cc} test it is 4.605. Our out-of sample size is 2000. The average VaR values are in terms of log returns. The asterisk denotes rejection of the method for the particular test and the double asterisk denotes rejection of the method for the final test of conditional coverage at the 10% significance level.

From Table 5, for the case of 99% VaR we see that the accepted methods are the EVT, the HS-100 and the two FHS methods. However the HS-100 method proves to be the least conservative judging by the average VaR value. In addition, we see that along with the accepted methods, the HS-250 and the MA methods also exhibit a number of exceedances that falls in the 'green' BIS zone. However, the MA method fails to pass the unconditional coverage test and the HS-250 gives clustered exceptions, as it can be seen from the table. In the 95% VaR case, almost all methods pass the test, all except the two HS methods. What is worth of mentioning here is that although the AR(1)GARCH(1,1) method fails to pass the unconditional coverage test, it finally overcomes that failure due to the extremely low clustering in the exceedances it yields.

Figure 3. 99% and 95% VaR for Heating oil spot against actual losses.
In the y-axis are the negative log returns, which correspond, to losses and in the x-axis is the out of sample period. The panels from A to D present the 99% VaR plots and the panels from E to H present the 95% VaR plots.

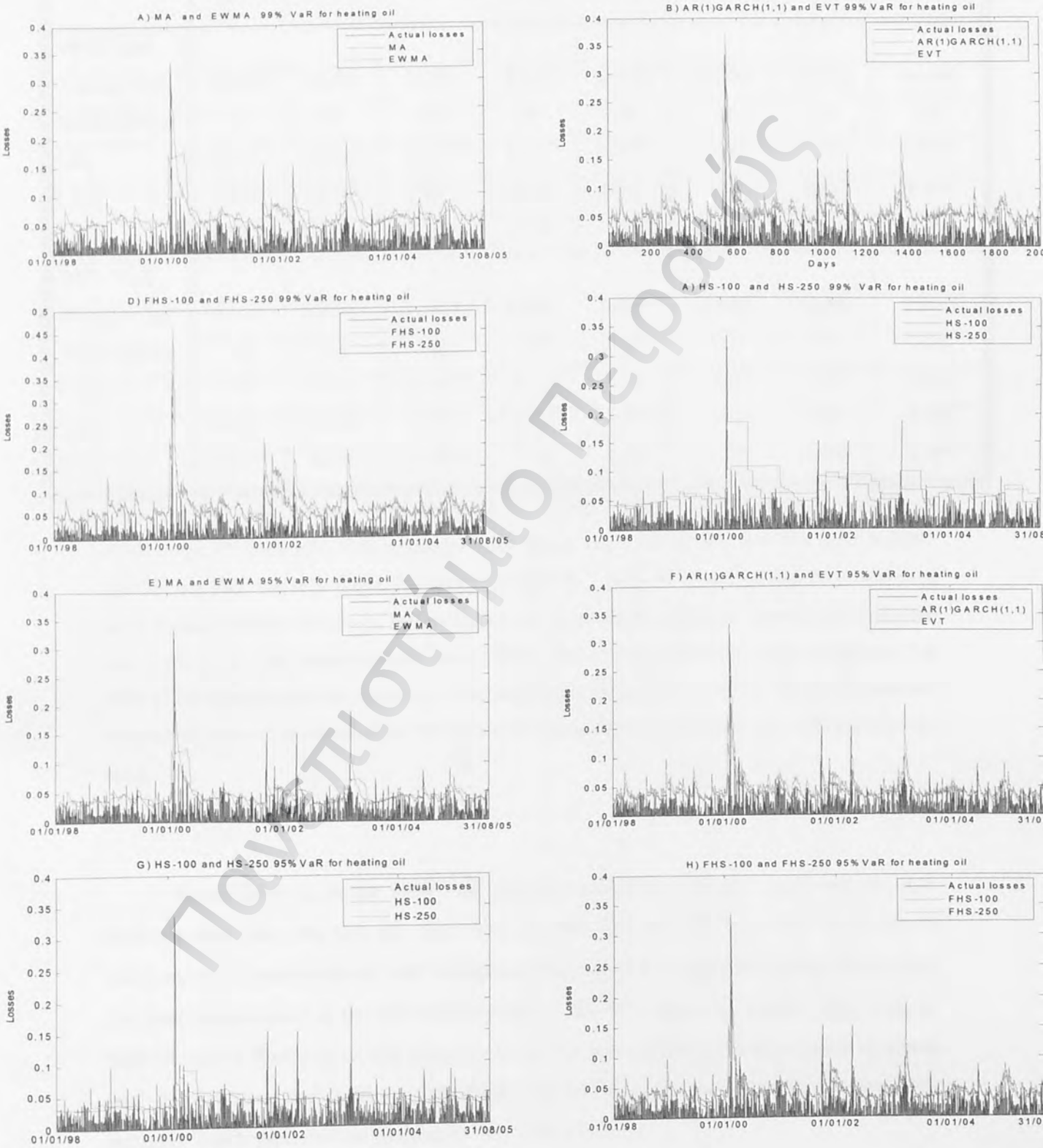


Table 6. VaR backtesting results for unleaded gasoline spot

	MA	EWMA	GARCH(1,1)	HS-100	HS-250	FHS-100	FHS-250	EVT
99% VaR								
Average VaR	0.0676	0.0674	0.0589	**0.0720**	0.0776	**0.0729**	**0.0781**	**0.0763**
No.of exceptions	47	53	68	28	33	22	24	29
LR_{uc}	26.712*	37.890*	71.655*	2.882*	7.1498*	0.197	0.763	3.600*
LR_{ind}	0.010	3.448*	2.561	0.683	0.317	1.374	0.583	0.853
LR_{cc}	26.723**	41.331**	74.217**	**3.566**	7.467**	**1.572**	**1.346**	**4.454**
95% VaR								
Average VaR	0.0478	**0.0476**	0.0417	**0.0473**	**0.0483**	**0.0492**	**0.0503**	**0.0490**
No.of exceptions	113	116	149	104	111	102	112	115
LR_{uc}	1.724	2.585	22.162*	0.170	1.243	0.044	1.474	2.279
LR_{ind}	4.527*	0.788	0.001	3.551*	1.300	1.455	0.494	0.307
LR_{cc}	6.251**	**3.374**	22.163**	**3.721**	**2.544**	**1.498**	**1.968**	**2.587**

Table 6: This table shows the components of Christoffersen's test applied to our family of methods for 99% and 95% VaR estimation. The critical value for the LR_{uc} and LR_{ind} tests is 2.706 and for the LR_{cc} test it is 4.605.The abbreviations $LR_{uc,}$ LR_{ind}, LR_{cc} are for the likelihood ratio tests of unconditional coverage, independence and conditional coverage respectively (equations 3.6, 3.10, 3.12). Our out-of sample size is 2000. The average VaR values are in terms of log returns. The asterisk denotes rejection of the method for the particular test and the double asterisk denotes rejection of the method for the final test of conditional coverage at the 10% significance level.

From Table 6, in the 99% VaR case the backtesting results show that the four methods that pass the test are the EVT, the HS-100 and the two FHS methods. In addition, only those methods yield exceptions that fall in the 'green' BIS zone. From those the least conservative is the HS-100 method. In the 95% case we observe again that no method that is finally accepted manages to give a number of exceptions equal or lower than the expected one. However the EVT, the EWMA, the two HS and the two FHS methods finally pass the test judging by their LR_{cc} values.

Figure 4. 99% and 95% VaR for Unleaded Gasoline spot against actual losses.
In the y-axis are the negative log returns, which correspond, to losses and in the x-axis is the out of sample period. The panels from A to D present the 99% VaR plots and the panels from E to H present the 95% VaR plots.

