Value-at-Risk Estimation and Extreme Risk Spillover between Oil and Natural Gas Markets

By

Dmitriy Pigildin

Submitted to

Central European University

Department of Economics

In partial fulfillment of the requirements for the degree of Master of Arts

Supervisor: Professor Péter Kondor

Budapest, Hungary

2009

Abstract

Based on the closing spot prices of Western Texas Intermediate (WTI) crude oil and Henry Hub natural gas spanning from 1994 to 2009, this work examines the performance of four risk quantification methodologies (widely known as Value-at-Risk) and assesses them against several accuracy and efficiency criteria. The results indicate that at 95% confidence level GED-GARCH method performs better than any other alternative: the VaR series obtained with this method are statistically accurate and are the least likely to result in forgone profits from speculation. The VaR series calculated with GED-GARCH are used further to investigate extreme risk spillover between the respective markets. Using the Hong's concept of Granger causality in risk (2002), we show that there is a significant risk spillover from oil market to natural gas market during the period in consideration, while there is no spillover in the reverse direction. Moreover, the upside risk spillover from oil to gas markets is found to be more statistically (and economically) significant and protracted than the downside risk spillover.

Table of contents

1. Introduction

The OPEC oil embargo of 1973, the Iranian Revolution of 1979, and deregulation of energy markets in the eighties put an end to the era of controlled stable energy prices that prevailed before. Indeed, it was then that energy markets have become so volatile that quantification and management of the price risk turned into a critical issue.

To fully understand the patterns of energy commodities risk it is not enough to merely calculate it; it would also be helpful to determine its origins. While there are always many independent factors affecting price dynamics in oil or gas market (commodity storage capacity, demand prospects), it may be the case that the risk in one market can generate risk in another. The information about such risk spillover can scarcely be overestimated and is of practical importance for commodity market participants as well as for other economic agents.

Using the time series of closing spot prices of Western Texas Intermediate (WTI) crude oil and Henry Hub natural gas spanning from 1994 to 2009, this work sets out as a purpose to examine four methodologies of risk estimation (widely known as Value-at Risk), test their performance against several accuracy and efficiency criteria, and compare the findings with those obtained in the literature of the field. The results of the best-found approach are used further for investigating extreme risk spillover between the respective energy markets.

1.1. Value-at-Risk (VaR) and extreme risk spillover

As a measure of risk, we employ VaR methodology, which nowadays has been accepted as a frontline defense versus price risk. Its main idea was established by JP Morgan under the name Riskmetrics^{TM} in 1989 and was popularized in the technical document of 1994, which is considered to have launched the use of VaR by financial and public institutions. In 1996 the

concept found its place in the Amendment to Basel Accord and has made the existing rigid 8% rule of capital requirement much more efficient since it allowed the institutions to choose own models to quantify the maximum loss over ten days with 99% confidence level (Giot and Laurent, 2003).

In its essence, VaR is a universal tool that assesses a potential loss within a pre-specified period of time at a pre-specified confidence level. For example, if the value of VaR for a time series of prices is equal to $23,000 over one day at 95% confidence level, a market participant can assert: "I am 95% sure that tomorrow the price will not fall below $23,000 (if tomorrow is a trading day)." What made VaR approach quite appealing is its parsimoniousness – the entire price risk can be represented with a single number.

Generally, one might use different confidence levels and VaR time periods. For example, 99% confidence level and ten day horizon is the specification advised by Basel Committee (Nylund, 2001). Lower confidence levels and frequencies are used as well. In this work, 95% confidence level and one day period VaR specification is employed (although 99% specification will also be partly discussed).

Although the importance of energy market risk management per se is out of question, the choice of a better VaR methodology remains less unambiguous. Plenty of ways to measure VaR exist, each having own advantages and shortcomings. While, for example, the methods of Historical Simulation family are easy to implement and are very intuitive, their assumption of time-constant returns distribution is at odds with continuous structural changes in energy markets, which can lead to very rigid VaR estimations. On the other hand, a standard GARCH method is better equipped to model returns volatility; however, it imposes the normality assumption that contradicts the empirical observation that the distribution of energy commodity

returns is often leptokurtic. As a result, the VaR measures are underestimated. Hence, augmented by methodological speculations, the task of finding a better VaR approach becomes that of purely empirical nature.

The extreme risk spillover between oil and natural gas markets is of great interest since the two energy commodities are substitutes for particular groups of non-residential consumers; so, for example, a positive price shock in the oil market can lead to larger demand in the gas market, and prices can start to converge. On the other hand, there is a big impediment to their substitutability. While oil is traded internationally, natural gas markets are regional due to imperfect mobility – natural gas requires developed pipeline infrastructure to be freely transported[1]. That's why, for example, in case of a positive oil price shock, the profitable substitution of expensive oil by cheap gas cannot be so pronounced. Also, since the price of oil is subject to changes in international demand (as opposed to regional demand in case of natural gas), we are effectively speaking about the risk spillover from oil market to natural gas market, since there is no any feasible way of regional natural gas market risk to translate into international oil market risk.

1.2. Main findings and contributions

The assessment indicates at the complete failure of Historical Simulation methods for both energy commodities and at the rigidity of the assumption that future risk can be approximated by past risk. Generalized Error Distribution (GED) GARCH method proves to be the best VaR method at 95% confidence level due to its higher efficiency relative to other methods. Quite surprisingly, GARCH-N, the method that does not account for non-normality of returns'

[1] The world of energy is changing however. Liquefied Natural Gas (LNG) is the future of mobile gas since its transportation does not require pipelines. Currently accounting for 7% of the world natural gas demand, LNG is expected to grow annually on average by 6.7% in trade by 2020 (Cook, 2005).

distribution, is found to perform better than GED-GARCH at 99% confidence level, and the reason is that the relative benefit from its efficiency is larger than relative cost from its inaccuracy.

With respect to extreme risk spillover results, we find statistically significant risk spillover running from WTI to Henry Hub, while none is found for the reverse direction. In the Granger causality sense, WTI risk can therefore be a good basis for risk prediction in Henry Hub market but not the other way around. The robustness check provides rough evidence that the source of extreme risk in Henry Hub market can indeed be the risk in WTI market.

Finally, we find that there exists asymmetry in opportunities between sellers and buyers of gas because the upside risk spillover from WTI to Henry Hub is much more statistically significant and protracted than the downside one. We show that an extreme positive oil price shock advantageous for gas sellers can translate to natural gas market shortly (five days) and can accumulate over time (up to two weeks), while a negative shock advantageous for gas buyers is much less significant and, if happens, is shorter (up to five days).

The work extends the existing expertise of the field in a number of aspects. First, VaR concept has traditionally been perceived as a precaution measure against downside risk, that is, falling prices. However, the risk in financial markets (and more so in case of commodity markets) comes from rising prices as well[2], and it also needs quantification. That's why VaR measures for both upside and downside risk are analyzed in this work. To the best of our knowledge, only two papers considered risk in both directions.

Second, regarding the criteria of VaR approaches assessment, only few papers considered both accuracy and efficiency criteria. While it is important for a risk manager to use an accurate

[2] In stock markets rising prices have adverse consequences for short-sellers; while in commodity markets the roles of winners and losers are even more pronounced: falling prices favor buyers, while rising prices favor sellers of commodities.

VaR method, which would not underpredict a price shift for tomorrow, it is also important to avoid conservative VaR estimates overpredicting the risk, since this would lead to forgoing potential profits. As such, efficiency criteria are known for long since the work of Hendricks (1996); however, their use in energy market risk literature surprisingly has been limited. The VaR methodology assessment in this work uses both accuracy and efficiency criteria for inference.

Third, the sample used here for VaR forecasts is nontrivial for the fact that it covers both calm and volatile periods, and provides quite challenging environment for assessing predictive performance of the VaR methods. Finally, the Granger causality in risk concept used in this work for extreme risk spillover investigation is applied to oil and natural gas dimension for the first time. Fan et al. (2008) use this methodology for WTI and Brent crude oil while Hong et al. (2009) apply it to Euro/Dollar and Yen/Dollar rates.

The structure of the thesis is as follows. We start with the Section 2, which reviews the existing relevant literature of the field. Section 3 speaks about the chosen VaR methods, their assessment, and the Granger causality in risk approach for studying the extreme risk spillover. In Section 4, we describe our data and sample as well as prepare the data for further use. The findings of the research are covered in Section 4, while in Section 5 we present concluding remarks, implications, and describe some limitations of our research, which could be accounted for in future extensions.

2. Literature Review

In this section we review the existing background of the field. First, we describe the general taxonomy and basic characteristics of VaR approaches developed so far. Then we shift the focus to the literature on VaR approaches used in energy markets. Finally, we examine the literature on the risk spillover between oil and natural gas markets.

2.1. VaR measurement: A general sketch[3]

Methods for VaR measurement are categorized into three groups: *Variance-Covariance* (VCV) (also known as delta-normal or analytic VaR), *Historical Simulation*, and *Extreme Value Theory (EVT)*. In what follows, we describe each separately.

2.1.1. Variance-Covariance (VCV)

The basic assumption of VCV is that the returns are normally distributed and the errors are not serially correlated. With this, the potential loss changes in proportion to the standard deviation of a variable under consideration (prices, returns, or portfolio values):

$$\mathrm{VaR}_t = \alpha\sigma_t ,$$

where α is the value of Normal CDF for a given probability (0.95 (or 0.05) and 0.99 (or 0.01) in this work), and σ_t is the sample standard deviation[4]. Calculation of VaR therefore pins down to calculation of the sample standard deviation since α values are predetermined. VCV has four ramifications to model the standard deviation for time *t*.

[3] This sub-section follows the typical structure that majority of papers on VaR follow when considering VaR taxonomy. For more formal categorization the reader is referred to Ch.10 "Approaches to measuring VaR" of Philippe Jorion's "Value at Risk: The New Benchmark for Controlling Market Risk" (Jorion, 1997).

[4] This is the specification for single financial series. A more general would be the one for a portfolio: $\mathrm{VaR}_t = \alpha\sigma_{p,t} = \alpha\sqrt{\sum_{i=1}^{n} w_{i,t}^2\sigma_{i,t}^2 + 2\sum_{i=1}^{n}\sum_{j<i}^{n} w_{i,t}w_{j,t}\sigma_{ij,t}}$, that is, including covariance among the variables. This research, however, is limited to single variables only, that's why the aspect of covariance is irrelevant here.

First, standard deviation may be assumed constant over time. So, once it has been estimated for one sub-sample, it is used for all future observations as well. This rigidity in the assumed volatility is without doubt a very unrealistic assumption because volatility of some returns tends to be time-variant[5]. The following three ways of modeling standard deviation suppress the constancy assumption.

One could use Equally Weighted Moving Average approach to get the value of standard deviation (Hendricks, 1996). Some fixed amount of historical data (k days of prices or returns) is used recursively to generate the value of σ_t but the weights assigned to the variables in time from s to $t-1$ are assumed equal regardless of their temporal remoteness:

$$\sigma_t = \sqrt{\frac{1}{k-1}\sum_{s=t-k}^{t-1}\left(x_s - \mu\right)^2},$$

where x_s is the change in the variable value on the day s, and μ is the average change.

Another alternative is to use the Exponentially Weighted Moving Average approach first presented by Roberts (1959). Intuitively, it assigns smaller weights to more remote observations:

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

So, as s approaches $t-1$, the assigned weight grows larger. $0<\lambda<1$ is a "decay factor", which shows how fast these weights diminish as they become more remote.

Finally, (G)ARCH models by Engle (1982) and Bollerslev (1986) are another way to model volatility of returns, and will be presented in detail in the methodology section.

The last three specifications of VCV relax the assumption of constant volatility but one still has to bear in mind that standard versions of these models do impose the assumption of

[5] Such non-constant behavior of volatility is nicely characterized by Hopper (1996) as playing different roulette wheels every evening with different ranges of returns.

normality. It is considered as the worst shortcoming of VCV since many financial series (Van den Goorbergh and Vlaar, 1999) including those of energy commodities proved to be having thicker tails than what Normal distribution would presume. Hence, the extreme deviations, which are naturally concentrated in the tails of distributions, are underpredicted if the assumption of normality is imposed. The rest two approaches described below do not make any assumptions about the distribution of variables. Nevertheless, they are not ideally applicable either.

2.1.2. Historical Simulation (HS)

This is conceptually the simplest method of calculating VaR. It hinges on the actual distribution of the variable and therefore makes no theoretical assumption about it; the only assumption is the constancy of this distribution over time. Its core idea is that future resembles the past, and tomorrow's risk can be modeled by the information set based on the past distribution of returns of certain length (so called window).

HS has some common characteristics with Equally Weighted Moving Average approach considered shortly before. It also assigns equal weights to past returns and uses their distribution for inference. Yet they should not be confused. The former method uses history only to estimate the standard deviation and imposes normality assumption, while HS does not make such assumption; rather it readily uses the actual percentiles of the distribution.

2.1.3. Extreme Value Theory (EVT)

In its essence, EVT is a combination of VCV and Historical Simulation. It explicitly targets statistical modeling of the tails distribution thereby eliminating the need to assume normality. First, a threshold percentile is determined, which separates non-tail from tail distribution, and

then the tail distribution is approximated by known functions. A comprehensive coverage of EVT practical aspects is given by Këllezi and Gilli (2000).

2.2. VaR measurement: Energy markets

The need for energy markets risk quantification is stipulated by the volatile environment of these markets and significant importance they hold. Movements in energy prices can have profound impact on all facets of human activity, not only well-being of those involved in commodity trading. For example, Sadorsky (1999) finds that oil price shocks have asymmetric effects on an economy as a whole. Thus, he finds that oil price shocks have full impact on the economic activity, while changes in economic activity barely have any impact on oil prices. The importance of this finding is more pronounced for the commodity export-oriented countries whose budget revenues could be exposed to a real disaster once oil prices unexpectedly plunge.

Although not as crucial on a global level, the natural gas prices are no less important on the regional scale. Inherently larger volatility of gas markets complicates the decision-making of local municipalities in finding the best time to buy gas. Consequent losses from poor timing purchases are passed on individual consumers even when the gas futures prices are falling (Davis, 2006). Such inefficiency also calls for means of managing market risk.

Naturally, however, risk in energy commodities was first investigated together with risk in other commodities since all commodities share the "physical" aspect. Moreover, prices of all commodities (energy, agriculture, metals) are affected by supply and demand imbalances, storage constraints, and seasonality effects (Giot and Laurent, 2003). VaR calculation for individual energy commodities is a relatively recent research ramification.

From the beginning the vein of literature admitted the non-normality of the commodities distributions (fat tails and sometimes negative skewness) and concentrated on the search for

models capable of capturing these facts. Traditionally, researchers have been taking Riskmetrics™ as a reference method, literally trampling it down and proving the superiority of newly proposed approaches. Thus, Giot and Laurent (2003) test the performance of skewed Student ARCH and skewed Student APARCH (Asymmetric Power ARCH) methodologies (Student distributions are intended to account for fat tails while skewed versions – naturally for skewness) for a number of commodities including WTI and Brent crude oil. They conclude with superiority of Student APARCH specification over Student ARCH and Riskmetrics™ and explain this by its greater flexibility: when modeling the second moment of the returns distribution, it relaxes the requirement of conditional *variance* modeling as they find that for many commodities – especially energy – one should model the conditional standard deviation rather than variance. Student ARCH methodology, however, is still found to perform quite well for "small" percentiles, and, besides, it attracts with its relative simplicity and applicability in a plain spreadsheet setup without programming.

Rather than inventing new techniques, some researchers opt for improving the extant ones. Cabedo and Moya (2003) advance the Historical Simulation (HS) approach by adding the element of ARMA forecasts (HSAF)[6]. For the daily spot Brent crude oil prices, they show that HSAF methodology is both more accurate (VaR values are closer to the assumed likelihood level) and efficient (less rigid) than the standard HS methodology. The findings also indicate that HSAF is more efficient than ARCH methodology (not as accurate though).

Employing the same methodology and research structure, Sadeghi and Shavvalpour (2006) similarly show that for OPEC weekly oil prices, HSAF outperforms both HS and ARCH-type methodology at 99% confidence level. Neither group of the authors, however, do not consider

[6] The approach will explicitly be covered in the methodology section.

the sensitivity of the method performance to the choice of fitted ARMA model, so their results cannot be taken universally unambiguous.

One of the first works exclusively dedicated to VaR in energy commodities on a more structural level is the one of Hung et al. (2008). Besides analyzing crude oil (both WTI and Brent) markets, they consider various refined products such as propane, gasoline, and heating oil for more robust results. They use the family of GARCH models and evaluate them based on the accuracy and efficiency criteria. Specifically, GARCH-HT (Heavy Tail) model is introduced for the first time and is compared to GARCH-t and standard GARCH models. The first rationale for GARCH-HT is again the evidence for heavy tails in the distributions of commodities' returns in question. Second, GARCH-t, which is often used to account for non-normality, is unable to simultaneously capture the heavy-tails and positive excess kurtosis due to its distributional properties, in contrast to GARCH-HT. Thus, GARCH-t model is found the least while GARCH-HT the most accurate at both low and high confidence levels. In terms of efficiency, the findings have been mixed but generally, the authors summarize GARCH-HT as the more preferred model.

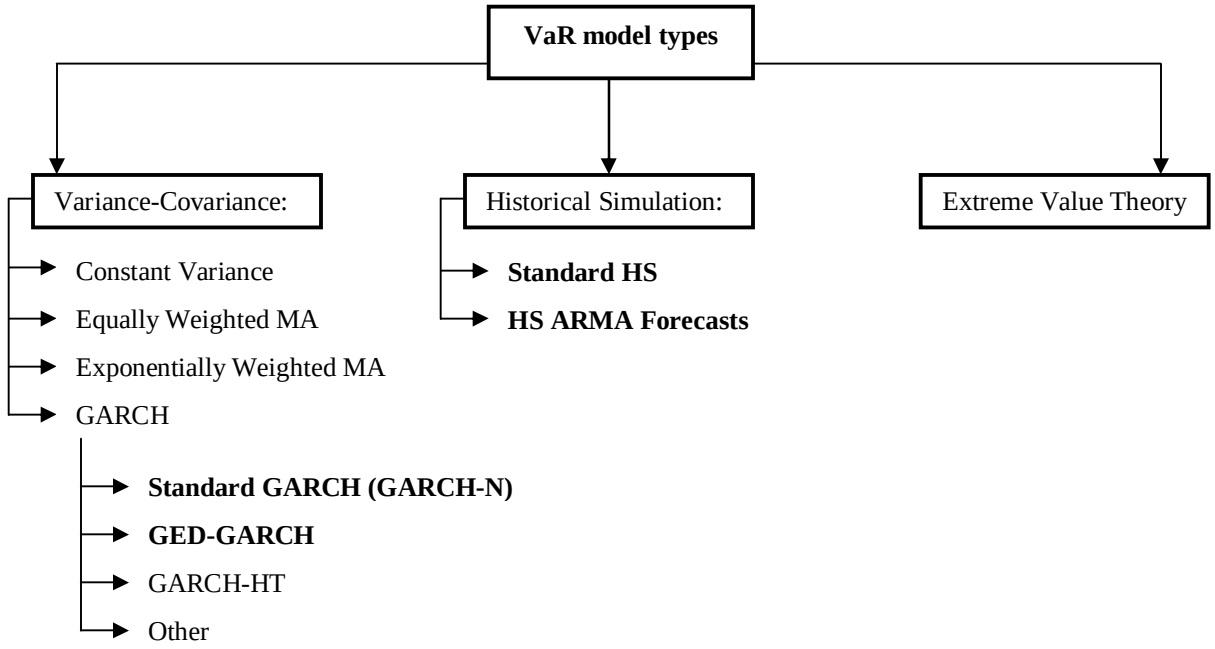
Finally, Fan et al. (2008) use another tool for modeling fat tails – Generalized Error Distribution (GED), which at some parameters converges to Normal distribution. They also employ GARCH models and apply standard GARCH and GED-GARCH to the daily series of WTI and Brent prices. GED-GARCH is found to perform better than standard GARCH at 99% confidence level, while at 95% the two models are statistically identical.

An interesting fact: the overwhelming majority of the authors when emphasizing their contributions keep saying that former literature assumed that the returns of commodities are normally distributed, although the search for tools accommodating heavy tails has been going for

more than a decade already. Non-normality of the commodities returns' distribution is a well-established fact, and its "discovery", in our opinion, cannot be considered a contribution anymore.

To summarize general as well as energy market specific VaR categorization, we reflect all methods in the Diagram 1 below. The highlighted methods are used in this research and are considered in greater detail in 3.1.

Diagram 1.
Structural presentation of VaR taxonomy



2.3. Extreme risk spillover

The literature on this phenomenon is rather scarce. To the best of our knowledge, the earliest relevant piece is the one of Ewing et al. (2002) where they consider volatility transmission in oil and natural gas markets rather than extreme risk spillover. They use oil and natural gas indices, which are comprised of stock prices of fifteen largest widely held oil and gas companies in the US. In addition to volatility persistence in both markets, they find that "oil return volatility depends on past natural gas return volatility (as well as its own past volatility)" (p.536). This fact

is explained as the result of some substitutability between these two commodities; however, no clear reason is given why the volatility transmission goes in the gas-to-oil direction.

A fundamentally different approach of risk spillover investigation is proposed in Hong (2002) and is implemented with respect to Chinese stock markets in Hong (2003). The author proposes the concept of Granger causality in risk and focusing solely on the left tail probabilities (thereby concentrating on extreme downside risk) he finds significant intra-China risk spillover[7]. The spillover between Chinese stock markets and Asian (Korea, Taiwan, Singapore) as well as international (US, Japan, Germany) markets is less evident: some groups of Chinese shares are found to have risk links to those markets while some are not. The cited reason is the stock market segmentation in China: at that time, local and foreign investors had access to separate markets, which effectively restricted the propagation of extreme shocks from international stock markets.

Granger causality in risk methodology has been applied to energy markets as well. Fan et al. (2008) investigate the extreme risk spillover between WTI and Brent crude oil returns. They establish two-way Granger causality in the commodities' extreme risk. More specifically, WTI risk helps predict the risk in Brent returns for both downside and upside shocks; however, Brent risk carries no information for predicting downside risk in WTI, though can predict upside. The reason for the very existence of causality in risk, the authors say, is increasing globalization, while the direction of risk causality from WTI to Brent is explained by the US dollar being the major invoicing currency and the US oil consumption being largest on the world scale.

With respect to risk spillover task, our work combines the economic logic of Ewing et al. (2002) with the methodology introduced by Hong (2002). With this in mind, we now turn to methodology section, which will feature a detailed review of some of the cited models of 2.2 as well as Hong's concept of Granger causality in risk.

[7] We note that it is the methodology of Hong (2003) but not its subject what is relevant for our work.

3. Methodological Issues

This section provides the theoretical framework for the research. The first sub-section describes the methodology of the implemented VaR approaches; the second one features several criteria along which the approaches are assessed, while the third sub-section speaks about the methodology of Granger causality in risk to investigate the extreme risk spillover between the two markets.

3.1. VaR methodology

Overall, four approaches are employed in this work: Historical Simulation (HS), Historical Simulation with ARMA Forecasts (HSAF), GARCH Normal Distribution (GARCH-N), and GARCH Generalized Error Distribution (GED-GARCH).

There are two rationales to use these methods. First, HS and HSAF methods constitute one family (Historical Simulation), while GARCH-N and GED-GARCH – another (GARCH); and this dimension tempts to ask which family of approaches performs better in our exercise.

Second, each method represents an extreme case of HS and GARCH groups in terms of evolution and flexibility. To the best of our knowledge, HSAF and GED-GARCH are the latest versions of Historical Simulation and GARCH families respectively, while standard HS and GARCH-N are the simplest versions of those. Thus, it is also interesting to test the intra-group superiority of these methods. The first pair of the methods to be described is Historical Simulation group.

3.1.1. Standard Historical Simulation (HS) approach

As it was said in the literature review section, this is conceptually the simplest method, which presumes that the recent past of certain length explains the present risk. So, basically, VaR for time *t* is the value of a return r at a given percentile p in the past returns distribution:

$\mathrm{VaR}_{t|t-1} = r_{t-1}^{p}$.

The calculation routine is repeated over time adding a new observation to the window and dropping the most distant one. That's why this method is "just taking sample percentiles over a moving sample" (Van den Goorbergh and Vlaar, 1999, p.22). For example, if we choose the window size to be 150 days, then the 95% confidence level *negative* VaR for the 151st day will be the return corresponding to the 5th percentile based on the distribution or returns spanning from day 1 to day 150. VaR for the 152nd day, however, will be based on the 2 to 151 distribution of returns, etc.

Just like a negative shift for producers, an upward shift in the commodity prices is a source of risk for consumers. That's why we calculate two series of VaR for each commodity – upside and downside VaR[8]. Both upside and downside VaR measures are calculated for all methods we describe below.

How many observations to include into the distribution of past returns is the main analytical question of the method. A practitioner may be interested in capturing short-term risk and choose the small size of the window (say, 150 days). The undesirable consequence of this choice is over-sensitivity of VaR values to accidental changes. If, however, the purpose is to estimate historical distribution percentiles as properly as possible, one might decide to use a window of a larger size

[8] The separation of upside and downside VaR is important in methodological sense as well. If we uniformly consider both positive and negative returns, the series of upside and downside VaR would be identical and would lead to unreliable VaR estimates. After all, for example, why would one need to be cautious downward for tomorrow if today there is a positive shock but no negative?

(say, 1500 days). This is not flawless as well: in this case, the two-three year ago outcomes would be kept in the distribution, while they might not be relevant already. As a result, the series of VaR estimates can remain unchanged for long periods, which is inefficient (we illustrate the relevance of this issue in section 4).

3.1.2. Historical Simulation with ARMA Forecasts (HSAF)

Developed by Cabedo and Moya (2003), HSAF methodology uses the distribution of fitted residuals obtained from ARMA model of past returns. To implement it, one should go through the following four steps.

First, as in any procedure involving ARMA modeling, the data should be checked for non-stationarity. Non-stationary series should be modified to get stationary ones, and then their autocorrelation behavior should be investigated. If there is no statistically significant autocorrelation in the series, there is no need to go further as the whole process can be modeled by the standard HS method. Only if we spot statistically significant autocorrelation in the series, can we proceed to the second step of the method.

The second step is the past returns ARMA model estimation using the standard Box-Jenkins methodology, which posits that the behavior of any time series can be approximated by AR, MA, or ARMA models. Several ways to evaluate a model can be employed at this stage such as the analysis of Q-statistics and Akaike Information or Schwarz Criteria.

After obtaining a satisfactory model, one should get the “in-sample” forecasts, which are simply fitted values. Fitted *residuals* are then calculated using these fitted values (or they can be obtained independently of fitted values). The distributions of these residuals (positive and negative separately) is analyzed and the percentiles corresponding to the assumed confidence level are calculated.

Finally, we use the ARMA model again; this time in order to get the "out-of-sample" forecasts. These forecasts form the basis for VaR measures, but to finally calculate the VaR series we must correct these forecasts by the percentiles calculated in the previous stage. We separately add the percentiles for negative and positive residuals to the "out-of-sample" forecasts; so, intuitively, the deviations in returns should not exceed the obtained VaR measures because we assume that the distribution of the *residuals* in the "out-of-sample" period will remain the same as it was in the "in-sample" period.

3.1.3. GARCH-N (Normal distribution)

The model has traditionally been used to account for volatility clustering. The mean equation takes the following form:

$$r_t = X_t'\phi + \varepsilon_t,\ \varepsilon_t = \sigma_t u_t,\ u_t \mid \Omega_{t-1} \sim N(0,1),$$

where X_t' is the vector of independent variables, which can include the lags of r, and ϕ is the coefficient vector. The variance of a (G)ARCH (p , q) model is explained by p lags of the past values of variance, and q lags of past squared errors:

$$\sigma_t^2 = \omega + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 .$$

The key assumption of the method is that the error u_t is normally distributed.

Practically, the approach works as follows. First, the above model is estimated for the "in-sample" stationary data (the mean equation should essentially be taken the same as it was in HSAF methodology). Then, "out-of-sample" forecasts and GARCH standard errors for these forecasts are obtained[9]. Again, the standard errors for upside and downside forecasted returns are analyzed separately. These errors are then multiplied by the Normal cumulative distribution

[9] Eviews routinely calculates these values.

function values corresponding to the assumed confidence levels[10] (this way the Normal distribution is assumed; this step is the principal difference from GED-GARCH, which is the next approach to be described). Finally, the forecasted returns series are corrected by the obtained products of CDF values and model standard errors, and we get two VaR series for the "out-of-sample" positive and negative returns[11]:

$$\mathrm{VaR}_t^{upside} = r_t + \alpha_N \hat{\sigma}_t,$$

$$\mathrm{VaR}_t^{downside} = -r_t - \alpha_N \hat{\sigma}_t .$$

3.1.4. GED-GARCH (Generalized Error Distribution)

This specification of GARCH has been developed to account for a leptokurtic distribution (Nelson, 1991), which is at odds with Normal distribution. With respect to our work, GED-GARCH presents an excellent theoretical basis to check the fit of non-Normal setup in energy markets.

The modeling of GARCH element (mean and variance equations) remains the same as in GARCH-N. The difference concerns the generalization of distribution in the following fashion. The probability density function obtains the form of:

$$f(\varepsilon) = \frac{k e^{-\frac{1}{2}\left|\frac{\varepsilon}{\lambda}\right|^k}}{2^{\left(\frac{k+1}{k}\right)}\Gamma\left(\frac{1}{k}\right)\lambda}, \left(0 \le k \le \infty\right), \text{ where } \lambda = \left(\frac{2^{\left(-\frac{2}{k}\right)}\Gamma\left(\frac{1}{k}\right)}{\Gamma\left(\frac{3}{k}\right)}\right)^{\frac{1}{2}}, \text{ and } \Gamma(\bullet) \text{ is the gamma function.}$$

The principal components of GED are scale parameter λ and shape parameter *k*. For us, it is exactly parameter *k* that brings difference. Indeed, when *k*=2 (plus if μ=0 and σ=1), GED

[10] For 95% level the value is 1.645, while for 99% it is 2.33.