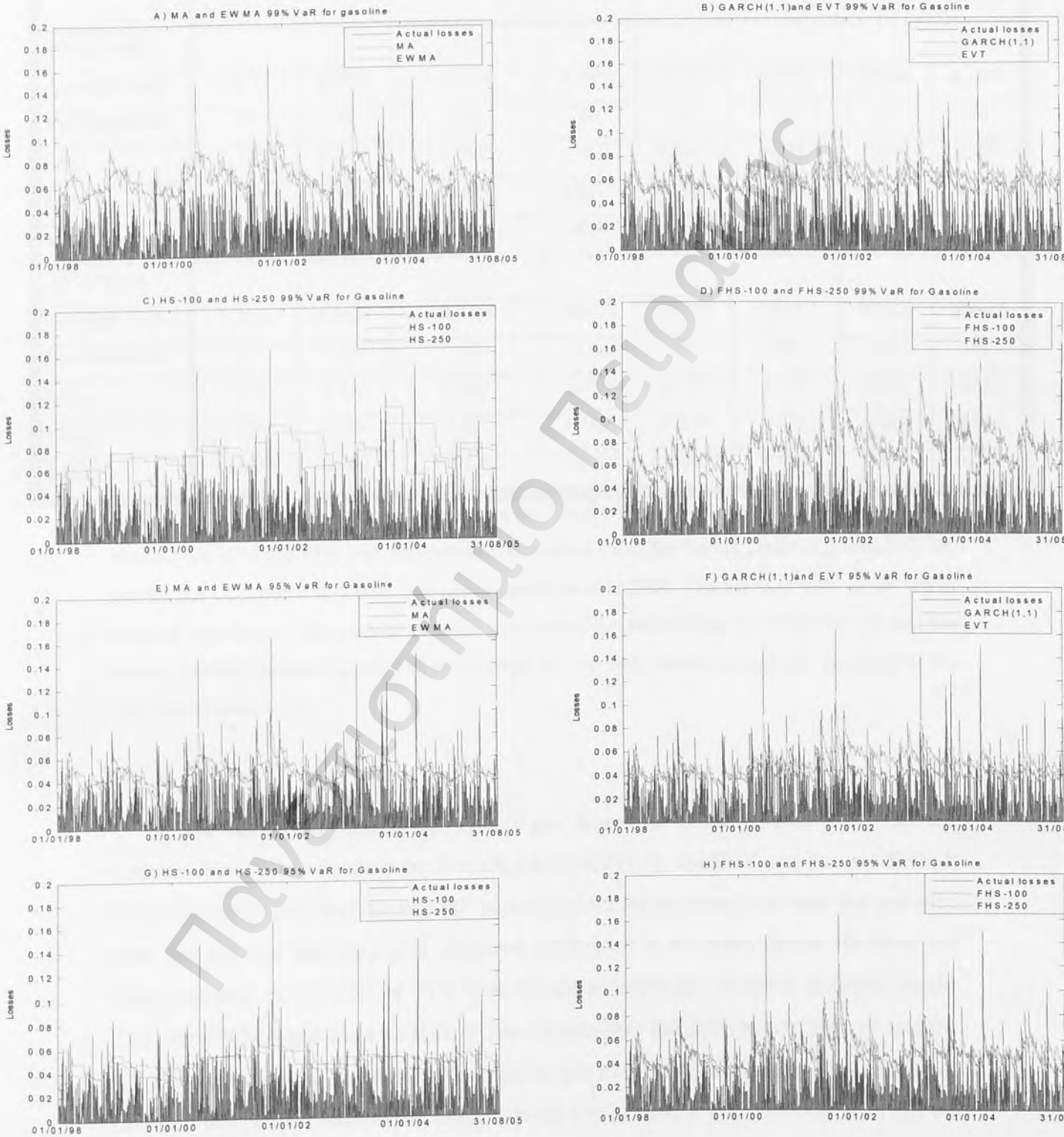


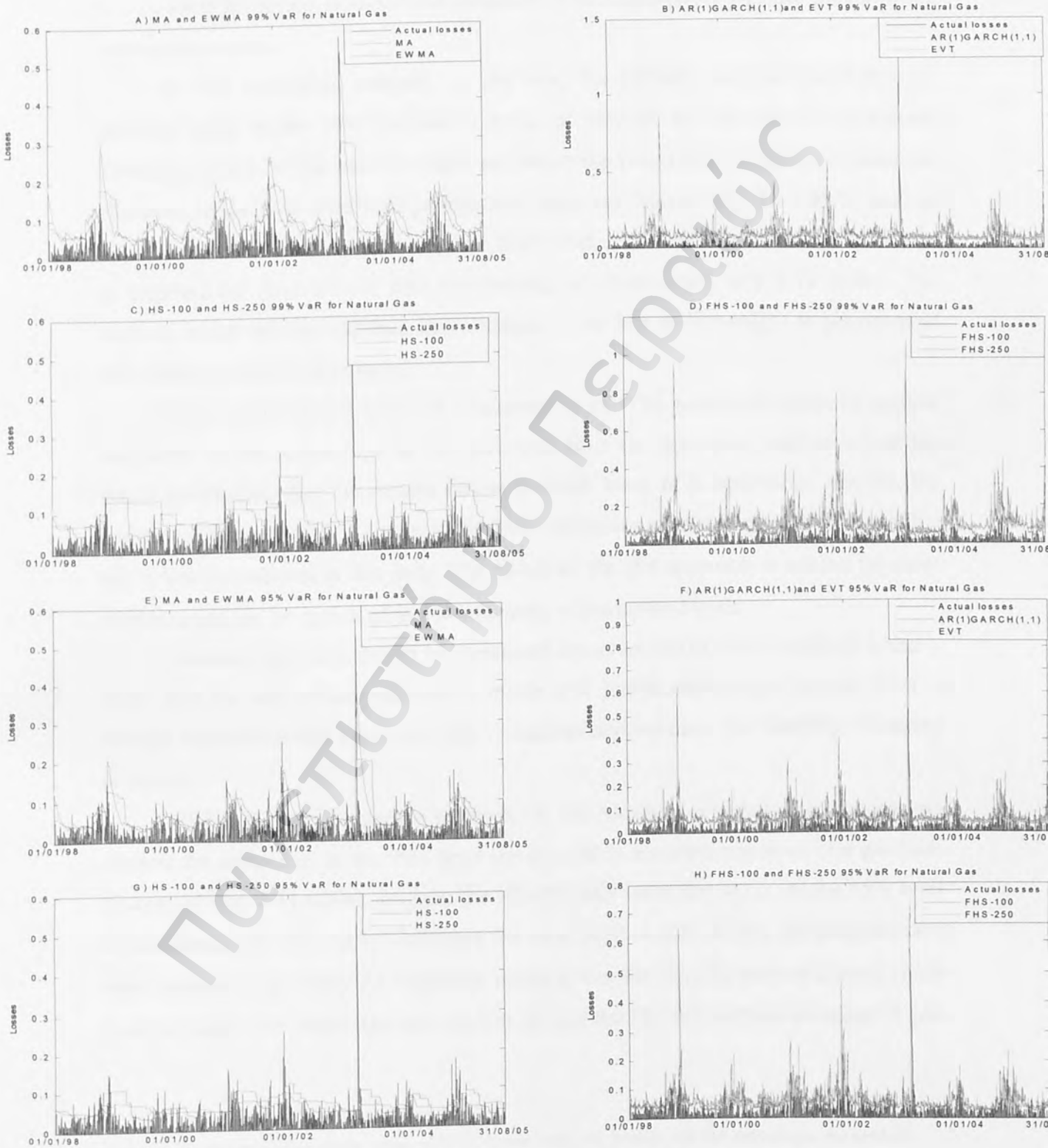
Table 7. VaR backtesting results for natural gas spot

	MA	EWMA	AR(1)GARCH(1,1)	HS-100	HS-250	FHS-100	FHS-250	EVT
99% VaR								
Average VaR	0.0937	0.0920	**0.0999**	0.0974	0.1126	**0.0974**	**0.1062**	**0.1105**
No.of exceptions	43	34	28	26	28	21	23	22
LR_{uc}	20.121*	8.196*	2.882*	1.667	2.882*	0,050	0.436	0.197
LR_{ind}	9.549*	5.397*	0.795	4.007*	12.356*	0.445	0.535	0.489
LR_{cc}	29.671**	13.593**	**3.678**	5.674**	15.239**	**0.496**	**0.972**	**0.687**
95% VaR								
Average VaR	0.0662	**0.0651**	0.0706	0.0594	0.0590	**0.0623**	**0.0618**	**0.0645**
No.of exceptions	115	96	79	113	93	98	103	94
LR_{uc}	2.279	0.166	4.905*	1.724	0.520	0.040	0.097	0.380
LR_{ind}	7.280*	2.299	2.087	6.112*	10.536*	0.854	0.020	0.045
LR_{cc}	9.560**	**2.466**	6.992**	7.836**	11.056**	**0.894**	**0.117**	**0.425**

Table 7: This table shows the components of Christoffersen's test applied to our family of methods for 99% and 95% VaR estimation. The critical value for the LR_{uc} and LR_{ind} tests is 2.706 and for the LR_{cc} test it is 4.605. Our out-of sample size is 2000. The average VaR values are in terms of log returns. The asterisk denotes rejection of the method for the particular test and the double asterisk denotes rejection of the method for the final test of conditional coverage at the 10% significance level.

The backtesting results for natural gas show that in the case of 99% VaR the methods that are not rejected are the AR(1)GARCH(1,1), the EVT and the two FHS. It should be mentioned that the two HS methods yield fewer exceptions than the expected ones, but the fact that they give clustered exceptions is the main reason for them not being accepted. In the case of 95% VaR we observe that the accepted methods are the EVT, the EWMA and the two FHS. It can be seen that the AR(1)GARCH(1,1) and the HS-250 methods yield few absolute exceptions but the first fails to pass the unconditional coverage test and the second the independence test in such a way that they both fail the conditional coverage test.

Figure 5. 99% and 95% VaR for Natural Gas spot against actual losses.
In the y-axis are the negative log returns, which correspond, to losses and in the x-axis is the out of sample period. The panels from A to D present the 99% VaR plots and the panels from E to H present the 95% VaR plots.



8.2 Remarks on VaR results

From the results obtained and presented in the tables above we have the following comments to make:

i) The parametric methods i.e. the MA, the EWMA, and the GARCH(1,1)[13] perform badly at the 99% confidence level, as they all fail the test of conditional coverage, except for the case of natural gas where the GARCH(1,1) method is accepted. However, in the 95% level their performance improves. Specifically the EWMA method is accepted for all our data set except for Brent crude, whereas the GARCH(1,1) method is accepted for three of our data set (heating oil, brent crude and WTI crude). The method, which exhibits the least improvement, is the MA as it manages to get accepted only when applied for heating oil.

The results do not come as a surprise to us. The parametric methods assume normality for the returns, but as we have seen from the descriptive statistics tests the actual distribution that the returns follow is much more of a leptokurtic one. So, the reason that these methods improve in the 95% confidence level is that the thickness of the tail is less pronounced in that level. The failure of the MA approach to exhibit the same improvement can be explained as a shortcoming of the method itself.

Another thing that should be mentioned about the GARCH(1,1) method is that it never fails the test of independence no matter if it gets finally accepted or not. That is another verification that the GARCH(1,1) successfully captures the volatility clustering of returns.

ii) For the non-parametric methods i.e. the historical simulation approaches we observe the following: In the 99% level the HS-100 is accepted tree times (for gasoline, heating oil and WTI crude), whereas the HS-250 only once (for WTI). At the 95% level of confidence HS-100 and HS-250 have the same performance as they get accepted only when applied to gasoline. An important result is that the HS-100 method always yields lower average VaR value than the HS-250. In fact the HS-100 method whenever it gets

[13] As we mentioned before in the cases of WTI, Brent crude oil, heating oil and natural gas we used an AR(1)GARCH(1,1), but for simplicity we do not mention it in these remarks.

accepted it yields the lowest average VaR value among the other accepted methods. So, the rolling window's length plays an important role only at the 99% confidence level.

iii) Moving now to the semi parametric methods and starting with the FHS we observe that both the FHS methods (FHS-100 and FHS- 250) are accepted when applied in all our data sets at both confidence levels. However, whenever they are compared in terms of average VaR values with the HS-100 and the GARCH (1,1), they prove to be more conservative. In addition, the FHS-100 method yields lower average values than the FHS-250 except for the case of natural gas at the 95% level, where the FHS-250 yields the lowest average VaR value among the other accepted methods. So generally the two FHS methods give quite equivalent results.

iv) The other semi-parametric approach i.e. the EVT produce quite interesting results. The specific method is accepted when applied in all our data sets and at both the confidence levels. It generally yields quite low average VaR values, but it never manages to give the lowest one among the other accepted methods. The fact that the EVT method is never rejected is not unexpected. It is very natural that since the distribution of results has fatter tails than the normal distribution, the tails of these distributions are best described by the GPD, which is assumed by the EVT approach.

Concluding, we can say that in the 99% confidence level the methods that have the most 'acceptances' are the EVT, the FHS-100 and FHS-250 (5/5) followed by the HS-100 (3/5), with the HS-100 yielding the lowest average VaR values whenever compared to another accepted method. At that confidence level, the methods that assume normality, as expected, perform purely.

In the 95% confidence level, the EVT, the FHS-100 and the FHS-250 method get the most 'acceptances' (5/5), followed by the EWMA method (4/5) and the GARCH(1,1) method (3/5). The GARCH(1,1) method yields the lowest average VaR values whenever it is accepted. The HS-100 method is accepted only once but at that one time yields the lowest average VaR value.

Our ultimate purpose is to determine the best VaR model. And the best model should be the one that, among the others that have passed the backtesting tests, has the lowest average VaR value. This is crucial because the lowest the VaR the less amount of capital needs to be tied up. So, for each of the petroleum products and natural gas, the best models based on their average value, for VaR calculations are shown in the following table.