[11] One should note, however, that the two equations are absolutely identical. The signs in the equation are only indicative of the actual signs. For example, the second equation is written in negative only because the returns and the values of cumulative distribution function in that case are negative by definition.

becomes Standard Normal distribution and one can say that Normal distribution is a special case of GED with *k*=2:

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{\varepsilon^2}{2}\right)}.$$

Figures A.1a-d (Appendix I) visualize the range of GED possibilities by illustrating probability density functions with different *k* parameters ceteris paribus. As *k* gets smaller than 2, the tails of the distribution get thicker and the peak gets sharper.

The first step in calculating VaR using GED-GARCH is estimating the model (luckily, Eviews has a built-in GED estimation) for the "in-sample" period. Along with the results, the software returns a GED-parameter, which is exactly *k*. Here comes the principal difference from GARCH-N method. With GARCH-N, we would now be ready to use the critical values of CDF. Yet, here, we must first take use of the *k* parameter by replicating a distribution with the obtained *k* value.

As we get the distribution, we are ready to follow the same procedure as with GARCH-N: obtain the fitted "out-of-sample" returns and correct them by the product of corresponding standard errors and the assumed GED cumulative distribution values:

$$\mathrm{VaR}_t^{upside} = r_t + \alpha_{GED}\hat{\sigma}_t,$$

$$\mathrm{VaR}_t^{downside} = -r_t - \alpha_{GED}\hat{\sigma}_t.$$

3.2. VaR assessment criteria

The performance of a VaR model can be estimated only by backtesting, that is, checking how well our predictions performed in the past. Our choice of the most satisfactory model will be based on several criteria, which can be classified into two groups. The *accuracy* criteria focus on

whether a VaR daily measure is able to cover the realized daily loss, that is, whether VaR is able to "keep its promise". The *efficiency* criterion assesses how conservative VaR is, or whether VaR "overworks". The consideration of both accuracy and efficiency criteria is vital for objective judgment because even if a VaR model is accurate, it can be "too accurate", or conservative, so that investors may have unrealized profits.

We start the discussion with the three accuracy criteria.

3.2.1. Binary loss function (BLF)

Based on our "out-of-sample" results, we assess the success of our models. There are two possible daily realizations: we can name them "norm" (0) – when VaR is able to cover the realized loss and "exception" (1) – when VaR fails to do that. Then the BLF for a method *i* looks as follows:

$$\mathrm{BLF}_{i,t+1} = \begin{cases} 1, \text{ if } r_{i,t+1} < \mathrm{VaR}_{i,t} \\ 0, \text{ if } r_{i,t+1} \geq \mathrm{VaR}_{i,t} \end{cases} \text{ for } r < 0 ; \; \mathrm{BLF}_{i,t+1} = \begin{cases} 1, \text{ if } r_{i,t+1} > \mathrm{VaR}_{i,t} \\ 0, \text{ if } r_{i,t+1} \leq \mathrm{VaR}_{i,t} \end{cases} \text{ for } r > 0 .$$

As we get these ones and zeros for every day, we calculate the average binary loss function (ABLF) and get a number between 0 and 1. Ideally, VaR would *just* "keep its promise", and the percentage of "exceptions" in the "out-of-sample" period would be exactly equal to (1-c) where c is the assumed confidence level (95% or 99% in this work). That is, if 95% is the assumed confidence level and the returns intersect VaR series exactly in 5% of the cases, the model is considered adequate. Intuitively, the value of ABLF shows the probability of the returns intersecting the value of VaR.

3.2.2. Quadratic loss function (QLF)

Unlike BLF, QLF treats "exceptions" differently. It assigns different weights to those of different magnitude:

$$\mathrm{QLF}_{i,t+1} = \begin{cases} 1+\left(r_{i,t+1} - \mathrm{VaR}_{i,t}\right)^2, \text{ if } r_{i,t+1} < \mathrm{VaR}_{i,t} \\ 0 \qquad\qquad\qquad\quad , \text{ if } r_{i,t+1} \geq \mathrm{VaR}_{i,t} \end{cases} \text{ for } r < 0;$$

$$\mathrm{QLF}_{i,t+1} = \begin{cases} 1+\left(r_{i,t+1} - \mathrm{VaR}_{i,t}\right)^2, \text{ if } r_{i,t+1} > \mathrm{VaR}_{i,t} \\ 0 \qquad\qquad\qquad\quad , \text{ if } r_{i,t+1} \leq \mathrm{VaR}_{i,t} \end{cases} \text{ for } r > 0.$$

The application of QLF could be crucial in cases where the judgment based on BLF produces marginally different results when the number of "exceptions" in one method almost equals that in another method. QLF can single out the methods where "exceptions" are much larger than in others in terms of magnitude.

3.2.3. Kupiec LR test

The likelihood ratio test suggested by Kupiec (1995) is a formal statistical check of VaR models accuracy. We denote the desired confidence level as (1-c), number of "exceptions" as M, and a sample size as N, hence, the rate of "exceptions" (also called as the rate of failures) equals $\frac{M}{N}$. The statistic of the Kupiec test is:

$$\mathrm{LR} = 2\ln\left(\left(1-\frac{M}{N}\right)^{N-M}\left(\frac{M}{N}\right)^{M}\right) - 2\ln\left(\left(1-c\right)^{N-M} c^{M}\right) \sim \chi^2(1),$$

and the null hypothesis is that the rate of "exceptions" is the true probability ($\frac{M}{N} = c$). If we reject this hypothesis (that is, if LR statistic is larger than the critical value), we say that the model is inadequate. We should also note that we treat Kupiec test as a superior indicator of

accuracy over BLF and QLF since we are interested in the *statistical*, rather than absolute precision. BLF and QLF are taken as indicators of the deviation magnitude to have understanding of how methods over- or underestimate risk in absolute terms.

3.2.4. Mean relative scaled bias (MRSB)

This is an efficiency assessment measure proposed by Hendricks (1996). A risk manager may forgo profits if he chooses a too conservative model. Using MRSB, one can find out which model produces the smallest average VaR. Naturally, since this is a relative measure, *all* VaR models should be estimated before making inference about MRSB.

The process goes in the following way. First, we should calculate an average VaR measure $\bar{X}_t$ for a date *t* across all *M* methods:

$$\bar{X}_t = \frac{1}{M}\sum_{i=1}^{M} X_{i,t} .$$

Second, we measure the percentage deviation of each method daily VaR from $\bar{X}_t$. This gives us the *daily relative bias* for each method and each date. Finally, we average these deviations across all *N* dates and obtain MRSB – a single number – for each method:

$$\text{MRSB}_i = \frac{1}{N}\sum_{t=1}^{N} \frac{X_{i,t} - \bar{X}_t}{\bar{X}_t} .$$

When making inference about the MRSB results, one should completely abstract from accuracy measures. MRSB criterion is purely relative; it has no external benchmark to assess against.

Once we get the results of all four criteria, we have to proceed with the choice; so the main question is how the criteria are ranked in their importance. Of course, an ideal scenario would be to have a method with a "fail-to-reject" Kupiec statistic and the smallest MRSB to proclaim that method the best. But what if a method does not satisfy the accuracy criteria but has the smallest

MRSB? Shall we leave it in consideration at all? That is, will accuracy or efficiency consideration dominate? We adopt the idea expressed in Hung et al. (2008): the accuracy of a model is a more important criterion for a risk manager than efficiency. They rationalize it by the fact that "Basel Committee on Banking Supervision (BCBS) allows banks to adopt the internal VaR model if their models can pass the backtests" (p.1188), that is, accuracy tests. In addition, we think that one should consider the object of research in question – here we are speaking about commodities, which bring "physical" value; so the necessity to secure supplies for consumers or profit for producers may be more pronounced than the potential for speculative windfalls. Overall, however, we believe that the issue of criteria superiority is open to debate: after all, it pins down to the problem of relative importance of hedging versus speculation in energy commodity markets. To the best of our knowledge, no research explicitly considered this matter.

3.3. Extreme risk spillover

To measure the risk spillover between WTI and Henry Hub markets we employ the concept of Granger causality in risk coined by Hong (2002).

The most popular concept in Granger causality sense has been Granger causality in mean, and there also exists the concept of Granger causality in variance. However, in the *extreme* risk spillover investigation we concentrate on the comovements of distribution tails – either left (for downside risk) or right (for upside risk). Even Granger causality in variance is irrelevant since variance is a "two-sided risk measure" (Hong, 2009, p. 3), while the VaR framework presumes one-sided risk estimation (that is, upside and downside risk considered separately). Moreover, the sole consideration of either first or second moments is insufficient because comovements of distribution tails can be brought about by skewness and kurtosis as well. Thus, Granger causality in risk in the VaR setup can rise even if there is no Granger causality in first moments.

In its nature, the method checks if the extreme risk in one market (the times when actual returns cross VaR values) carries predictive power for the extreme risk in another market in future. As a measure of extreme risk we take the VaR series of the method that we find to be the best at 95% confidence level following the methodology in 3.2.

In what follows we describe the method of Granger causality in risk in detail. First, we generate the risk indicator $Z_{lt} \equiv 1(Y_{lt} > \mathrm{VaR}_{lt}),\ l = 1,2$ for upside risk and $Z_{lt} \equiv 1(Y_{lt} < \mathrm{VaR}_{lt}),\ l = 1,2$ for downside risk where $1(\bullet)$ is the indicator function and *l* represents WTI or Henry Hub markets. When the returns exceed the VaR measure, Z_{lt} takes the value of 1, otherwise it is 0. Practically, this function is the same as Binary Loss Function described earlier.

As we get these risk indicator series, the whole essence of the method pins down to testing the null hypothesis $\mathrm{H}^0 : E(Z_{1t} \mid I_{1t-1}) = E(Z_{1t} \mid I_{t-1})$ against the alternative $\mathrm{H}^A : E(Z_{1t} \mid I_{1t-1}) \neq E(Z_{1t} \mid I_{t-1})$. Here $I_{t-1} \equiv (I_{1t-1}, I_{2t-1})$ where $I_{1t-1} = \{Y_{1t-1}, ..., Y_{11}\}$ and $I_{2t-1} = \{Y_{2t-1}, ..., Y_{21}\}$ are $t-1$ information sets for WTI and Henry Hub, which contain the time series of returns backwards to the first observation.

Then we estimate the cross-covariance function between the vectors $\{\hat{Z}_{1t}\}$ and $\{\hat{Z}_{2t}\}$:

$$\hat{C}(j) \equiv \begin{cases} T^{-1} \sum_{t=1+j}^{T} (\hat{Z}_{1t} - \hat{\alpha}_1)(\hat{Z}_{2t-j} - \hat{\alpha}_2),\ 0 \le \mathrm{j} \le \mathrm{T}-1, \\ T^{-1} \sum_{t=1-j}^{T} (\hat{Z}_{1t+j} - \hat{\alpha}_1)(\hat{Z}_{2t} - \hat{\alpha}_2),\ 1-\mathrm{T} \le \mathrm{j} < 0, \end{cases}$$

where $\hat{\alpha}_l \equiv T^{-1} \sum_{t=1}^{T} \hat{Z}_{lt}$, that is, the means of WTI and Henry Hub risk indicators. The *j* variable can be referred to as the lag indicator. Basically its meaning depends on the meaning of *l* values. In this work *l*=1 designates Henry Hub market while *l*=2 designates WTI market; that's why the

values of $0 \le j \le T-1$ represent the lags, with which WTI risk translates into Henry Hub risk, while the values of $1-T \le j < 0$ represent the lags, with which Henry Hub risk translates into WTI risk.

We use the sample cross-covariance function in order to get the sample cross-correlation function $\hat{\rho}(j) \equiv \frac{\hat{C}(j)}{\hat{S}_1 \hat{S}_2},\ j = 0, \pm 1, ..., \pm(T-1)$ where $\hat{S}_l \equiv \sqrt{\hat{\alpha}_l(1-\hat{\alpha}_l)}$ is the sample standard deviation of $\left\{\hat{Z}_{lt}\right\}$ series[12]. In a spreadsheet environment like Excel we can present this cross-correlation function as two series of correlations between various lags of one commodity returns and the contemporaneous values of another's returns and vice versa.

However, the assessment of risk spillover solely based on cross-correlation functions would be meaningless as even if there is no risk spillover at, say, the 100th lag, there still can be positive correlation just because the intersection of VaR in one market took place 100 days after the intersection of VaR in another. However, we know that markets are largely affected by recent events rather than distant ones. To avoid the mechanical inference we need to find a way to sort relevant VaR intersections from irrelevant ones. One way to do it is to use a kernel function, which would impose a larger weight on the nearby lags relative to the higher order lags[13]. Following Hong et al. (2009) we use the Daniell kernel $k(x) = \frac{\sin(\pi x)}{\pi x}$ to separate the relevant lags.

[12] Such specification is due to the fact that $\left\{\hat{Z}_{lt}\right\}$ follows Bernoulli distribution.

[13] There exist truncated and non-truncated versions of kernel functions. A truncated (uniform) function imposes equal weight on the values in its range, while a non-truncated distinguishes them by different weights (e.g. downward-weighting for more distant lags). Hong et al. (2009) considers several kernel function specifications, which could be applied in this method. He finds that as long as non-truncated versions of kernel function are used, a choice of a particular function has little substantive impact on the results.

The test statistic (for H^0 against H^A), which asymptotically converges to N(0,1) is as follows:

$$Q(M) \equiv \frac{\left(T\sum_{j=1}^{T-1} k^2(\frac{j}{M})\hat{\rho}^2(j) - C(M) \right)}{\sqrt{D(M)}} \quad \text{where} \quad C(M) \equiv \sum_{j=1}^{T-1}\left(1-\frac{j}{T}\right)k^2\left(\frac{j}{M}\right) \text{is the centering}$$

moment and $D(M) \equiv 2\sum_{j=1}^{T-1}\left(1-\frac{j}{T}\right)\left(1-\frac{j+1}{T}\right)k^4\left(\frac{j}{M}\right)$ is the standardization moment. The argument *M* basically represents the truncation lag; it denotes how many lags we are considering for further analysis of risk spillover. For example, taking *M*=3 we test if the risk spillover from one market to another is statistically significant within 3 lags[14].

Intuitively, the larger the value of *Q*(*M*) the more likely we are to reject H^0 since *Q*(*M*) is basically comprised of standardized, summed and weighted correlations across *j*. If this sum of cross-correlations is sufficiently large, we can suggest that information about risk in one market can help in predicting risk in another.

One might wonder, however, why care about this method if we could simply apply a basic Granger causality regression-based test to the risk indicators $\{\hat{Z}_{1t}\}$ and $\{\hat{Z}_{2t}\}$? Hong et al. (2009, p.3) motivate it by the fact that "the risk indicator Z_{lt} has to be estimated, and parameter estimation uncertainty has a nontrivial impact and should be taken care of properly". They also find that the finite sample performance of a regression-based method is inferior to the performance of the downward-weighting-based method. Since this work sets out to apply a method that proved to perform better than others in the field, the comparison of methodologies is left for further research.

[14] The graphs of Daniell kernel for different *M* are given in Appendix II in order to provide a better visual perception about the nature of this function.

4. Data and Sample Description

4.1. Data

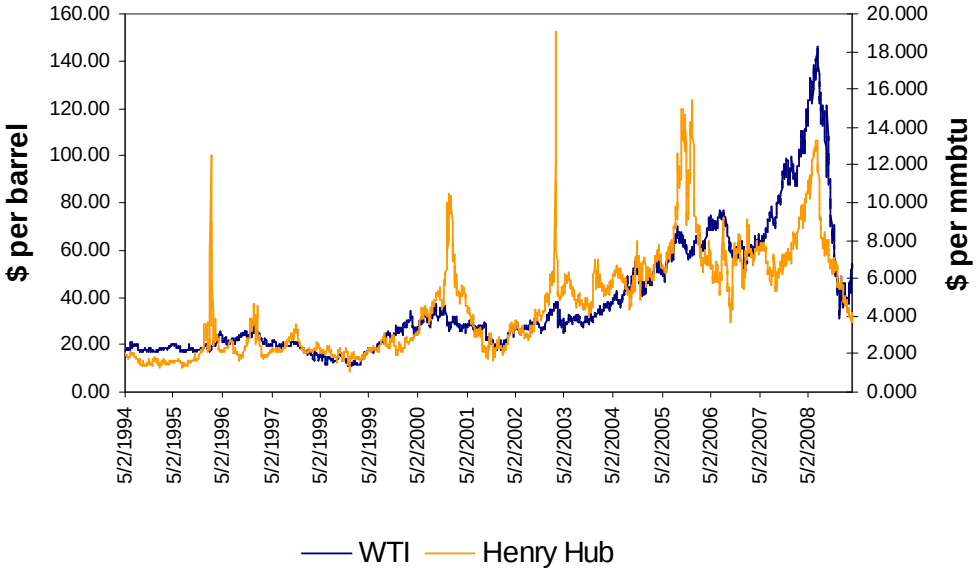
The research uses the daily time series of spot prices of West Texas Intermediate (WTI) crude oil ($ per barrel) and Henry Hub natural gas ($ per mmbtu[15]) obtained from the official website of the State of Utah citing Wall Street Journal as a source of these daily series. WTI, also known as Texas Light Sweet, is produced and refined in the US. It is high quality oil with low content of sulfur and low API gravity[16], which makes it a pricing benchmark for oil products not only in the US but in the whole North American region (EIA, 2006). Henry Hub is the central connection of natural gas pipelines located in Louisiana. It connects to other 13 pipelines in the US, and, like WTI in oil pricing, is a major benchmark for natural gas pricing in North America.

The choice of these two commodities for the research is not arbitrary. Both are traded on New York Mercantile Exchange (NYMEX) and, as has been said, are basis-forming energy commodities in the region; hence, they are not subject to endogenous shocks. Second exogeneity concern relates to the risk spillover between the two markets. The thing is that there is no ex ante mechanism of tying gas prices to oil prices in North America; so, gas price should solely be the result of supply and demand factors. On the contrary, the explicit integration of oil prices into gas prices is an issue in Continental Europe (Energy Charter, 2007). Furthermore, the choice of North American commodities is due to the sample issue as well. The European Energy Exchange (EEX) – the counterpart of NYMEX – has only three incomplete years of available natural gas price data, which is not enough for any sort of reasonable inference.

[15] Millions of British thermal units. 1 cf (cubic foot) equals 1028 btu.
[16] American Petroleum Institute (API) gravity measures the density of oil relative to water.

The sample spans from May 1994 to March 2009 and covers 3890 observations. May 2, 1994 is the date when Henry Hub natural gas price was first recorded. So, although WTI had been traded long before May 1994, this date is the starting observation of the sample common for these two commodities. The sample is divided into two periods: "in-sample" (May 1994 – December 2006) and "out-of-sample" (January 2007 – March 2009). The "in-sample" data are used to estimate the coefficients, while the "out-of-sample" data are used for forecasting.

Figure 1. Spot prices of WTI crude oil and Henry Hub natural gas



The dynamics of prices for the whole sample are depicted in Figure 1. The series tend to move together with occasional divergences. It is consistent with the fact that oil and natural gas are substitutes for some non-residential consumers; so, during gas shortages (e.g. 2001 winter "gas shock") prices start to converge since those who can substitute gas for oil find it profitable to do so. The comovement, however, is neither ideal nor should it be. Oil is traded internationally and its price is subject to shifts in international demand; while natural gas cannot be freely transported, so its markets are regional, and its prices are mostly affected by domestic demand.

We start formal investigation of the prices by testing whether the series are stationary. If the series are non-stationary (integrated of order *n*) one should difference those *n* times to get stationary ones. Otherwise further consideration is meaningless since means of the series are statistically non-constant. The Augmented Dickey-Fuller (ADF) test shows that oil and gas prices series in *levels* are non-stationary (see Table 1). Thus we generate the series of returns $r_{it} = \ln P_{it} - \ln P_{it-1}$ where i denotes oil or gas. Since the returns series are stationary, we can conclude that the original price series are integrated of the first order. Henceforth the returns series are used for our analysis.

Table 1.
ADF statistics for oil and gas

Series (in logs)	ADF statistics for levels (p-value)	ADF statistics for returns (p-value)
WTI crude oil	-1.45 (0.559)	-63.98 (0.000)
Henry Hub natural gas	-2.32 (0.165)	-58.54 (0.000)

The stationary returns series are shown in Figures 2 and 3. A note of caution: Henry Hub returns are delusively illustrated as being less volatile. This is because its scale is made wider than the one of WTI in order to fit the extreme deviations of 1996 and 2003 into the graph.

Figure 2. WTI returns

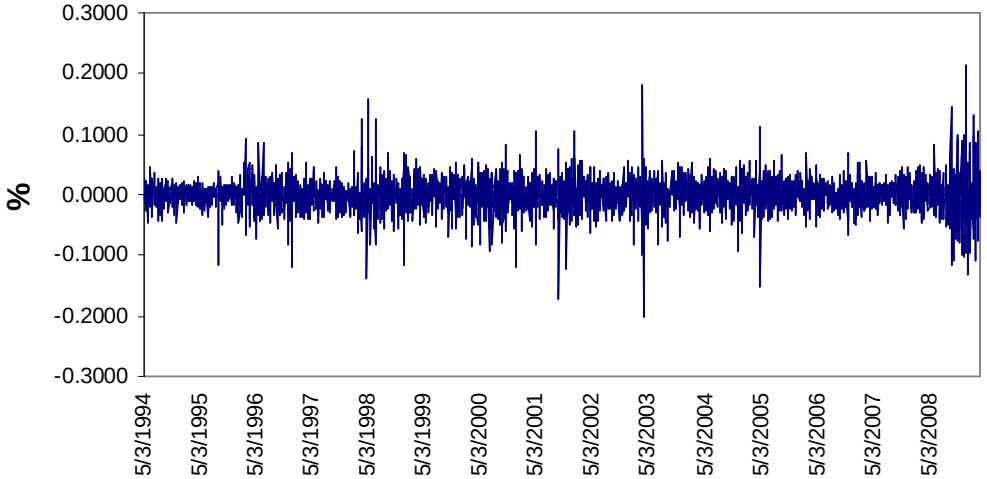
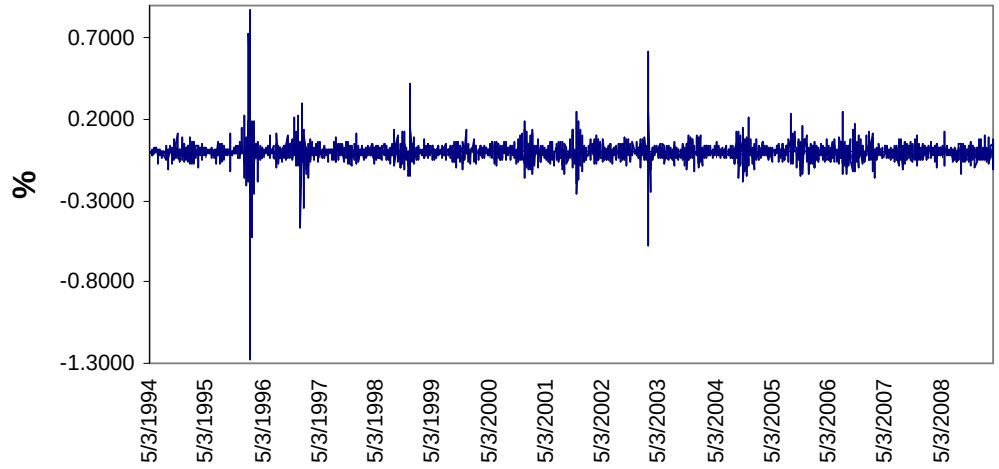


Figure 3. Henry Hub returns



The sample descriptive statistics with returns' distributions are given in Figures 4 and 5.

Figure 4. WTI descriptive statistics

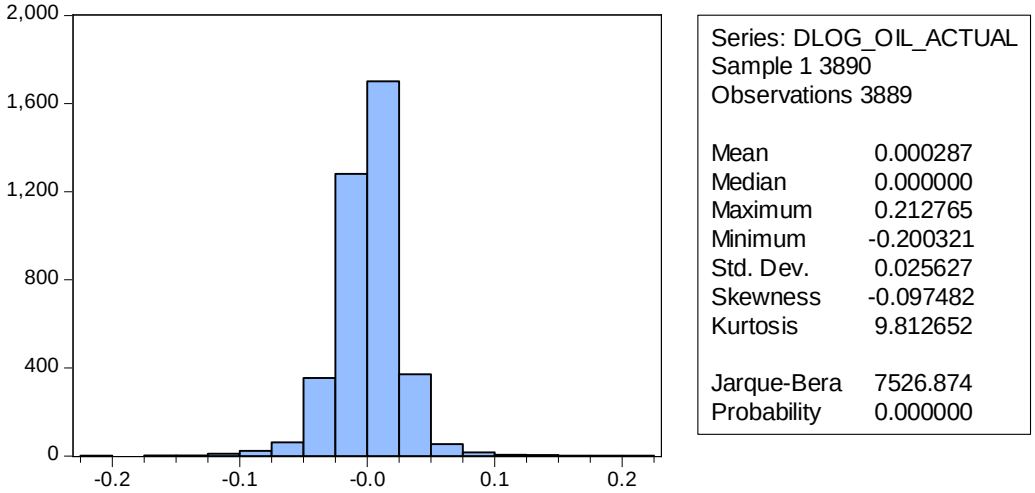
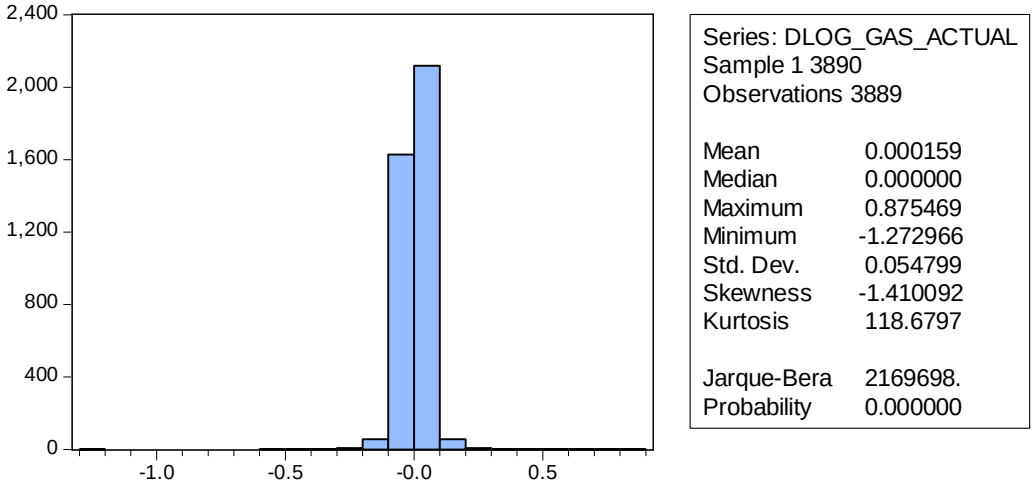


Figure 5. Henry Hub descriptive statistics



Neither commodity returns' distribution is normal. Both feature negative skewness and positive excess kurtosis. The former indicates at the asymmetry of the returns, while the latter is the evidence of fat tails. Another noteworthy fact is the larger volatility of the natural gas market. The maximum and minimum gas returns are larger than the respective values of WTI in absolute terms, and its standard deviation is two times larger as well.

4.2. Sample

As was mentioned, the sample has been divided into the modeling and forecasting sub-samples. The main characteristic of both sub-samples is their heterogeneity. The resultant impediment with respect to the "in-sample" period is the difficulty to fit a single model as it is usually easier to fit several different models for several homogeneous sub-samples. Despite this, the research employs the large single "in-sample" period since our objective is to test several models under the same conditions rather than tailor a certain model to a sub-sample and achieve its best possible performance.

The same feature applies to the "out-of-sample" period, however, again, disregarding this fact, we use the single sample since its heterogeneity presents an excellent opportunity to test models in two structurally different volatility environments – calm (January 2007 - September 2008) and unsteady (September 2008 - March 2009).

4. Research Findings

We first consider the choice of our models for both commodities for 95% confidence level. The 99% confidence level results are covered in Appendix III, however, the comparative aspects will be explicitly considered in this section. We then proceed to the extreme risk spillover results.

4.1. VaR estimation results

4.1.1. Standard Historical Simulation (HS) approach

Following the methodological description we estimate two VaR series – upside and downside. The initial HS VaR value is based on the sub-sample spanning from May 2nd 1994 to December 29th 2006. Further on, the most distant days are removed and the upcoming days are included into the distribution one by one. The "out-of-sample" estimation results for both commodities are given in Figures 6a and 6b.