Table 8. Best models for VaR computation

Spot	*VaR*	*Models*
WTI Crude oil	**99%**	HS VaR with 100 days rolling window
	95%	AR(1)GARCH(1,1) VaR
Brent Crude oil	**99%**	FHS VaR with 100 days rolling window
	95%	AR(1) GARCH(1,1) VaR
Heating oil	**99%**	HS VaR with 100 days rolling window
	95%	AR(1) GARCH(1,1) VaR
Unleaded Gasoline	**99%**	HS VaR with 100 days rolling window
	95%	HS VaR with 100 days rolling window
Natural Gas	**99%**	FHS VaR with 100 days rolling window
	95%	FHS VaR with 250 days rolling window

Table 8: The methods that have the lowest average VaR value among the other accepted methods from Christoffersen's conditional coverage test.

As we can see from Table 8, at the 99% level of confidence as expected no method that assumes normality is selected as the best, whereas at the 95% level of confidence a parametric method -the GARCH(1,1)- is selected as the best in three occasions. Finally we observe that only for unleaded gasoline exists a common method that is characterized as the best at both the confidence levels.

8.3 Conditional VaR results

Table 9. Results from backtesting the ES using the quantratic score function

Spot	99%			95%		
	Method	QS (rank)	average ES	Method	QS (rank)	average ES
				EWMA	0.0053 (2)	0.0508
				AR(1)GARCH(1,1)*	0.0048 (1)	0.0488
WTI crude oil	HS-100*	0.0118 (1)	0.0738			
	HS-250	0.0146 (3)	0.0567			
	FHS-100	0.0119 (2)	0.0644	FHS-100	0.0076 (4)	0.0589
	FHS-250	0.0156 (5)	0.0825	FHS-250	0.0075 (3)	0.0603
	EVT	0.0155 (4)	0.0866	EVT	0.0077 (5)	0.0607
				AR(1)GARCH (1,1)*	0.0048 (1)	0.0488
Brent crude oil	FHS-100*	0.0119 (1)	0.0642	FHS-100	0.0076 (4)	0.0590
	FHS-250	0.0156 (3)	0.0825	FHS-250	0.0075 (3)	0.0602
	EVT	0.0134 (2)	0.0890	EVT	0.0052 (2)	0.0506
				MA	0.0071 (3)	0.0517
				EWMA	0.0070 (2)	0.0559
Heating oil				AR(1)GARCH(1,1)*	0.0058 (1)	0.0515
	HS-100	0.0177 (3)	0.0809			
	FHS-100*	0.0137 (1)	0.0658	FHS-100	0.0094 (5)	0.0614
	FHS-250	0.0204 (4)	0.0910	FHS-250	0.0104 (6)	0.0666
	EVT	0.0158 (2)	0.0831	EVT	0.0080 (4)	0.0597
				EWMA	0.0081 (2)	0.0642
	HS-100	0.0147 (2)	0.0838	HS-100*	0.0072 (1)	0.0598
Gasoline				HS-250	0.0086 (4)	0.0666
	FHS-100	0.0153 (3)	0.0779	FHS-100	0.0096 (5)	0.0685
	FHS-250	0.0193 (4)	0.0927	FHS-250	0.0097 (6)	0.0698
	EVT*	0.0145 (1)	0.0845	EVT	0.0084 (3)	0.0656
				EWMA	0.0177 (2)	0.0598
Natural gas	AR(1)GARCH (1,1)	0.0389 (3)	0.0676			
	FHS-100*	0.0174 (1)	0.0779	FHS-100*	0.0160 (1)	0.0685
	FHS-250	0.0320 (2)	0.0927	FHS-250	0.0193 (3)	0.0698
	EVT	0.0526 (4)	0.1340	EVT	0.0251 (4)	0.0931

Table 9: In the first column there are the petroleum products and natural gas, in the second and fifth column appear the methods that have been accepted from the first backtesting stage at the 99% and 95% confidence level, respectively. In third and sixth column appear each method's score values using the formula (5.3) and in parenthesis their rank among the other methods. Finally, in fourth and seventh column we present the average value of the ES. An asterisk denotes the method that achieved the lowest score.

Summing up the results shown in Table 9, we present in the following table the models that rank first at both confidence levels and for each of the petroleum products and natural gas.

Table 10. Best methods for VaR computation based on the ES backtesting results

Spot	*VaR*	*Models*
WTI Crude oil	**99%**	HS VaR with 100 days rolling window
	95%	AR(1)GARCH(1,1) VaR
Brent Crude oil	**99%**	FHS VaR with 100 days rolling window
	95%	AR(1) GARCH(1,1) VaR
Heating oil	**99%**	FHS VaR with 100 days rolling window
	95%	AR(1) GARCH(1,1) VaR
Unleaded Gasoline	**99%**	HS VaR with 100 days rolling window
	95%	HS VaR with 100 days rolling window
Natural Gas	**99%**	FHS VaR with 100 days rolling window
	95%	FHS VaR with 100 days rolling window

Table 10: The methods that yield the lowest score function value (equation 5.3) for each data set at both confidence levels.

8.4 Remarks on CVaR Results

So, if we were to comment on the results, we would have to say that in the 99% confidence level not even one method that assumes a distribution -Normal or GPD- is selected as the best. The FHS-100 ends up first as it is characterized as best for 3/5 cases, with the HS-100 coming as second, having outperformed in 2/5 cases.

In the 95% confidence level appears quite the opposite case. The methods with the assumption of normality are characterized as best in 3/5 cases and moreover in the cases of gasoline and natural gas where the HS-100 and the FHS-100 method, respectively, are the best, the EWMA method comes second and very close in both cases.

Finally, in our search for the best model for each of the petroleum products and natural gas, we observe that only gasoline and natural gas have a method that is characterized best for both the confidence levels (the HS-100 for gasoline and the FHS-100 for natural gas).

8.5 Remarks on results from both evaluation stages

The results of the first and second evaluation stages regarding the best method for VaR computation agree in almost all cases. The exceptions to that agreement is first the case of heating oil at the 99% confidence level and second the case of natural gas at the 95% level of confidence. Specifically, the first evaluation stage propose as the best model- the model that passes the conditional coverage test and has the lowest average VaR value among the other accepted methods- for heating oil at the 99% level the HS-100 method, whereas the second evaluation stage proposes the FHS-100 method. In the case of natural gas the first stage gives us as the best method the FHS-250 whereas the second stage the FHS-100 method. It should be noted though that in both the cases where we get this disagreement, the methods that are proposed as the best from the ES backtest

rank second in the first evaluation stage as they have the second lowest average VaR values. So, we could say that although the disagreement in the results is evident for these cases, it is not harsh.

9. Results for the nearby futures

9.1 VaR Results

Tables from 11 to 14, present the results obtained for all our data sets and for all the methods employed for VaR calculation. We present the average VaR for each method and for each confidence level. In addition the number of exceptions is presented and by number of exceptions we mean the number that shows how many times the actual losses exceeded our VaR prediction. Finally the results of the likelihood ratio tests we used for backtesting are also shown. In addition, after each table, we present for all the methods employed, the figures of the calculated out-of sample 99% and 95% VaRs, respectively, against the actual losses.

Table 11. VaR backtesting results for WTI crude oil nearby futures

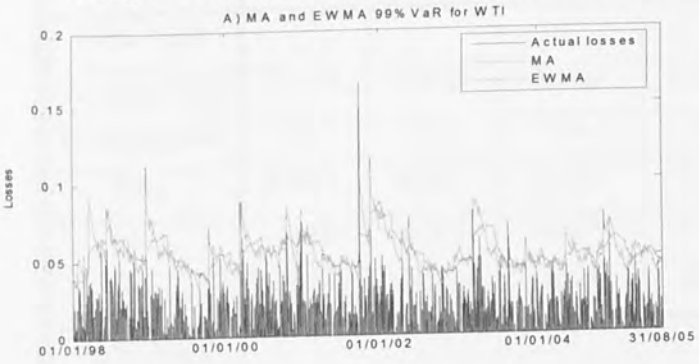
	MA	EWMA	GARCH(1,1)	HS-100	HS-250	FHS-100	FHS-250	EVT
99% VaR								
Average VaR	0.0538	0.0534	0.0522	**0.0563**	0.0577	**0.0590**	**0.0608**	**0.0600**
No.of exceptions	35	45	43	19	26	26	27	26
LR_{uc}	9.302*	23.326*	20.121*	0.050	1.667	1.667	2.237	1.667
LR_{ind}	2.071	0.792	0.990	1.857	8.349*	0.879	0.777	0.879
LR_{cc}	11.373**	24.118**	21.112**	**1.907**	10.016**	**2.546**	**3.015**	**2.546**
95% VaR								
Average VaR	**0.0380**	**0.0378**	**0.0368**	**0.0386**	0.0368	**0.0372**	**0.0379**	**0.0390**
No.of exceptions	107	105	108	104	118	100	108	93
LR_{uc}	0.512	0.264	0.665	0.170	3.251*	0.001	0.665	0.520
LR_{ind}	0.907	0.413	0.814	1.219	2.289	0.001	0.246	0.110
LR_{cc}	**1.419**	**0.678**	**1.480**	**1.389**	5.540**	**0.002**	**0.911**	**0.630**

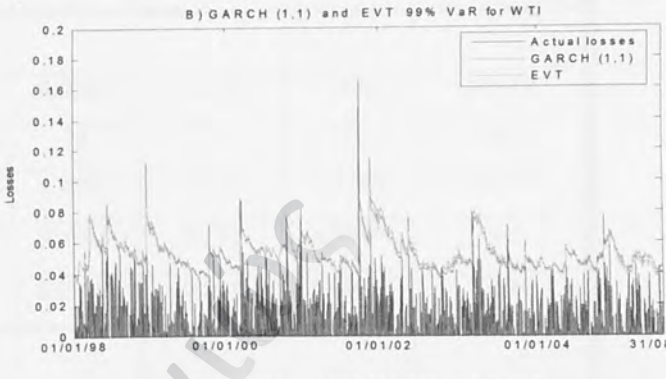
Table 11: This table shows the components of Christoffersen's test applied to our family of methods for 99% and 95% VaR estimation. The critical value for the LR_{uc} and LR_{ind} tests is 2.706 and for the LR_{cc} test it is 4.605. The abbreviations LR_{uc}, LR_{ind}, LR_{cc} are for the likelihood ratio tests of unconditional coverage, independence and conditional coverage respectively (equations 3.6, 3.10, 3.12). Our out-of sample size is 2000. The average VaR values are in terms of log returns. The asterisk denotes rejection of the method for the particular test and the double asterisk denotes rejection of the method for the final test of conditional coverage at the 10% significance level.