Figure 6a. VaR for positive and negative WTI oil returns, HS approach

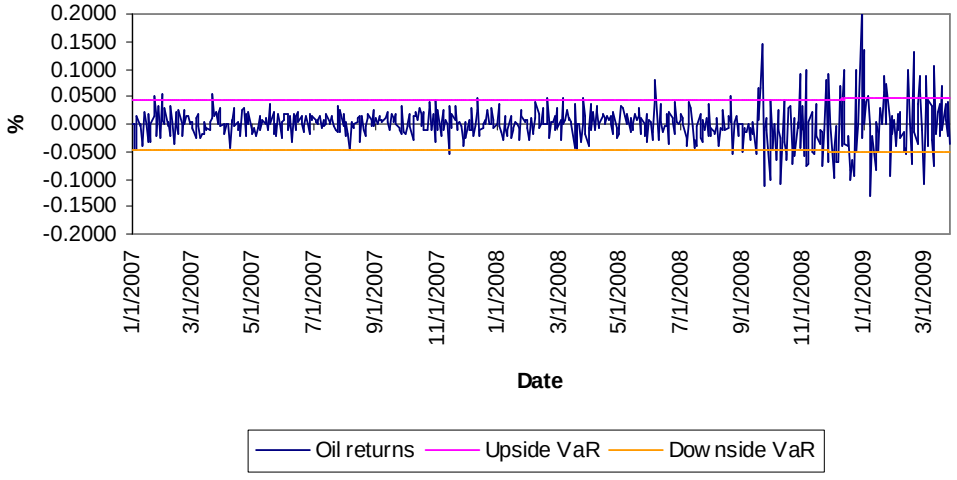
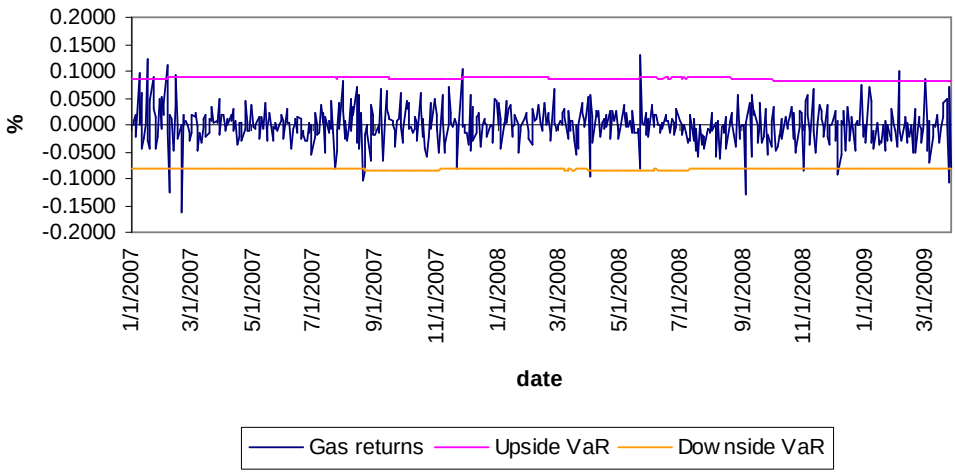


Figure 6b. VaR for positive and negative Henry Hub gas returns, HS approach



From the first visual examination, it is evident that the model exhibits permanent inflexibility. Both series for both commodities "tear" through the return peaks and bottoms. In the oil market, the explosion of volatility in September 2008 led to slight widening of the VaR series; however, since thousands of remote observations still have their impact on the distribution, this widening was disproportionately small relative to the size of return deviations. In the gas market, the situation is exactly the opposite: those are the extreme deviations of 1996, which have finally been dropped from the window by September 2008, and since then did not have any effect on the returns distribution. That's why we can see a slight VaR range contraction after the elimination of extremes.

Both figures indicate at the failure of the main assumption of HS method – that history repeats itself. The results show that neither relatively calm ("in-sample") past of oil market nor the volatile past of gas market can be extrapolated to the future.

One should still bear in mind, however, that if we took a smaller size of the window in this method, the weight of a single observation in the distribution of returns would be smaller, and there would be more shifts in the VaR series. So, again, the larger the size of the window the less

flexible a VaR is. However, since the purpose of this work is to assess the performance of different methods within identical conditions, we do not consider alternative sizes of HS window further.

4.1.2. Historical Simulation with ARMA forecasts (HSAF)

We have to estimate the ARMA model of the returns, and need to use stationary data. The results of stationarity tests have been reported in the data description section, that's why we readily start investigating the returns series.

We follow the Box-Jenkins methodology to identify the best model for the behavior of oil and natural gas returns. One has to bear in mind that the performance of HSAF method relies solely on the fit of the obtained ARMA model since unlike in GARCH methods here we do not model the returns' volatility. That's why all reasonable model alternatives deserve careful consideration.

The criteria that we evaluate our models against are Akaike Information Criterion (AIC), the general fit of the model (R-squared), statistical significance of the coefficients, as well as Breusch-Godfrey serial correlation test. Similar to Cabedo and Moya (2003) who apply HSAF methodology to Brent crude oil returns, we find that the models for WTI and Henry Hub returns require both AR and MA elements and can be approximated by ARMA(1,1) specification. The estimation results for the chosen models of WTI and Henry Hub returns are given in Table 2 below.

In both cases ARMA(1,1) model is one of those preferred by AIC and has the highest R-squared. Both AR and MA terms are economically and statistically significant at any conventional level of acceptance in both equations.

Table 2.
Estimation results for ARMA(1,1) model of WTI crude oil and Henry Hub returns.

	WTI returns			Henry Hub returns		
c	0.00036 (0.00031)			0.00031 (0.00107)		
AR(1)	0.86747[c] (0.06830)			-0.66257[c] (0.14757)		
MA(1)	-0.92040[c] (0.05763)			0.79043[c] (0.09867)		
R^2	0.012241			0.028312		
AIC	-4.643444			-2.891556		
Breusch-Godfrey LM test	0.13	0.25	0.26	0.93	0.09	0.15

Notes: a) Standard errors are in parentheses; b) *a*, *b*, and *c* superscripts denote significance at 10%, 5% and 1%; c) LM test entries are p-values for the Breusch-Godfrey serial correlation LM test for the 2nd, 3rd, and 4th lags. Insignificance at 5% means independent errors.

Since the sub-samples are heterogeneous, the choice of the best model requires uneasy trade-offs. In particular, a higher order ARMA model leads to less reliable coefficients for us (more so in case of Henry Hub returns), though the fit of such model is (naturally) larger. Thus, for example, the ideal model for natural gas equation in terms of AIC and R-squared is ARMA(4,3); however, we decide not to overcomplicate the choice, and opt for a simple-cum-parsimonious model. Had our sub-sample been more homogeneous, fewer compromises would be required.

After obtaining the "out-of-sample" forecasts and correcting them with the 0.05 and 0.95 percentiles of "in-sample" residuals' distribution, we get VaR series shown in Figures 7a and 7b.

Figure 7a. VaR for positive and negative WTI crude oil returns, HSAF approach

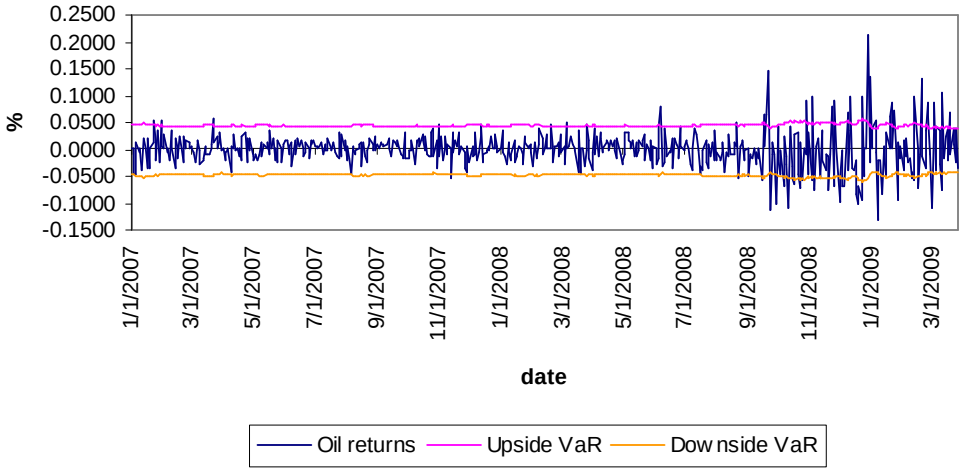
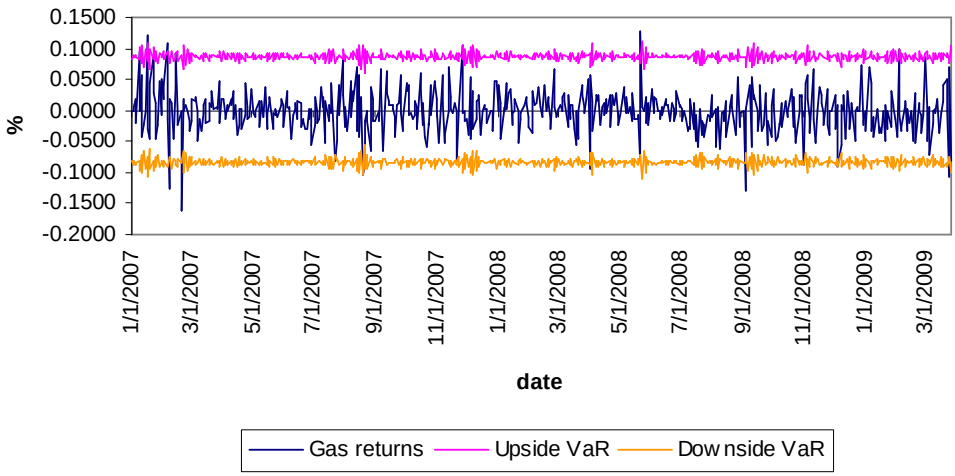


Figure 7b. VaR for positive and negative Henry Hub gas returns, HSAF approach



What we see on the figures is a "good effort" to have flexible predictions of risk but not more. ARMA behavior can clearly be viewed as an improvement over the rigid standard HS; yet, the general behavior remains the same – no major deviations from the VaR mean over time.

4.1.3. GARCH-N (Normal distribution)

First of all, further examination of the data indicates that there is significant ARCH effect in both returns series, which requires variance modeling. We estimate a GARCH model for WTI and Henry Hub returns. Based on the significance of ARCH terms in the variance equations, AIC values and ARCH LM test, we fit GARCH(1,1) model to both returns:

Table 3.
Estimation results for the GARCH-N (1,1) variance equation of WTI and Henry Hub returns.

	WTI returns			**Henry Hub returns**		
c	1.72E-05[c] (5.24E-06)			4.99E-05[c] (1.23E-05)		
ARCH(-1)	0.079500[c] (0.019887)			0.162706[c] (0.020200)		
GARCH(-1)	0.893265[c] (0.023400)			0.840058[c] (0.016268)		
R^2	0.000994			0.013444		
AIC	-4.73551			-3.629910		
ARCH LM test	0.07	0.13	0.22	0.82	0.93	0.98

Notes: a) Standard errors are in parentheses; b) *a*, *b*, and *c* superscripts denote significance at 10%, 5% and 1%; c) LM test entries are p-values for the ARCH LM test for serial correlation in squared residuals for the 2nd, 3rd, and 4th lags. Insignificance at 5% means independent errors.

Following manipulations with the "out-of-sample" forecasts and the standard errors described in 3.1.3, we obtain the VaR series shown in Figures 8a and 8b.

Figure 8a. VaR for positive and negative WTI crude oil returns, GARCH-N approach

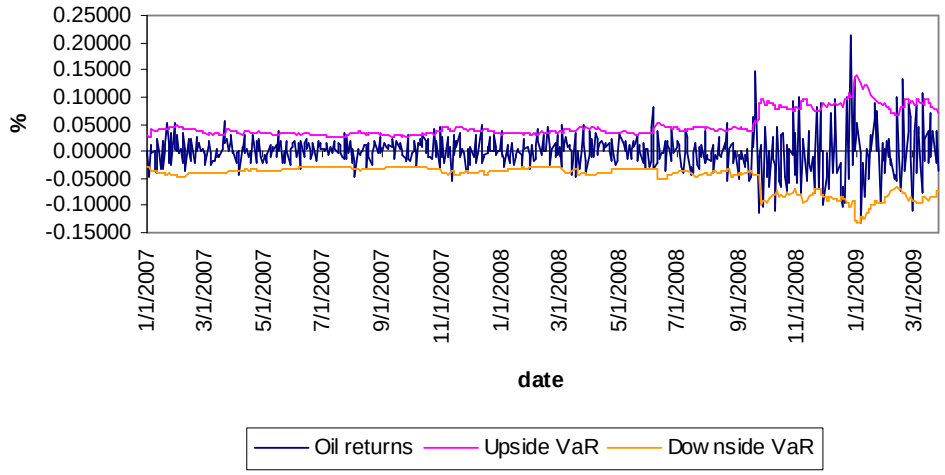
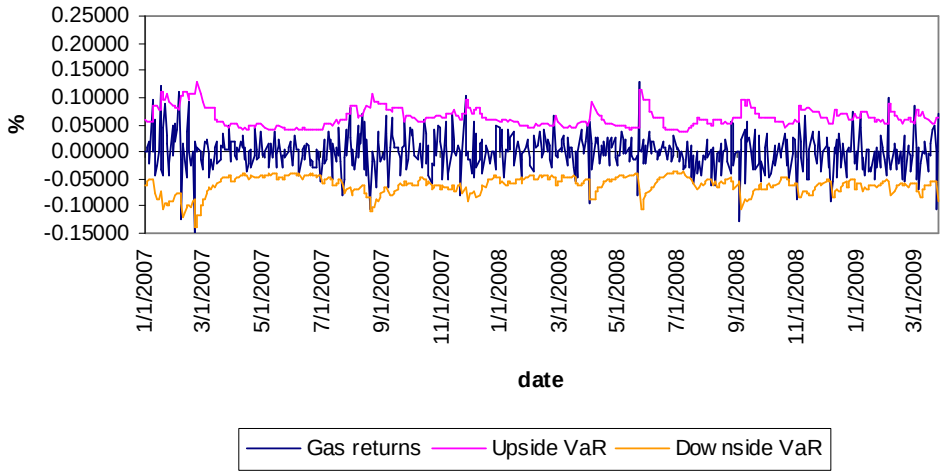


Figure 8b. VaR for positive and negative Henry Hub gas returns, GARCH-N approach



Volatility modeling brought about a clear improvement over the approaches of Historical Simulation family. There is more flexibility and, in general, the VaR series are able to correctly "guess" the dynamics of the "out-of-sample" returns. At least, visually, the models' performance

in predicting Henry Hub risk is better than WTI risk. Still, since the returns of both commodities exhibit fat tails, it might be that accounting for non-normality could introduce some improvement. GED-GARCH approach is the one to try.

4.1.4. GED-GARCH (Generalized Error Distribution)

In a similar fashion, we estimate the GARCH(1,1) model with the Generalized Error Distribution specification. Apart from traditionally provided output, Eviews reports the value of GED parameter (*k* in 3.1.4). The statistics of both models are given in Table 4 below.

Table 4.
Estimation results for the GED-GARCH (1,1) variance equation of WTI and Henry Hub returns.

	WTI returns			**Henry Hub returns**		
c	7.15E-06[c] (2.28E-06)			0.000102[c] (1.58E-05)		
ARCH(-1)	0.035640[c] (0.005797)			0.171933[c] (0.021343)		
GARCH(-1)	0.951428[c] (0.008325)			0.773284[c] (0.021077)		
R^2	0.003325			0.000959		
AIC	-4.826942			-3.863363		
ARCH LM test	0.22	0.33	0.23	0.89	0.97	0.99
GED parameter	1.14509			0.98424		

Notes: a) Standard errors are in parentheses; b) *a*, *b*, and *c* superscripts denote significance at 10%, 5% and 1%; c) LM test entries are p-values for the ARCH LM test for serial correlation in squared residuals for the 2nd, 3rd, and 4th lags. Insignificance at 5% means independent errors.

The fact that both GED parameters are very different from 2 is a vivid demonstration of the normality assumption irrelevance with respect to these two commodities. As expected, GED parameter of WTI returns model is closer to 2 than that of Henry Hub returns because the distribution of Henry Hub is much more leptokurtic. After replicating a theoretical distribution with each GED parameter in Excel, we can find the exact CDF critical values at 95% and 99%:

Table 5.
Critical values of cumulative density function with different *k* parameters.