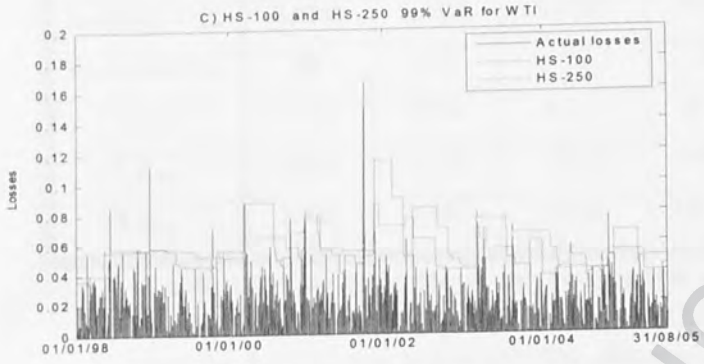
From Table 11 we see that in the 99% confidence level only the HS-100 and the two FHS methods are accepted. These methods yield a number of exceptions that falls in the 'green' BIS zone together with the HS-250 method, which though fails to pass the independence test. Moreover, the HS-100 yields the fewest exceptions and has the lowest average VaR value.

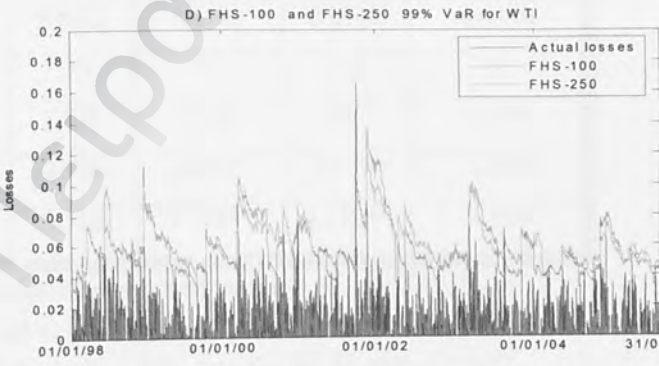
In the 99% confidence level the method that is not accepted is the HS-250. From the accepted methods the EVT and the FHS-100 yield a number of exceptions not bigger than the expected one. Finally the method that has the lowest average value among the accepted is the GARCH(1,1).

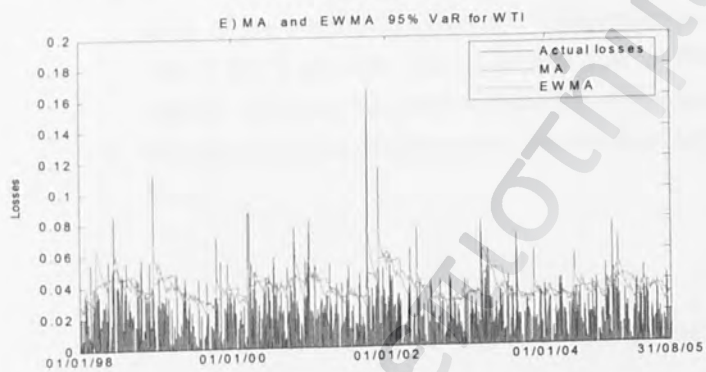
Figure 6. 99% and 95% VaR for WTI nearby futures against actual losses.
In the y-axis are the negative log returns, which correspond, to losses and in the x-axis is the out of sample period. The panels from A to D present the 99% VaR plots and the panels from E to H present the 95% VaR plots.

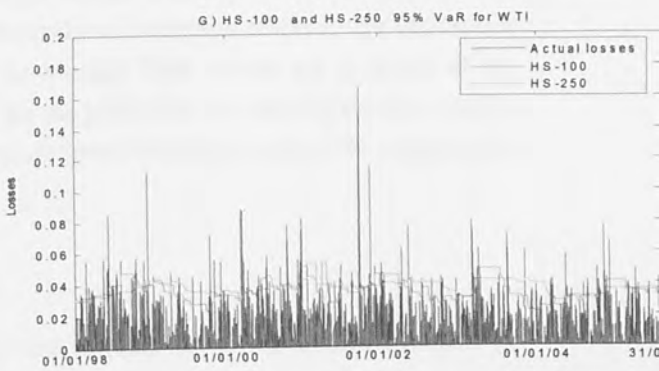


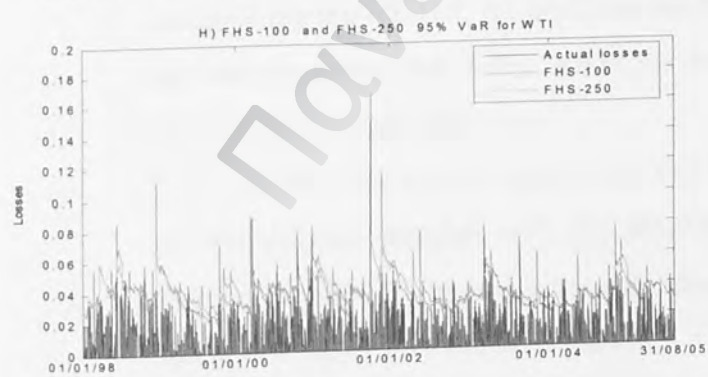












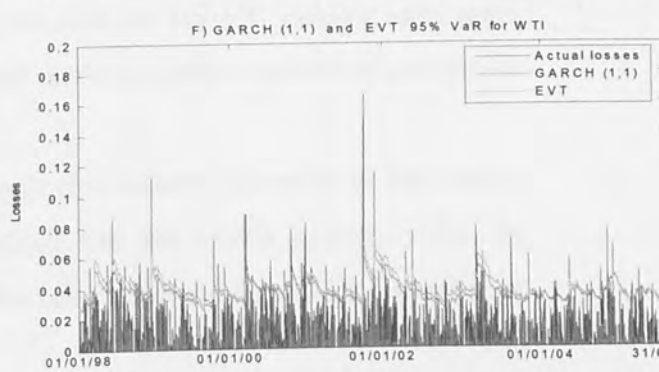


Table 12. VaR backtesting results for Brent crude oil nearby futures

	MA	EWMA	GARCH(1,1)	HS-100	HS-250	FHS-100	FHS-250	EVT
99% VaR								
Average VaR	0.0549	0.0534	0.0529	**0.0511**	0.0546	**0.0528**	0.0564	**0.0610**
No.of exceptions	32	35	34	23	28	27	25	20
LR_{uc}	6.165*	9.302*	8.196*	0.436	2.882*	2.237	1.1749	0.001
LR_{ind}	2.616	0.213	0.262	1.236	3.493*	0.777	4.286*	1.683
LR_{cc}	8.782**	9.516**	8.458**	**1.673**	6.376**	**3.015**	5.460**	**1.684**
95% VaR								
Average VaR	**0.0388**	**0.0378**	**0.0374**	0.0339	**0.0362**	0.0332	0.0340	**0.0370**
No.of exceptions	95	87	95	104	113	111	110	97
LR_{uc}	0.262	1.843	0.262	0.170	1.724	1.243	1.031	0.092
LR_{ind}	2.461	0.191	3.910	6.863*	1.994	8.534*	3.739*	3.497
LR_{cc}	**2.723**	**2.034**	**4.172**	7.033**	**3.718**	9.778**	4.771**	**3.589**

Table 12: This table shows the components of Christoffersen's test applied to our family of methods for 99% and 95% VaR estimation. The critical value for the LR_{uc} and LR_{ind} tests is 2.706 and for the LR_{cc} test it is 4.605. The abbreviations LR_{uc}, LR_{ind}, LR_{cc} are for the likelihood ratio tests of unconditional coverage, independence and conditional coverage respectively (equations 3.6, 3.10, 3.12). Our out-of sample size is 2000. The average VaR values are in terms of log returns. The asterisk denotes rejection of the method for the particular test and the double asterisk denotes rejection of the method for the final test of conditional coverage at the 10% significance level.

From Table 12 we see that in the 99% confidence level the methods that are accepted are the HS-100, the EVT and the FHS-100 with the HS-100 yielding once again the lowest average VaR value. All of the accepted methods yield a number of exceptions that fall in the 'green' BIS zone.

In the 95% level of confidence the methods that assume normality of the returns are not rejected together with the HS-250, which has the lowest average value. In addition the HS-250 method is the only among the other excepted that yields a number of exceptions bigger than the expected one.

Figure 7. 99% and 95% VaR for Brent crude oil nearby futures against actual losses. In the y-axis are the negative log returns, which correspond, to losses and in the x-axis is the out of sample period. The panels from A to D present the 99% VaR plots and the panels from E to H present the 95% VaR plots.

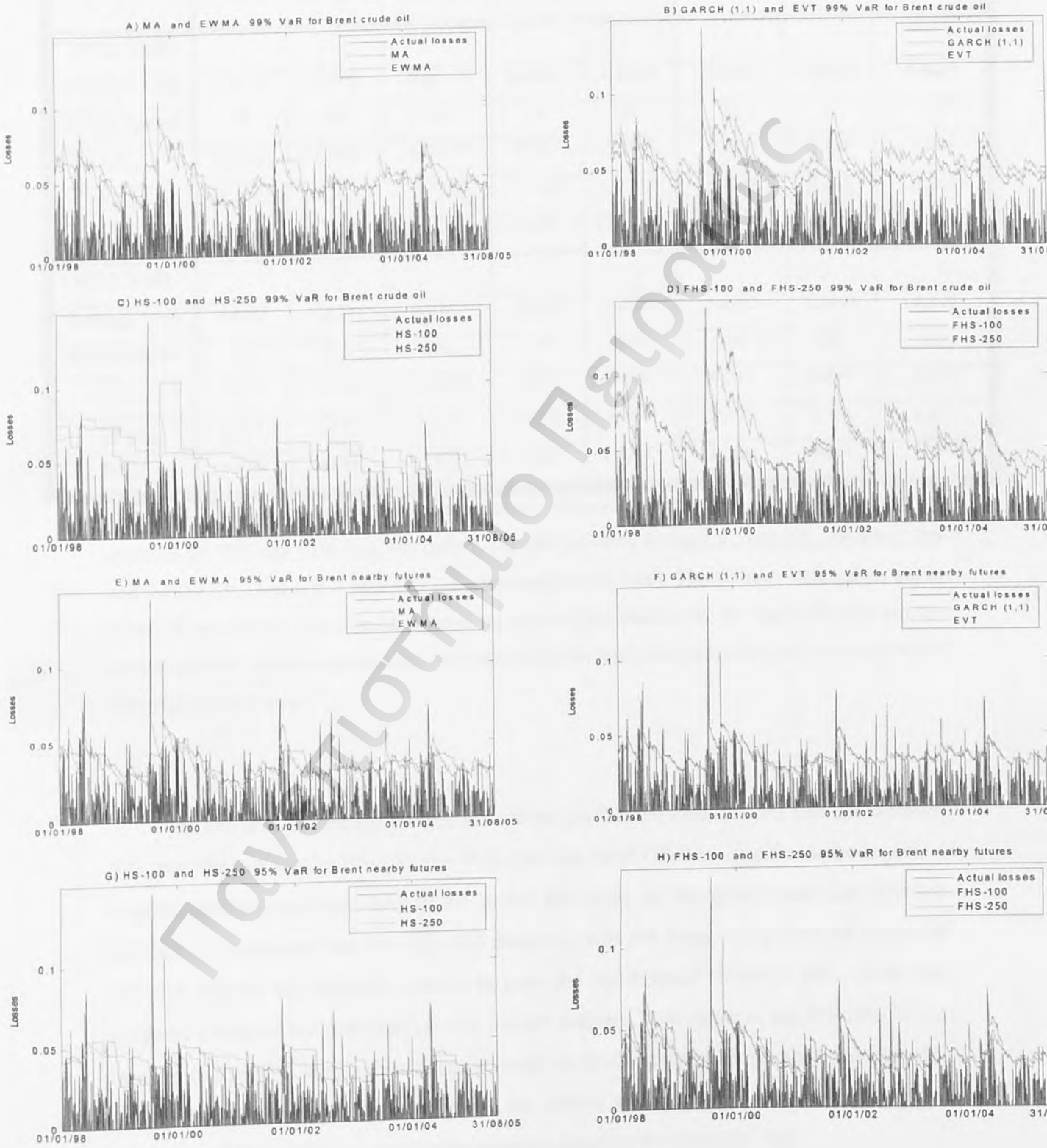


Table 13. VaR backtesting results for heating oil nearby futures

	MA	EWMA	GARCH(1,1)	HS-100	HS-250	FHS-100	FHS-250	EVT
99% VaR								
Average VaR	0.0555	0.0553	0.0519	**0.0583**	0.0634	0.0614	**0.0636**	**0.0647**
No.of exceptions	38	37	45	24	31	22	23	19
LR_{uc}	12.963*	11.687*	23.326*	0.763	5.244*	0.197	0.436	0.050
LR_{ind}	1.607	1.753	2.724	1.108	6.378 *	5.223*	0.535	0.364
LR_{cc}	14.571**	13.441**	26.050**	**1.871**	11.623**	5.421**	**0.972**	**0.415**
95% VaR								
Average VaR	**0.0393**	**0.0391**	0.0367	**0.0373**	0.0377	**0.0393**	**0.0404**	**0.040**
No.of exceptions	107	115	126	100	108	101	103	102
LR_{uc}	0.512	2.279	6.625*	0.001	0.665	0.011	0.097	0.044
LR_{ind}	2.977	1.714	0.567	2.926	4.164*	0.708	0.550	1.455
LR_{cc}	**3.498**	**3.994**	7.192**	**2.927**	4.830**	**0.720**	**0.647**	**1.499**

Table 13: This table shows the components of Christoffersen's test applied to our family of methods for 99% and 95% VaR estimation. The critical value for the LR_{uc} and LR_{ind} tests is 2.706 and for the LR_{cc} test it is 4.605. Our out-of sample size is 2000. The average VaR values are in terms of log returns. The asterisk denotes rejection of the method for the particular test and the double asterisk denotes rejection of the method for the final test of conditional coverage at the 10% significance level.

In the case of heating oil and for a 99% confidence level we see that the methods that pass the test are the HS-100, the FHS-250 and the EVT. The specific methods yield a number of exceptions that falls in the 'green' BIS zone. In the 'green' zone also falls the number of exceptions that the FHS-100 method yields but those exceptions are clustered and that forbids the FH-100 method to pass the conditional coverage test. From the accepted methods the one that has the lowest average VaR value is the HS-100. When moving to the 95% level of confidence, more methods manage to get accepted but with the HS-100 method yielding once again the lowest average value as well as being the only method that yields a number of exceptions equal to the expected one.

Figure 8. 99% and 95% VaR for Heating oil nearby futures against actual losses.
In the y-axis are the negative log returns, which correspond, to losses and in the x-axis is the out of sample period. The panels from A to D present the 99% VaR plots and the panels from E to H present the 95% VaR plots

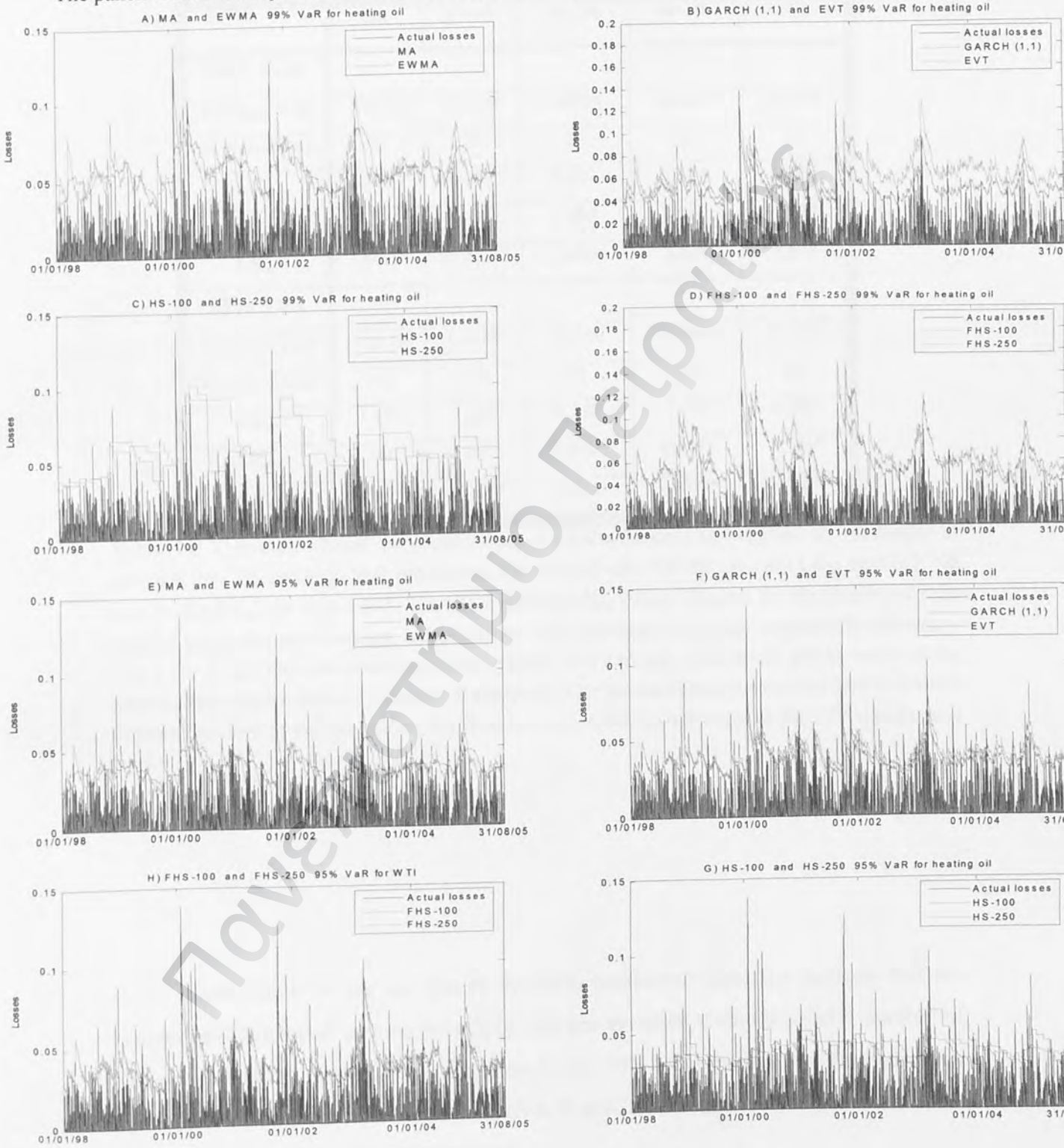


Table 14. VaR backtesting results for unleaded gasoline nearby futures

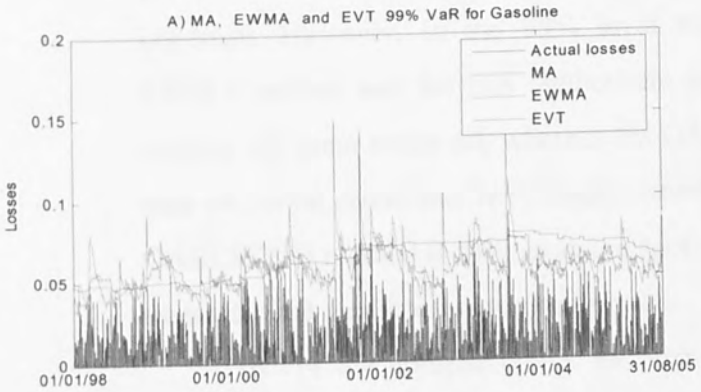
	MA	EWMA	HS-100	HS-250	EVT
99% VaR					
Average VaR	0.0579	0.0578	**0.0598**	**0.0625**	**0.0628**
No.of exceptions	33	39	20	25	24
LR_{uc}	7.149 *	14.292*	0.001	1.174	0.763
LR_{ind}	0.317	1.469	1.683	0.989	1.108
LR_{cc}	7.467**	15.762**	**1.684**	**2.164**	**1.871**
95% VaR					
Average VaR	0.0410	0.0409	**0.0396**	0.0408	0.0385
No.of exceptions	112	114	93	112	121
LR_{uc}	1.474	1.992	0.520	1.474	4.385*
LR_{ind}	3.340*	2.965*	2.803*	6.387*	4.161*
LR_{cc}	4.814**	4.958**	**3.324**	7.861**	8.546**

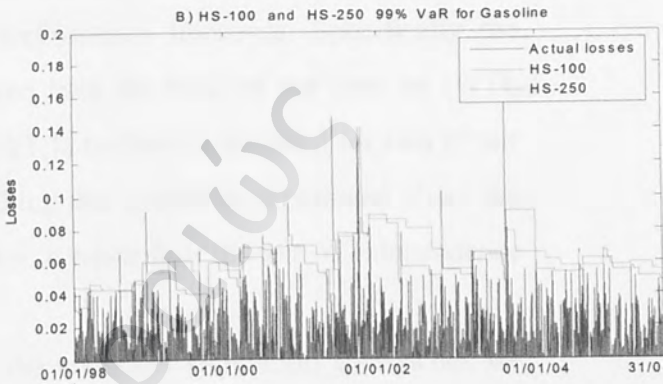
Table 14: This table shows the components of Christoffersen's test applied to our family of methods for 99% and 95% VaR estimation. The critical value for the LR_{uc} and LR_{ind} tests is 2.706 and for the LR_{cc} test it is 4.605. The abbreviations LR_{uc}, LR_{ind}, LR_{cc} are for the likelihood ratio tests of unconditional coverage, independence and conditional coverage respectively (equations 3.6, 3.10, 3.12). Our out-of sample size is 2000. The average VaR values are in terms of log returns. The asterisk denotes rejection of the method for the particular test and the double asterisk denotes rejection of the method for the final test of conditional coverage at the 10% significance level.