	95%	**99%**
Normal distribution (*k*=2)	1.645	1.96
WTI GED (*k*=1.14509)	1.646	2.68
Henry Hub GED (*k*=0.98424)	2.678	2.78

The following are the VaR series obtained after correcting forecasted returns by the product of obtained CDF critical values and standard errors of the model:

Figure 9a. VaR for positive and negative WTI crude oil returns, GED-GARCH approach

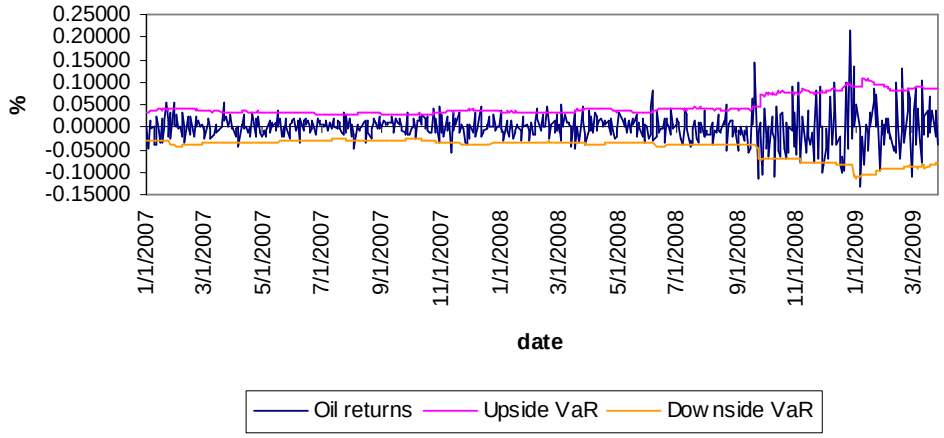
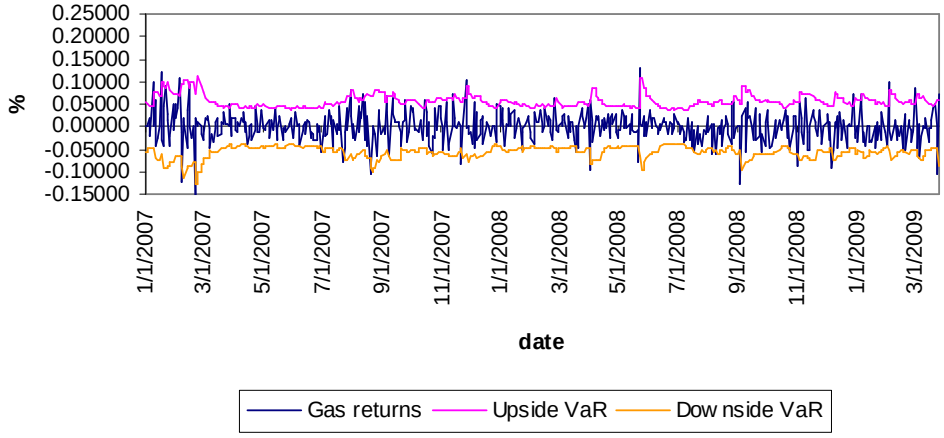


Figure 9b. VaR for positive and negative Henry Hub gas returns, GED-GARCH approach



The visual comparison of GARCH-N and GED-GARCH seems a bit complicated by the similarity of VaR dynamics. For more formal inference about the performance of each of the four described methods, we evaluate them according to the criteria cited in 3.2.

4.2. VaR performance assessment

We start the formal investigation with the 95% confidence level (Table 6)[17].

Table 6.
The indicators of methods' performance at 95% confidence level.

	Binary Loss function (upside/downside)		**Kupiec LR statistic (upside/downside)**		**Mean Relative Scaled Bias (upside/downside)**	
WTI						
HS	0.05651	0.06336	0.5142[a]	2.05794[a]	0.02522[W]	0.05043
HS ARMA	0.05822	0.05993	0.80807[a]	1.16442[a]	0.01454	0.05538[W]
GARCH-N	0.05832	0.05489	0.80818[a]	0.28467[a]	-0.00223	-0.0339
GED-GARCH	0.06346	0.06003	2.05805[a]	1.16453[a]	-0.03753[B]	-0.07191[B]
Henry Hub						
HS	0.01541	0.01712	19.87019	17.55720	0.18608	0.16902
HS ARMA	0.0137	0.01712	22.40933	17.55720	0.2016[W]	0.18390[W]
GARCH-N	0.03431	0.03602	3.38143[a]	2.64618[a]	-0.15227	-0.14777
GED-GARCH	0.04803	0.04460	0.04836[a]	0.37125[a]	-0.23541[B]	-0.20516[B]

Notes: a) *B* and *W* superscripts in the Mean Relative Scaled Bias denote "best" and "worst" across the methods respectively; b) *a* and *b* superscripts denote significance of Kupiec LR statistics at 5% and 1% respectively; c) the best modeling approach is highlighted in bold.

Consider oil market first. Even though all models underpredict 5% incidence of both upside and downside risk, they are still roughly correct to the extent that all of them formally pass the Kupiec test, and thus, are adequate. With respect to both Historical Simulation models, however, we must consider the real meaning of these statistics. From the graphs it's clear that if there were longer high volatility at the end of "out-of-sample" period, the performance of the models would deteriorate "in proportion" to the added period since VaR series are unable to capture any of the latest largest peaks. It is the calm period of January 2007 – October 2008 that positively told upon the performance of both Historical Simulation approaches. So the poor performance during the volatile period has been averaged out by good performance during the calm period.

[17] For the sake of brevity, the quadratic loss function (QLF) values are not reported since being marginally different from the binary loss function (BLF) values, they do not change the general pattern of the results.

The rigidity of Historical Simulation VaR measures is reflected in the efficiency criterion MRSB: on average, the Historical Simulation methods have been less efficient than GARCH methods. Combining common sense (graphs) and statistical results (tables) we can infer that HS and HSAF methods have been the least accurate and efficient VaR series for oil market at 95% confidence level.

Since we made sure of Historical Simulation methods sub-optimality, we now face the choice of GARCH-N against GED-GARCH. There is, however, no ex ante reason why any of them would be better than another at 95% confidence level. Both GARCH methods pass the Kupiec test, so we again consider the efficiency criterion. The table shows that GED-GARCH is more efficient at 95% because it has the smallest mean relative bias. This makes our choice: GED-GARCH is *accurate enough* and is also the most efficient method. A relevant consideration with respect to our way of selection is the fact that GARCH-N actually produces more accurate VaR in *economic* terms than GED-GARCH in both upside and downside cases. The thing is that we choose the method, which is able to produce *statistically* accurate indicators. Once the measures are found statistically significant we don't care *how* significant they are. In this way we limit ourselves to statistical rather than economic significance, just like the most previous papers did when employing the Kupiec test.

The assessment of natural gas returns at 95% is done in a similar fashion. This time, both the table and the graphs show that Historical Simulation methods are completely inadequate. Kupiec test rejects the $\frac{M}{N} = c$ hypothesis (see p. 21) because the models consistently overestimate VaR over the sample, and MRSB shows that HS and HSAF are more inefficient. So again the choice of a better model pins down to GARCH-N versus GED-GARCH, and following the same logic as for oil market we opt for GED-GARCH.

Now we shift to the consideration of 99% confidence level (Table 7)[18].

Table 7.
The indicators of methods' performance at 99% confidence level.

	Binary Loss function (upside/downside)		Kupiec LR statistic (upside/downside)		Mean Relative Scaled Bias (upside/downside)	
WTI						
HS	0.02568	0.01712	10.15742	2.48153[a]	0.04932[W]	0.12151[W]
HS ARMA	0.03082	0.01541	16.50315	1.49315[a]	0.01323	0.11795
GARCH-N	0.01544	0.01201	1.49318[a]	0.22288[a]	-0.08066[B]	-0.16303[B]
GED-GARCH	0.00686	0.00172	0.65202[a]	6.17427[b]	0.01811	-0.07643
Henry Hub						
HS	0	0	N/A	N/A	0.27921	0.27572
HS ARMA	0	0	N/A	N/A	0.39265[W]	0.28566[W]
GARCH-N	0.01029	0.01544	0.00496[a]	1.49318[a]	-0.36527[B]	-0.32588[B]
GED-GARCH	0.00858	0.01372	0.12540[a]	0.73096[a]	-0.30658	-0.23549

Notes: a) *B* and *W* superscripts in the Mean Relative Scaled Bias denote "best" and "worst" across the methods respectively; b) *a* and *b* superscripts denote significance of Kupiec LR statistics at 5% and 1% respectively; c) the best modeling approach is highlighted in bold.

Again, we start with oil market. The case of Historical Simulation methods failure is more pronounced here. Now both HS and HSAF VaR measures are insignificant; this time they actually *under*estimate the true risk. At the same time, HS is the most inefficient method. This gives a reason to question the findings of Cabedo and Moya (2003) that HSAF method, even if not as accurate, is more efficient than GARCH-N, and hence, they argue, is more preferred at the 99% confidence level for Brent returns. We find however that HSAF is both inaccurate and inefficient. We think the reason for such a discrepancy is general vulnerability of HSAF methodology to the model's fit. Since HSAF does not feature volatility modeling, the returns must be well approximated by AR and MA coefficients. However, since HSAF is constrained to use only own lags without other relevant exogenous variables, a single instance success of the methodology should not be generalized.

[18] See the VaR graphs in Figures A.3a-h in the Appendix III.

Between GARCH-N and GED-GARCH models we opt for GARCH-N this time. First, GED-GARCH overestimates the risk for negative returns, so it passes the Kupiec test only at 1% significance (in spite of a generally nice look of the VaR fit, the model "fails" to predict only 1 (!) returns excess). Second, GARCH-N turns out to be more efficient than GED-GARCH.

One might however be surprised by the fact that GARCH-N is preferred over GED at the 99% confidence level. After all, didn't we show that oil returns exhibit fat tails? The thing is that indeed GARCH-N underpredicts oil risk at this confidence level. But it does so within the confidence interval; so the failure rates in positive and negative returns of 1.54% and 1.2% respectively are statistically equal to the required 1%; thus, the model passes the Kupiec test. On the other hand, GED-GARCH fails less frequently at 99%, but in case of negative returns it actually "overperforms", and that's why not only is inefficient but also barely passes the Kupiec test at 1%. In other words, the relative benefit from GARCH-N efficiency is larger than the relative cost from its inaccuracy. Our model choice in this case is GARCH-N, and this finding is different from the one of Fan et al. (2008). They support the choice of GED-GARCH over GARCH-N at 99% confidence level for WTI and Brent returns, and they explain it solely by the Kupiec test statistics without consideration of any efficiency criteria.

Same applies to gas returns. This case is the harshest demonstration of Historical Simulation methods' deficiencies. Kupiec test statistic could not even be calculated because it would require division by zero (zero is the number of times actual Henry Hub returns exceed VaR measures in this case). The gas market is very volatile but its volatility was more or less homogeneous during the forecast period. That's why there were no volatility explosions which could produce some if any intersections. This is also the evidence of the GARCH modeling necessity: HSAF is of course less rigid than HS but the overall trend of its VaR still remains linear and the method is

not very helpful to fit the Henry Hub risk. Again, in this case GARCH-N is preferred over GED-GARCH since both methods pass the Kupiec test while GARCH-N is more efficient.

To sum up, Historical Simulation methods are found to be the least successful candidates to measure the risk in these two commodities. Between GED-GARCH and GARCH-N, the former is a better option for 95% confidence level, while the latter is better for 99% confidence level.

4.3. Risk spillover results

The previous sub-section is indispensable for the present one. As GED-GARCH is found the best performing method at 95% confidence level, we use the upside and downside VaR series of WTI and Henry Hub using GED-GARCH in order to investigate the extreme risk spillover between these two markets. Following the methodological steps in 3.3 we calculate the test statistics and corresponding p-values for upside and downside risk running in both directions. The results are given in Table 8.

Table 8.
The test statistics of Granger causality in risk for the two energy markets.

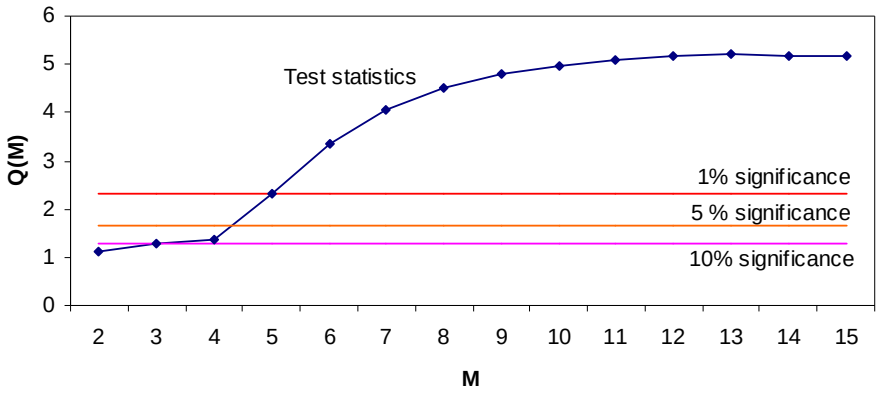
	Upside risk			**Downside risk**		
	M=5	M=10	M=15	M=5	M=10	M=15
WTI → Henry Hub	2.31^{b} (0.01)	4.98^{b} (0.00)	5.16^{b} (0.00)	1.37^{a} (0.09)	0.45 (0.33)	0.45 (0.33)
Henry Hub → WTI	-0.43 (0.67)	-0.30 (0.62)	-0.53 (0.7)	-0.28 (0.61)	-0.44 (0.67)	0.02 (0.49)

Notes: a) The reported values are the test statistics with corresponding upper-tailed p-values in parentheses; b) *a* and *b* superscripts denote significance of the statistics at 10% and 5% respectively;

First, consider the causality from WTI to Henry Hub. The p-values for the upside risk causality test are well below 5%, which indicates at significant extreme risk spillover from oil to natural gas market during the period under consideration. The case with downside risk is less clear-cut because we can confirm significant test statistics only at 10% and only at M=5. Though weaker, however, the spillover is still in place.

One should also note the dynamics of the test statistics at different values of M. In case of upside risk, growing M is associated with more statistically significant risk spillover, that is, the risk from the oil market is translated both shortly (within 5 days) and accumulates with time (within 15 days). On the contrary, the downside risk is not so protracted: weak within 5 days, it gets even weaker after 10 days and does not change 15 days after the negative return shock in the oil market. A more general representation of the risk spillover dynamics from WTI market to Henry Hub market is given in Figures 10a and 10b.

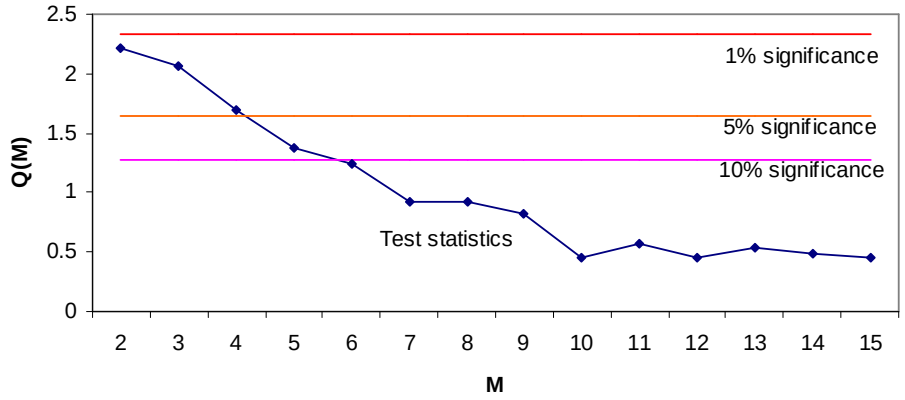
Figure 10a. The upside risk spillover dynamics in oil-to-gas direction



Mirroring the Table 8, the Figure 10a shows that the upside risk spillover becomes significant after 4-5 days and starts decaying after approximately 13 days. On the contrary, the downside risk (Figure 10b) is significant only at the beginning, then it strictly diminishes and after 6 days becomes insignificant even at 10%[19].

[19] The graphs however do not report the test statistics for M=1. The use of Daniell kernel function specification leads to heavily exaggerating test statistics for M=1; e.g. the M=1 statistics for the upside risk spillover from WTI to Henry Hub is equal to 107.5 (that is, very statistically significant); however, cases when the upside risk translates from oil to gas market after one day do not exist at all. The visual data examination shows that starting with M=2 the test statistics reflect the reality much more accurately. The absence of M=1 test statistics, however, does not contradict the general pattern of our results.

Figure 10b. The downside risk spillover dynamics in oil-to-gas direction



Quite expectedly, the Granger causality in risk from Henry Hub to WTI is non-existent according to the obtained results (Table 8). The p-values of the test statistics are larger than 10% in case of both upside and downside risk and at any M. The evolution of risk spillover over time is less clear-cut in this case, so the graphs are not provided.

4.3.1. Robustness check

One should note, however, that Granger causality does not presume causality in direct meaning. That's why the obtained results of Granger causality in risk should not be taken literally. While we find significant risk spillover from WTI to Henry Hub, the WTI risk can Granger-cause Henry Hub risk in case if a third process causes both of them with different lags. For example, both Henry Hub and WTI risk could be a result of general market volatility, but we cannot consider such possibility as a shortcoming of the method.

Nonetheless we check whether we can indeed speak about *direct* risk causality. We proxy general market volatility by General Volatility Index (designated as VIX). VIX is a measure of

S&P 500 expected volatility during next 30 days and is often referred to as "fear" index. Rather narrow in representing general market volatility, this index, we believe, is sufficient to account for major volatile episodes. If VIX is found to Granger-cause the dispersion in Henry Hub *and* WTI, this would cast doubt on our results.

Testing the impact of VIX on the size of returns, however, would be incorrect; that's why we use returns *in absolute value*. Intuitively, if VIX increases, the values of both WTI and Henry Hub returns are expected to increase in absolute value if they are affected by general market volatility. Another note of caution is that VIX series are non-stationary, so we use differenced series (Table A.1). We do the following regressions:

$$\begin{cases} \text{abs(oil)}_t = \alpha_0 + \sum_{j=1}^{n} \alpha_{1j} \text{abs(oil)}_{t-j} + \sum_{j=1}^{n} \alpha_{2j} \text{dVIX}_{t-j} + \varepsilon_t \\ \text{abs(gas)}_t = \beta_0 + \sum_{j=1}^{p} \beta_{1j} \text{abs(gas)}_{t-j} + \sum_{j=1}^{p} \beta_{2j} \text{dVIX}_{t-j} + \varepsilon_t \end{cases},$$

where abs(oil)_t and abs(gas)_t are the series of WTI and Henry Hub returns in absolute value, and *p* is not necessarily equal to *n*.

At *n*=4 and *p*=3, we find that neither Henry Hub nor WTI absolute returns are Granger-caused by VIX (the models are provided in Appendix IV). Undoubtedly, the nature of this robustness check concentrating on *dispersion* is different from the nature of *extreme risk* spillover methodology. Nonetheless, the fact that general market volatility does not Granger-cause any of the two returns series is believed to roughly confirm that the source of extreme risk in Henry Hub market can indeed be the risk in WTI market.

5. Concluding Remarks

Based on the prices of Western Texas Intermediate (WTI) crude oil and Henry Hub natural gas, we test the performance of four Value-at-Risk approaches and assess them against several accuracy and efficiency criteria. The results indicate that at 95% confidence level GED-GARCH method performs better than any other alternative. The VaR series obtained with this method are statistically accurate and are the least likely to result in forgone profits from speculation. The VaR series calculated with GED-GARCH are used further to investigate the extreme risk spillover between oil and natural gas markets. Using the Hong's concept of Granger causality in risk (2002), we find that there is significant risk spillover from oil market to natural gas market during the period in consideration, while there is no spillover in the reverse direction. Moreover, the upside risk spillover from oil to gas markets is found to be more statistically (and economically) significant and protracted than the downside risk spillover.

Our results have some important implications. First, it follows that participants of the natural gas market are relatively more vulnerable to the external risk than those of oil market. Of course, the results we obtained cannot be taken unambiguously since there are numerous other factors having impact on both natural gas and oil prices. Nonetheless, this evidence shows that ceteris paribus there is a point in being more cautious in trading natural gas than oil. And in spite of the absence of ex ante pricing mechanism tying the two markets in the US, WTI risk can indeed be a good basis for the risk estimation in Henry Hub.

Second, significant and protracted *upside* risk spillover from WTI to Henry Hub versus weak and short *downside* one is an indicator of opportunities asymmetry between buyers and sellers of natural gas. The positive oil price shocks are found to be translated to the gas market for a long while, which is advantageous for gas sellers. However, there is no "opportunity compensation"

for buyers since the transmission of negative oil price shocks is weak and short, which leaves gas buyers in a more risky environment than gas sellers. This asymmetry is magnified by the fact that gas sellers – large international companies – usually have full-fledged risk management departments and can hedge themselves against a potential misfortune. Yet when gas buyers are concerned with risk they tend to address companies specializing in risk management advice, and, to the best of our knowledge, such companies are quite scarce on the global market[20].

Among limitations and possible extensions we would highlight the following. The first one concerns HSAF methodology, which performed rather poorly in this work and was eliminated too easily. Since we tried to replicate the exact features of Cabedo and Moya work (2003), we did not opt for enrichment of the model by relevant exogenous variables. Rather we stuck to lags specification only. The success of this model is in its high fit; hence, we recommend considering broader HSAF model specifications in future works.

Second, due to space limitations we did not consider the reasons for drastic dissimilarity between upside and downside risk spillover running from WTI to Henry Hub. To our mind, a deeper analysis of these reasons involving each market's inherent features would definitely contribute to the topic.

Finally, we would like to bring up the policy aspect of risk management as well. From our personal communication with a representative of an energy risk management company in Budapest, we can infer that there is a detachment of academia and business with respect to this field. The benefit from using the VaR methods, which proved to work well (say, based on Monte-Carlo simulation), is said to be larger than the benefit from introducing and testing new ones. Thus it is no surprise that numerous advancements in energy risk management literature

[20] E.g. Encore International is one of only 4 (!) independent energy risk management services providers in Europe.

are left in dust. The integration of theory and practice is of utmost importance for successful risk management, especially, in today's unstable environment.

It is natural to conclude that irrespective of whether oil or gas, sellers or buyers are concerned, risk quantification and management in energy markets is an indispensable way of business survival and prosperity. With ever increasing liberalization and competition, the VaR methods had better not be used separately; rather, intelligent use of their combination would enable a risk manager to get a more comprehensive estimation of market risk.

Appendices

Appendix I. GED for different k parameters

Figure A.1a

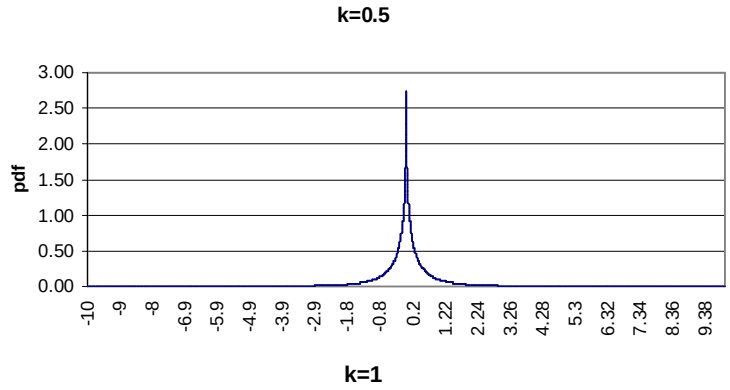


Figure A.1b

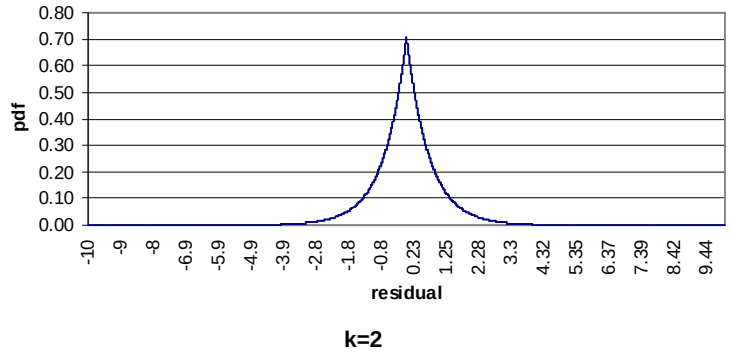


Figure A.1c

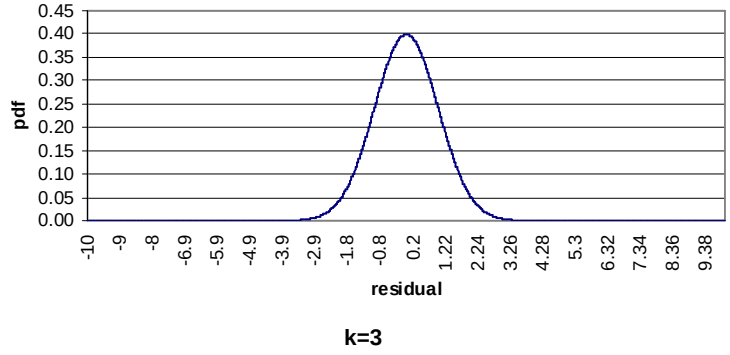
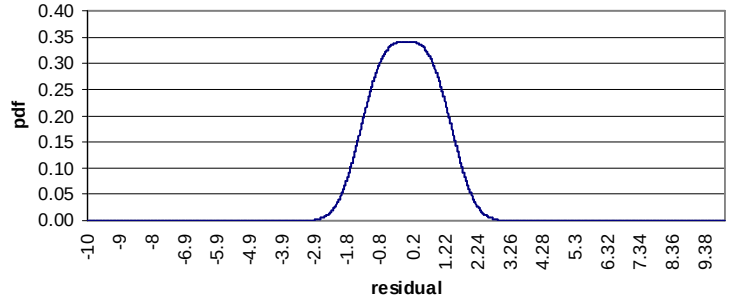
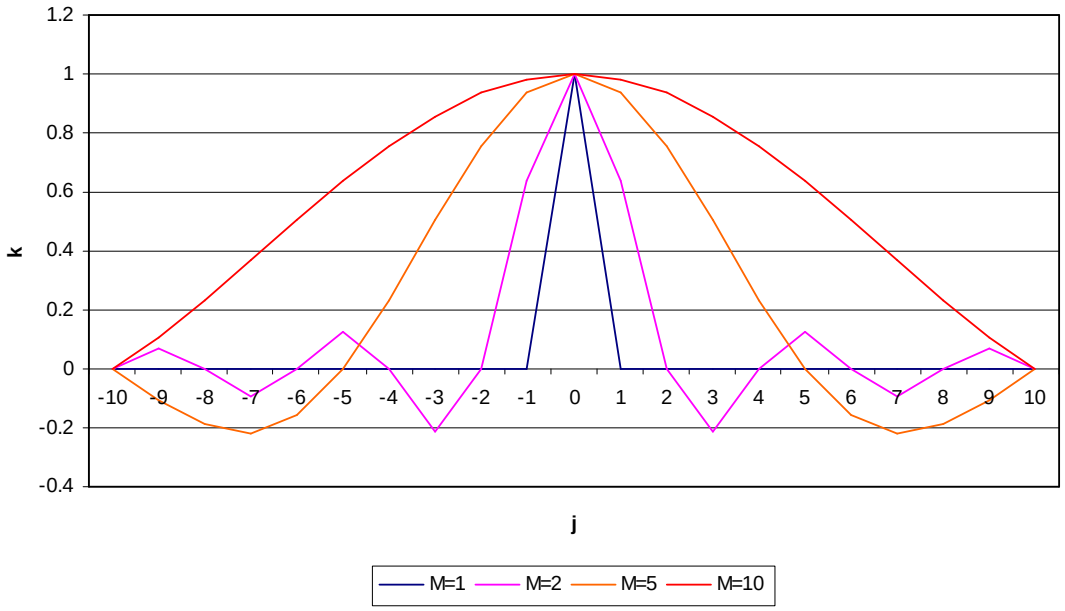


Figure A.1d



Appendix II. The shape of Daniell kernel function

Figure A.2. Daniell kernel function shape for different arguments



Note that the first intersection (to the left and to the right of zero) with the horizontal axis takes place at the corresponding M. E.g. the kernel function with M=5 makes the first intersection at $j = \pm 5$. The hump-shaped form of the function ensures that the remote lags (yet within M range) are given smaller weights.