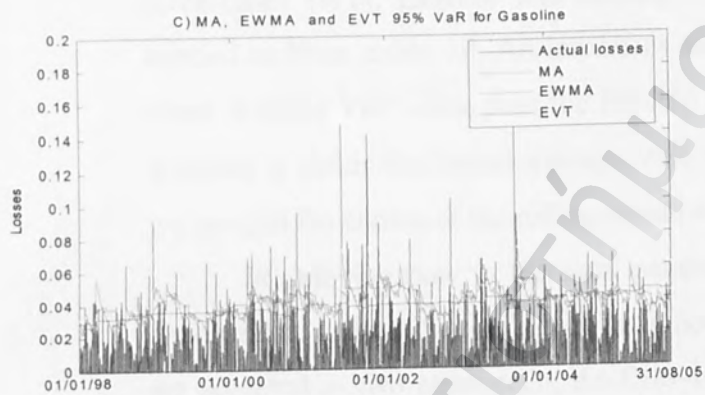
From Table 14 we see that in the 99% confidence level the methods that are rejected are those that assume normality. All the accepted methods yield a number of exceptions that fall in the 'green' BIS zone. In the 95% level from all the methods, only the HS-100 manages to pass the test, which is in addition the method that yields a number of exceptions lower than the expected one.

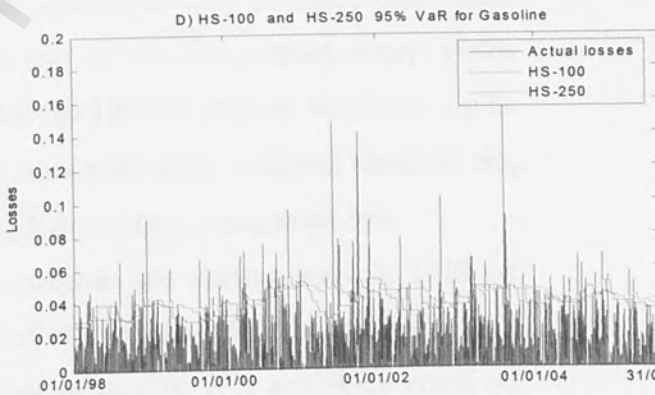
Figure 9. 99% and 95% VaR for Unleaded Gasoline nearby futures against actual losses.

In the y-axis are the negative log returns, which correspond, to losses and in the x-axis is the out of sample period. The panels from A and B present the 99% VaR plots and the panels C and D present the 95% VaR plots









9.2 Remarks on VaR results

i) The parametric methods i.e. the MA, the EWMA, and the GARCH(1,1) perform badly at the 99% confidence level, as they all fail the test of conditional coverage. However, in the 95% level their performance improves. Specifically the EWMA method and the MA method are accepted both for three of our data set (WTI, heating oil, brent crude oil, whereas the GARCH(1,1) method is accepted for two of our data set (brent crude and WTI crude). Another thing that should be mentioned about the GARCH(1,1) method is that, as in the spot market, it never fails the test of independence no matter if it gets finally accepted or not.

ii) For the non-parametric methods i.e. the historical simulation approaches we observe the following: In the 99% level the HS-100 is always accepted, whereas the HS-250 only once (for gasoline). At the 95% level of confidence the HS-100 is accepted in three cases (WTI, gasoline and heating oil) and HS-250 is only accepted only when applied to brent crude oil. An important result is that the HS-100 method always yields lower average VaR value than the HS-250. In fact the HS-100 method whenever it gets accepted it yields the lowest average VaR value among the other accepted methods. So, we see that the choice of the rolling window's length does play an important role.

iii) Moving now to the semi parametric methods and starting with the FHS we observe that at the 99% confidence level both the FHS methods (FHS-100 and FHS- 250) are accepted in two cases each, the FHS-100 when applied to WTI and Brent crude oil and the FHS-250 to Brent crude and heating oil. At the 95% level of confidence both the FHS-100 and the FHS-250 methods are accepted for WTI and heating oil. However, whenever they are compared in terms of average VaR values with the HS-100 method, they prove to be more conservative. In addition, the FHS-100 method yields lower average values than the FHS-250. So generally the two FHS methods give equivalent results and the choice of the rolling window's length does not play an important role.

iv) As far as the other semi-parametric approach i.e. the EVT is concerned, we make the following observations: The specific method is accepted when applied in all our data sets and for both the confidence levels, the exception is the case of gasoline at the 95% confidence level. It generally yields quite low average VaR values, but it never

manages to give the lowest one among the other accepted methods. The fact that the EVT method is almost never rejected and that it gets the most acceptances among the other methods is not unexpected. As in the case of the spot data set, similarly in the futures data sets this is very natural that since the distribution of results has fatter tails than the normal distribution, the tails of these distributions are best described by the GPD, which is assumed by the EVT approach.

Concluding, and having in mind all the results of all the methods when applied to all the petroleum products[14] we have the following comments to make: At the 99% confidence level the methods that have the most 'acceptances' are the EVT and the HS-100 (3/3) followed by the FHS-100 (2/3), with the HS-100 yielding the lowest average VaR values whenever compared to another accepted method. At that confidence level, the methods that assume normality, as expected, perform purely.

In the 95% confidence level, the EVT, the EWMA and the MA methods get the most 'acceptances' (3/3), followed by the HS-100, the FHS-100 and the FHS-250 methods, which get accepted two out of three times (2/3). The HS-100 method yields the lowest average VaR values whenever it is accepted among the other accepted methods.

[14] All except for gasoline, comments on which in this paragraph are not included. That is so because not all methods were applied in gasoline and in this paragraph we want to make a comparable analysis for the methods.

So, as in the case of the spot data, our ultimate purpose for the nearby futures data is to determine the best VaR model as well. And the best model should be the one that, among the others that have passed the backtesting stage, has the lowest average VaR value. So, for each of the petroleum products the best models based on their average value, for VaR calculations are shown in the following table.

Table 15. Best models for VaR computation

Nearby futures	*VaR*	*Models*
WTI Crude oil	**99%**	HS VaR with 100 days rolling window
	95%	GARCH(1,1) VaR
Brent Crude oil	**99%**	HS VaR with 100 days rolling window
	95%	HS VaR with 250 days rolling window
Heating oil	**99%**	HS VaR with 100 days rolling window
	95%	HS VaR with 100 days rolling window
Unleaded Gasoline	**99%**	HS VaR with 100 days rolling window
	95%	HS VaR with 100 days rolling window

Table 15: The methods that yield the lowest average VaR value among the other accepted methods from Christoffersen's conditional coverage test.

We can see from Table 15 that at both confidence levels the non-parametric methods outperform. Specifically, at the 99% level of confidence as expected no method that assumes normality is selected as the best. But at the 95% level the non-parametric methods outperform as well, in fact only once a parametric method is selected as the best. Moreover we can see that from the non-parametric methods, the HS-100 method outperforms in most of the cases.

Finally we observe that only for heating oil and unleaded gasoline exists a common method that is characterized as the best at both the confidence levels.

9.3 CVaR Results

Table 16. Results from backtesting the ES using the quantratic score function

Nearby futures	99%			95%		
	Method	**QS (rank)**	**average ES**	**Method**	**QS (rank)**	**average ES**
WTI crude oil				**MA**	0.00450 (2)	0.0045
				EWMA	0.00453 (3)	0.0473
				GARCH(1,1)*	0.00425 (1)	0.0463
	HS-100	0.0091 (2)	0.0647	**HS-100**	0.00482 (4)	0.0490
	FHS-100	0.0097 (3)	0.0591	**FHS-100**	0.00650 (7)	0.0537
	FHS-250	0.0125 (4)	0.0751	**FHS-250**	0.00619 (6)	0.0544
	EVT*	0.0086 (1)	0.0647	**EVT**	0.00600 (5)	0.0545
Brent crude oil				**MA**	0.00143 (2)	0.0231
				EWMA	0.00144 (3)	0.0225
				GARCH(1,1)*	0.00128 (1)	0.0223
	HS-100	0.0077 (2)	0.0583			
				HS-250	0.00474 (4)	0.0491
	FHS-100*	0.0073 (1)	0.0489			
	EVT	0.0122 (3)	0.0720	**EVT**	0.00670 (5)	0.0530
Heating oil				**MA***	0.0048 (1)	0.0492
				EWMA	0.0096 (6)	0.0693
	HS-100*	0.0093 (1)	0.0656	**HS-100**	0.0052 (2)	0.0508
				FHS-100	0.0062 (4)	0.0540
	FHS-250	0.0121 (2)	0.0724	**FHS-250**	0.0066 (5)	0.0569
	EVT	0.0122 (3)	0.0772	**EVT**	0.0059 (3)	0.0545
Gasoline	**HS-100***	0.0108 (1)	0.0705	**HS-100**	0.0057	0.0535
	HS-250	0.0134 (2)	0.0803			
	EVT	0.0175 (3)	0.0934			

Table 16: In the first column there are the petroleum products, in the second and fifth column appear the methods that have been accepted from the first backtesting stage at the 99% and 95% confidence level, respectively. In third and sixth column appear each method's score values using the formula (5.3) and in parenthesis their rank among the other methods. Finally, in fourth and seventh column we present the average value of the ES. An asterisk denotes the method that achieved the lowest score.

Summing up the results presented in Table 16, we demonstrate the best methods for VaR calculation as suggested by the second stage of evaluation.

Table 17. Best models for VaR computation based on the ES backtesting results

Nearby futures	*VaR*	*Models*
WTI Crude oil	**99%**	EVT
	95%	GARCH(1,1)
Brent Crude oil	**99%**	FHS with 100 days rolling window
	95%	GARCH(1,1)
Heating oil	**99%**	HS with 100 days rolling window
	95%	MA
Unleaded Gasoline	**99%**	HS with 100 days rolling window
	95%	HS with 100 days rolling window

Table 17: The methods that yield the lowest score function value (equation 5.3) for each data set and at both confidence levels.

9.4 Remarks on CVaR Results

We observe that in the 99% confidence level not even one parametric method is selected as the best. From the semi-parametric approaches the EVT in the case of WTI, and the FHS-100 in the case of Brent crude oil rank first, while the HS-100 ends up first as it is characterized as best for 2/4 cases (for heating oil and unleaded gasoline). Again we have to say that the underachievement of the methods that assume normality was quite expected due to the fat tails that the log returns' distribution exhibited.

In the 95% confidence level appears quite the opposite case. The methods with the assumption of normality are characterized as the best in 3/4 cases. Specifically, the GARCH(1,1) method for WTI and Brent crude oil and the MA method for heating oil. The fact that the thickness of the tail is less pronounced at the 95% level of confidence could be a reasonable explanation for the improvement of the parametric methods.

Finally, in our search for the best model for each of the petroleum products, we observe that only gasoline has a method that is characterized best for both the confidence levels (the HS-100).

9.5 Remarks on results from both evaluation stages

So far we have chosen the best models for VaR computation for each of the petroleum products' nearby futures and for each confidence level as proposed by the two evaluation stages. We observe though that the two stages propose different methods for the same data set. In fact only for gasoline we have a commonly proposed method-the HS-100-. In all the other cases the results are quite different. In fact the methods that rank first using the score function have average VaR values quite larger than the methods that rank lower judging by their score function value. So, in these cases we will have to choose between the two approaches if we want to end up with just one method for VaR computation that we can characterize it as the best.

10. Conclusion

We computed VaR and ES for WTI crude oil, Brent crude oil, heating oil, unleaded gasoline and natural gas spot market prices and for WTI crude oil, Brent crude oil, heating oil and unleaded gasoline nearby futures. We used parametric, non-parametric and semi-parametric methods to do that. In our effort to choose the best method we used two techniques and checked if the results were similar. Firstly, we used the backtesting approach proposed by Christoffersen and among the accepted methods we considered as the best the one with the lowest average VaR value. Secondly, for the accepted methods with the help of a loss function, we found their score function values and the one with the lowest value was considered as the best. The results of both of these evaluation stages were indeed similar in the case of spot market data but that wasn't true for the nearby futures.

In the question which approach we should follow in order to decide on the best method in each case for VaR computation the answer is not straightforward. All the methods that passed the conditional coverage test are for sure candidates. So selecting among them as the best the one that yields the lowest average VaR value is surely a good choice, since the selected method would have good forecasting power and we wouldn't have to tie up a large amount of capital. On the other hand if we use the approach with the score function value we will end up with a method that has a good forecasting power and in advance predicts the loss we would suffer if a VaR violation occurs. At the end we prefer to find a method that accurately predicts both the VaR and the ES measures even if that means that we would have to tie up a larger amount of capital. So, our final decision on the approach we should use in order to find the best method for VaR computation relies upon the trust we have for the loss function we use. If we trust the loss function that we define, then the approach that uses the score function should be preferred.

Summing up, we can say that although spot and nearby future prices are highly correlated the results concerning the best model for VaR computations are not the same. In addition, no method proved to be the best for all the data sets examined at both confidence levels, with one exception, the HS-100 for unleaded gasoline at both

confidence levels for both spot and futures. Moreover we see that our results concede with the results of previous studies if we consider only Kupiec's backtesting criterion as these studies did. But using more sophisticated backtesting methods, as we did, the results are quite different but more reliable.

In addition, we observe that the size of the rolling sample plays an important role and specifically the smaller one yields better results. Furthermore, an important result is that as we move to higher confidence levels (from 95% to 99%) the performance of the methods that assume normality for the log returns decrease. Finally, the EVT method is the most controversial one as the backtesting test always accepts it, but almost never does it yield the lowest average VaR value, nor the lowest score function value. Nonetheless, it may never rank first judging by those values, but in almost all cases it ranks either second or third.

Concluding we can say that many steps can be made in order to drive the price risk quantification of petroleum products and their futures a little further. The use of a GARCH model with Student-t innovations is a first step that may yield significant results and, the further exploration of the EVT method. This could be done by figuring out a way of calculating the threshold value that is more robust from a mathematical point of view.

We have no doubt that from a financial and a risk management aspect, the oil environment will prove to be a very hot topic and many of the next studies will focus on that.

References

1) Angelidis, T. and Degiannakis,S (2006). "Backtesting VaR Models: An Expected Shortfall Approach". *Working paper.*

2) Angelidis, T., Degiannakis,S and Benos A. (2004). 'The use of GARCH models in VaR estimation'. *Statistical methodology* 1(2), 105-128.

3) Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D. (1999)." Coherent Measures of Risk".*Mathematical Finance* 9, 203-228.

4) Bali, T. G. (2003). "An extreme value approach to estimating volatility and value at risk". *Journal of Business* 76: 83-108.

5) Barone-Adesi, G., Giannopoulos, K. and Vosper, Les. (2000)."Filtering historical simulation. Backtest analysis". *Working paper,* University of Westminster, March.

6) Bollerslev, T. (1986). "Generalized autoregressive conditional heteroskedasticity". *Journal of Econometrics* 31: 307-327.

7) Cabedo, J. David, Moya Ismael (2000)." Estimating oil price Value at Risk using the historical simulation approach". *Working paper.*

8) Christoffersen, F. Peter (2003). "Elements of financial risk management". *Academic press.*

9) Christoffersen, P. (1998). "Evaluating interval forecasts". International Economic Review 39: 841- 862.

10) Christoffersen, P. and F. X Diebold. (2000) "How relevant is volatility forecasting for financial risk management?".*Review of economics and statistics* 82: 12-22.

11) Danielsson, J. , De Vries, C.(1997)."Value at Risk and extreme events". Risk 85-106

12) Dowd, K. (2005). "Measuring Market Risk". *John Wiley & Sons Ltd.*, New York.

13) Engle, R. F. (1982). "Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation". *Econometrica* 50: 987-1008.

14) Finglewski S. (1997). "Forecasting volatility using historical data". New York University *Working paper no.13*

15) Geman H. (2005)."Commodities and commodity derivatives".*John Wiley & Sons Ltd.*

16) Gencay, R. and Selcuk, F. (2004)." Extreme value theory and Value-at-Risk: Relative performance in emerging markets". Forthcoming in International Journal of Forecasting.

17) Giot, P., Laurent, S., (2003b). "Market risk in commodity markets: a VaR approach". *Energy Economics* 25: 435 – 457

18) Hendricks, D. (1996). "Evaluation of Value-at-Risk models using historical data". Economic Police Review 2, 39-70.

19) Hoppe, R. (1998) "VaR and the unreal world". Risk 11 (July): 45-50.

20) Hull, J., and White, A.(1998). "Incorporating volatility updating into the historical simulation method for VaR". Journal of Risk 1: 5-19.

21) J.P. Morgan. 1995. RiskMetrics-technical manual. Third edition

22) Krehbiel, T and Lee C. Adkins, (2003). "Price risk in the NYMEX energy complex: an EVT approach". Ok;ahoma state University *Working paper*.

23) Kupiec, P.H. (1995). "Techniques for verifying the accuracy of risk measurement models". *The Journal of Derivatives* 3: 73-84.

24) Lambadiaris, G., Papadopoulou, L., Skiadopoulos, G. and Zoulis, Y. (2003). "VaR: history or simulation?" Risk 16 (September): 122-127.

25) Lopez, J.A. (1998). "Methods for evaluating Value-at-Risk estimates". Federal Reserve Bank of New York, Economic Policy Review.

26) McNeil, A.J. (1998) 'Calculating quantile risk measures for financial return series using extreme value theory.' Mimeo. ETHZ Zentrum, Zurich.

27) McNeil, A.J. and Frey, R. (2000)." Estimation of tail-related risk measures for heteroskedasticity financial time series: An extreme value approach". *Journal of Empirical Finance* 7: 271-300.

28) McNeil, Alexander J. (1999) "Extreme Value Theory for Risk Managers," Working Paper ETH Zurich..

29) Neftci, Salih N. (2000) "Value at risk Calculations, Extreme Events, and Tail Estimation," *Journal of Derivatives*, vol. 7, no. 3. 23-38.

30) Pritsker, M. (1999). "The hidden dangerous of historical simulation". Mimeo, Federal Reserve Board, Washington DC.

31) Sarma, M., Thomas, S. and Shah, A. (2003). 'Selection of VaR models.' *Journal of Forecasting* 22: 337-358.

32) Skiadopoulos G., Chantziara T. (2005). 'Can the dynamics of the term structure of petroleum futures be forecasted? Evidence from major markets'.

33) Tsay, S. Ruey (2002) " Analysis of financial time series". *John Wiley & Sons Ltd.*

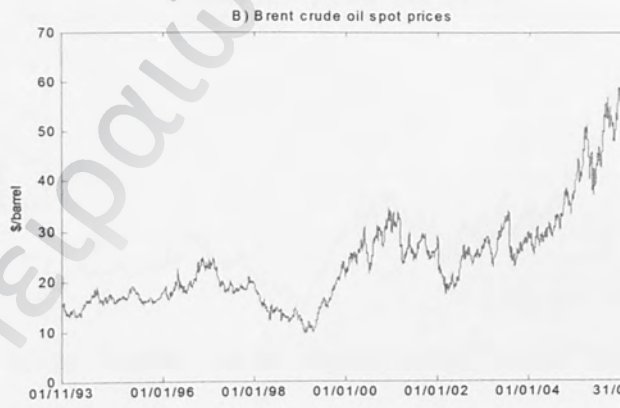
34) Van den Goorbergh, R.W.J. and Vlaar, P. (1999). "Value-at-Risk analysis of stock returns. Historical simulation, variance techniques or tail index estimation?" DNB Staff Reports 40, Netherlands Central Bank.

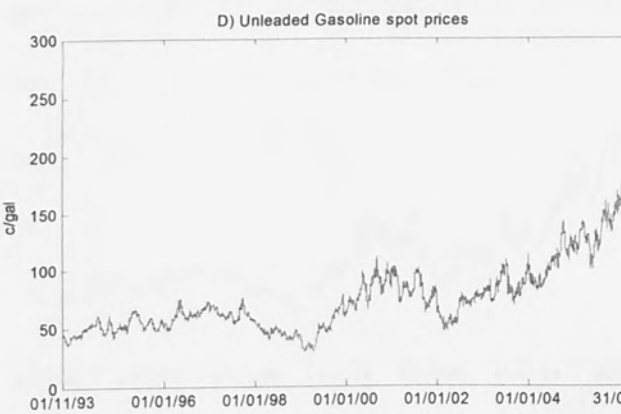
35) Vlaar, P (2000). "Value at Risk models for Dutch bond portfolios". *Journal of Banking and Finance* 24: 131-154.

Appendix A.

Figure 10. Price evolution for the spot market data.

Panels from A to E present the spot price evolution for WTI crude oil, Brent crude oil, heating oil, unleaded gasoline and natural gas respectively. In the x-axis is the observation period and in the y-axis the measuring units.





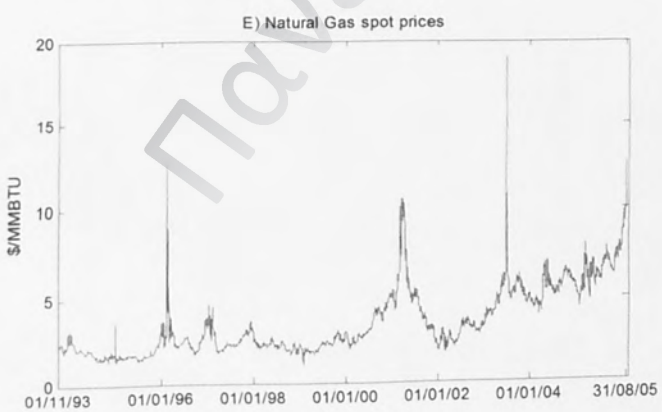
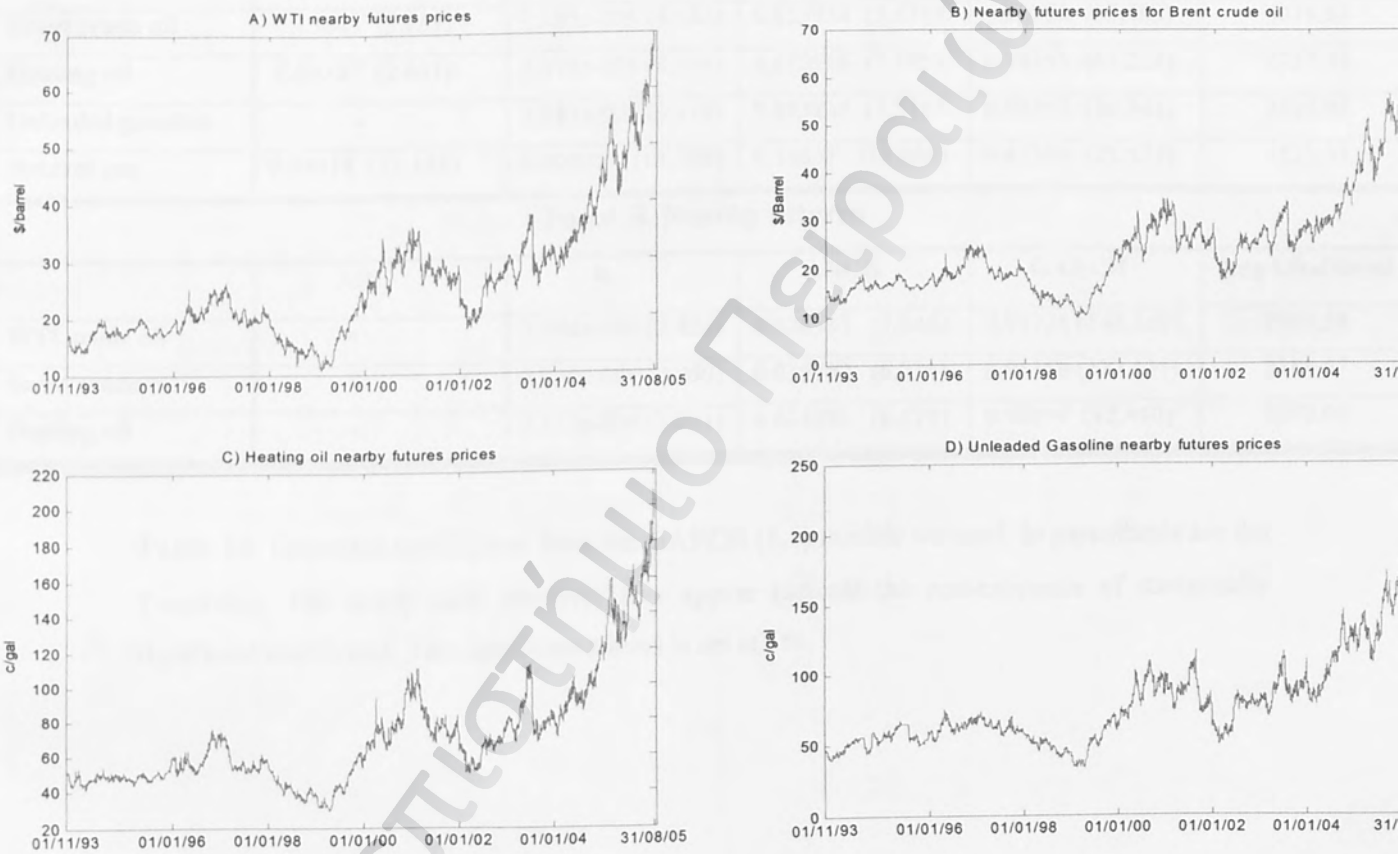


Figure 11. Price evolution for the nearby futures data.

Panels from A to D present the price of nearby futures evolution for WTI crude oil, Brent crude oil, heating oil and unleaded gasoline respectively. In the x-axis is the observation period and in the y-axis the measuring units.



Appendix B

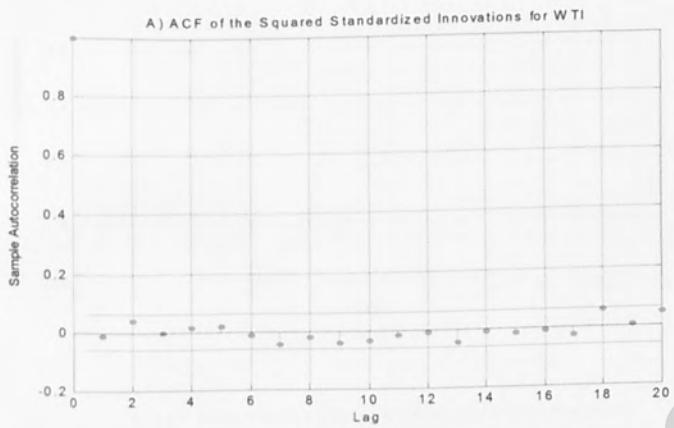
Table 18.

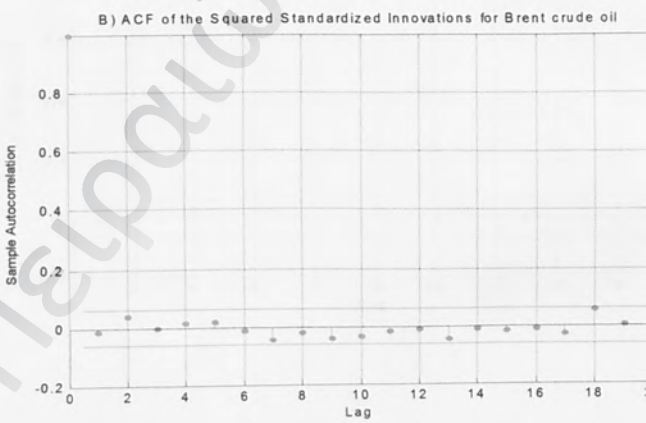
Panel A. Spot					
	AR	**K**	**ARCH**	**GARCH**	**Log Likelihood**
WTI crude oil	0.82669 (8.753)	1.198e-005 (4.452)	0.066203 (7.082)	0.90707 (85.428)	2733.54
Brent crude oil	0.15643 (5.089)	1.185e-005 (4.442)	0.053954 (5.479)	0.92758 (65.463)	3071.63
Heating oil	0.03367 (2.041)	1.376e-005 (4.539)	0.072656 (7.195)	0.89355 (63.224)	2757.98
Unleaded gasoline	-	2.583e-005 (3.118)	0.043214 (3.574)	0.90293 (36.944)	2625.05
Natural gas	0.24418 (11.161)	0.000297 (14.700)	0.54637 (19.908)	0.45363 (25.523)	1821.58
Panel B. Nearby futures					
	AR	**K**	**ARCH**	**GARCH**	**Log Likelihood**
WTI crude oil	-	1.994e-006 (2.424)	0.036533 (7.946)	0.95728 (148.369)	2904.58
Brent crude oil	-	3.236e-006 (2.960)	0.036762 (6.226)	0.95358 (122.535)	2886.67
Heating oil	-	5.155e-006 (2.801)	0.044391 (6.175)	0.93894 (82.460)	2892.68

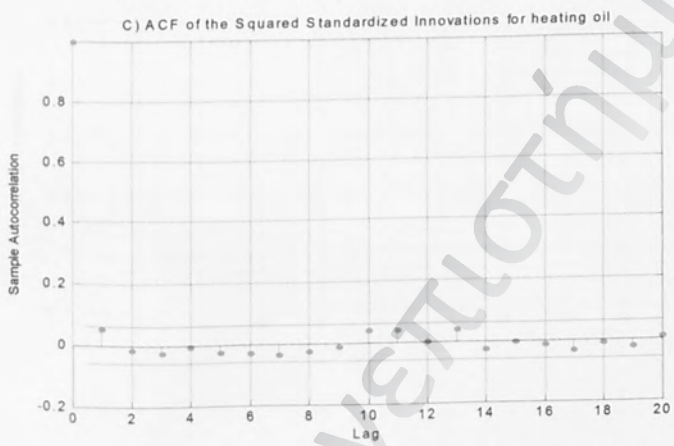
Table 18: Estimated coefficients from the GARCH (1,1) models we used. In parenthesis are the T-statistics. The empty cells wherever they appear indicate the non-existence of statistically significant coefficient. The significance level is set at 5%.

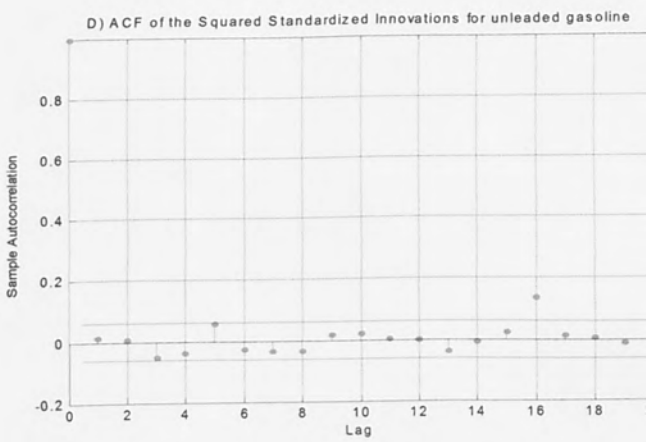
Appendix C

Figure 12. The ACF of the squared standardized innovations for the spot market data after implementing the GARCH(1,1) variance model.
Panels from A to E present the ACF of the squared standardized innovations for WTI, Brent crude, heating oil, unleaded gasoline and natural gas respectively.









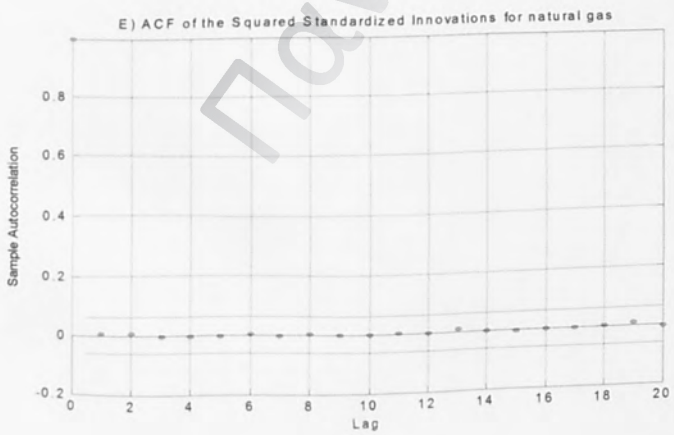


Figure 13. The ACF of the squared standardized innovations for the nearby futures data after implementing the GARCH(1,1) variance model.
Panels from A to C present the ACF of the squared standardized innovations for WTI, Brent crude and heating oil respectively.