Appendix III. Value-at-Risk at 99% confidence level

Historical Simulation – Standard Approach

Figure A.3a. VaR for positive and negative WTI oil returns

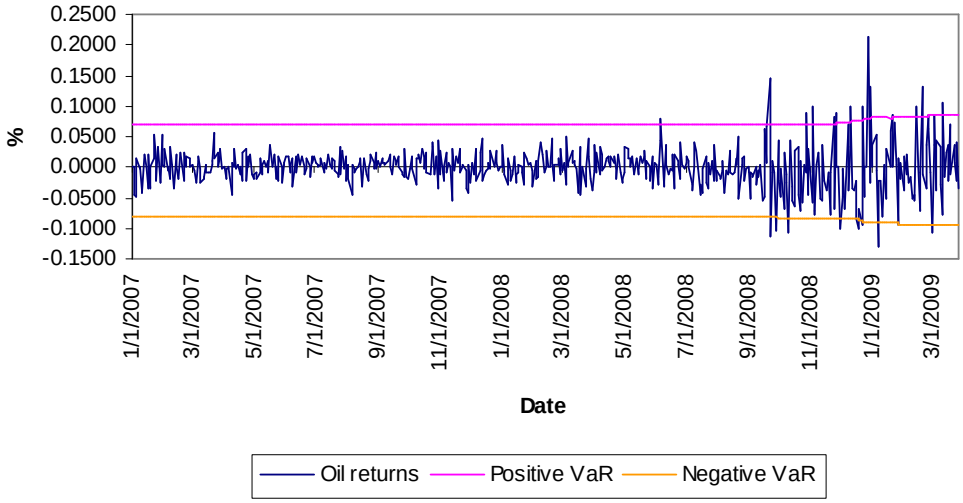
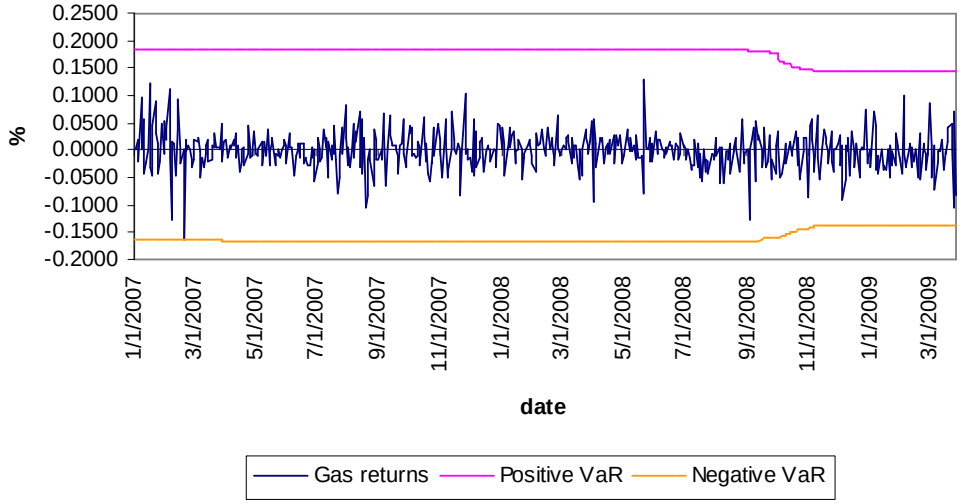


Figure A.3b. VaR for positive and negative gas returns



Historical Simulation – ARMA Forecasts

Figure A.3c. VaR for positive and negative oil returns

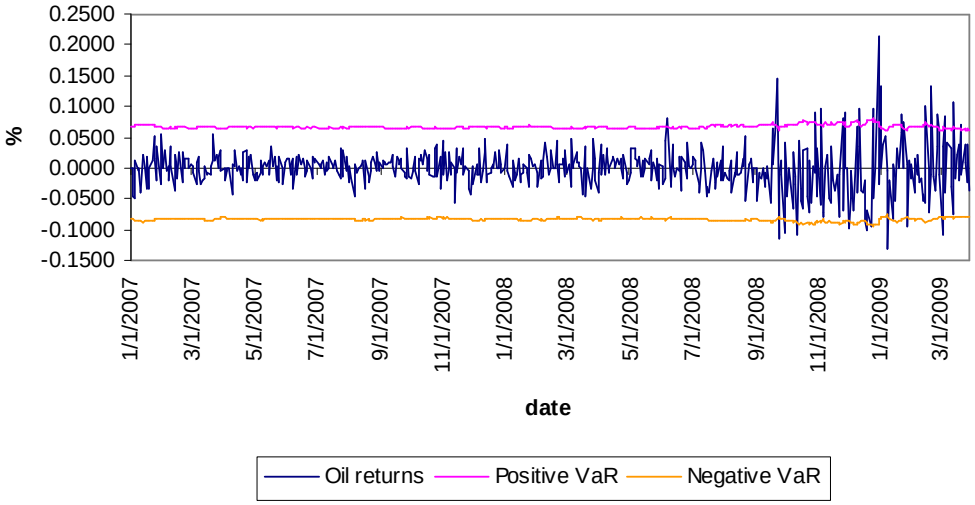
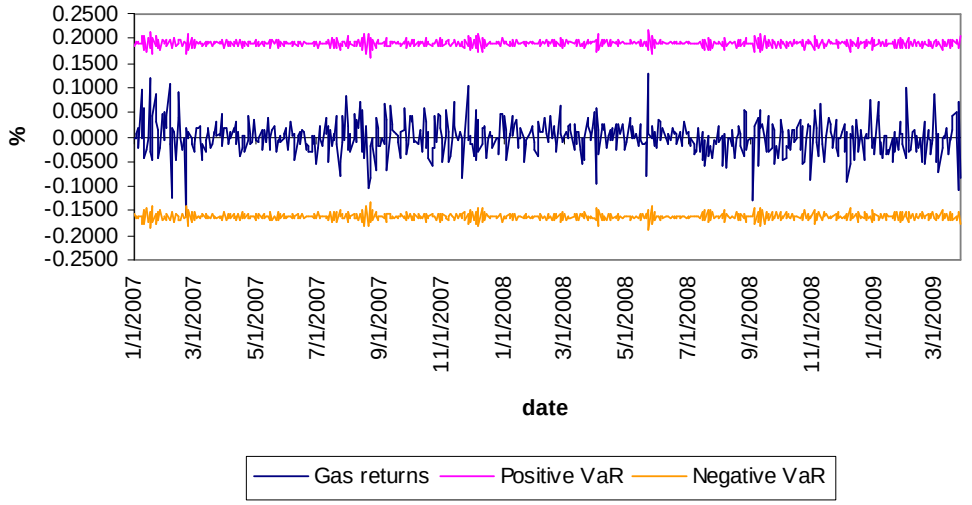


Figure A.3d. VaR for positive and negative gas returns



GARCH-N

Figure A.3e. VaR for positive and negative oil returns, GARCH-N approach

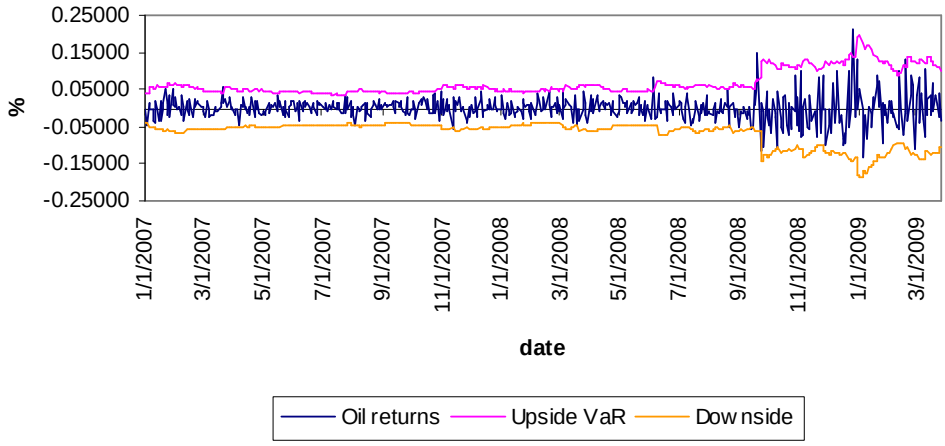
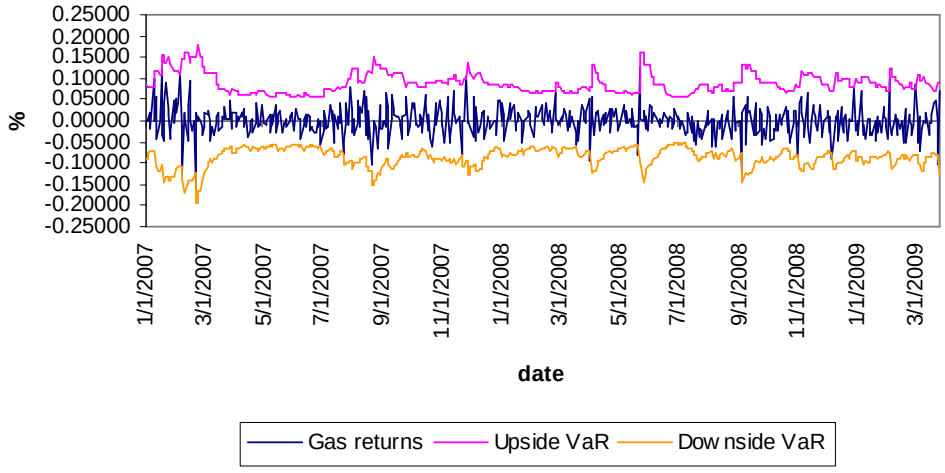


Figure A.3f. VaR for positive and negative gas returns



GED-GARCH

Figure A.3g. VaR for positive and negative oil returns, GED-GARCH approach

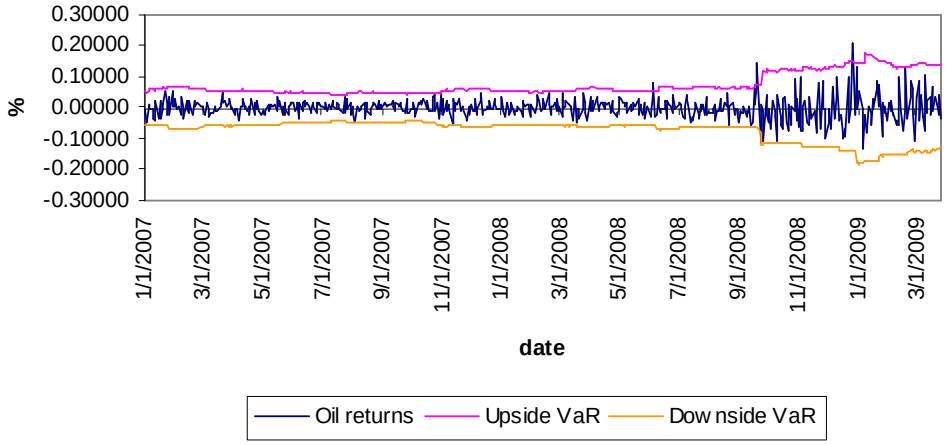
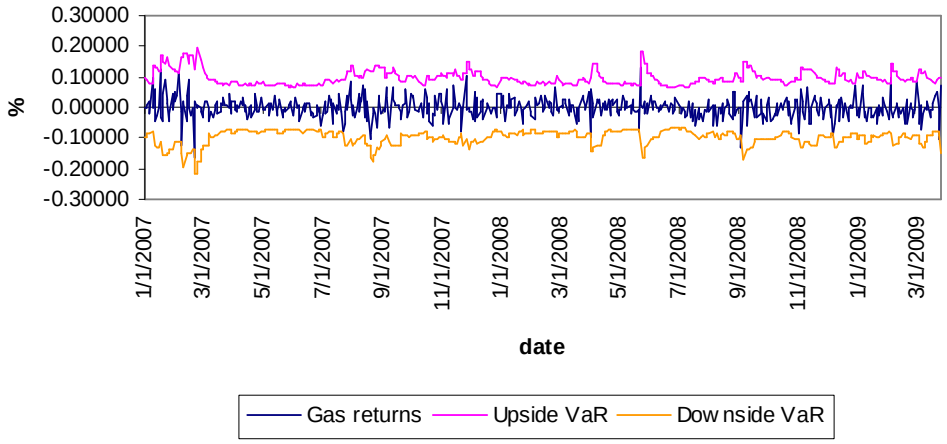


Figure A.3h. VaR for positive and negative Henry Hub gas returns, GED-GARCH approach



Appendix IV. Robustness check statistics

Table A.1.
ADF statistics for the General Volatility Index

Series (in logs)	ADF statistics for levels (p-value)	ADF statistics for differences (p-value)
VIX	-1.65 (0.454)	-21.80 (0.000)

Table A.2.
WTI Granger causality estimation results

abs(oil)	
c	0.009064[c] (0.00163)
abs(oil) (-1)	0.080762 (0.058149)
abs(oil) (-2)	0.157463[b] (0.076422)
abs(oil) (-3)	0.135724[b] (0.062599)
abs(oil) (-4)	0.234741[c] (0.079403)
dVIX (-1)	-0.037318 (0.02872)
dVIX (-2)	-0.04996[a] (0.027991)
dVIX (-3)	0.042259 (0.029801)
dVIX (-4)	0.026963 (0.030041)
R^2	0.173814
AIC	-4.68254
F-statistic for joint significance of dVIX (p-value)	2.137 (0.075)

Notes: a) Standard errors are in parentheses; b) *a*, *b*, and *c* superscripts denote significance at 10%, 5% and 1%.

Table A.3.
Henry Hub Granger causality estimation results

abs(gas)	
c	0.021696[c] (0.002007)
abs(gas) (-1)	0.009651 (0.047549)
abs(gas) (-2)	0.080596[a] (0.050362)
abs(gas) (-3)	-0.006597 (0.04174)
dVIX (-1)	0.039041 (0.025128)
dVIX (-2)	0.025229 (0.025668)
dVIX (-3)	0.00194 (0.026454)
R^2	0.011161
AIC	-4.676683
F-statistic for joint significance of dVIX (p-value)	0.978 (0.403)

Notes: a) Standard errors are in parentheses; b) *a*, *b*, and *c* superscripts denote significance at 10%, 5% and 1%.

References

Literature

Bollerslev, Tim. 1986. "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, 31: 307-327.

Cabedo, J. David and Ismael Moya. 2003. "Estimating Oil Price 'Value at Risk' Using the Historical Simulation Approach." *Energy Economics*, 25: 239-253.

Cook, Linda. 2005. "The role of LNG in a global gas market." Speech given at the Oil & Money conference, London.

Davis, Ann. 2006. "Blue Flameout: How Giant Bets on Natural Gas Sank Brash Hedge-Fund Trader." *Wall Street Journal.* September 19.

Energy Charter. 2007. "Gas Pricing" In *Putting a Price on Energy: International Pricing Mechanisms for Oil and Gas.* Available at: http://www.encharter.org/fileadmin/user_upload/document/Pricing_-_chapter_4.pdf, accessed May 30, 2009.

Energy Information Administration (EIA). 2006. Frequently Asked Questions. Available at: http://tonto.eia.doe.gov/ask/crude_types1.html, accessed May 30, 2009.

Engle, Robert. 1982. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of UK Inflation." *Econometrica*, 50: 987-1008.

Ewing, Bradley T., Farooq Malik, and Ozkan Ozfidan. 2002. "Volatility Transmission in the Oil and Natural Gas Markets." *Energy Economics*, 24: 525-538.

Fan, Ying, Yue-Jun Zhang, Hsien-Tang Tsai, and Yi-Ming Wei. 2008. "Estimating 'Value at Risk' of Crude Oil Price and Its Spillover Effect Using the GED-GARCH Approach." *Energy Economics*, 30: 3156-3171.

Giot, Pierre and Sèbastien Laurent. 2003. “Market Risk in Commodity Markets: A VaR Approach.” *Energy Economics*, 25: 435-457.

Hendricks, Darryll. 1996. “Evaluation of Value at Risk Models Using Historical Data.” *Federal Reserve Bank of New York Economic Policy Review*, April: 39-70.

Hong, Yongmiao. 2002. “Granger Causality in Risk and Detection of Risk Transmission between Financial Markets.” Cornell University Department of Economics and Department of Statistical Science Working Paper.

Hong, Yongmiao. 2003. “Detecting Extreme Risk Spillover between Financial Markets.” Cornell University Department of Economics and Department of Statistical Science Working Paper.

Hong, Yongmiao, Yanhui Liu, and Shouyang Wang. 2009. “Granger Causality in Risk and Detection of Extreme Risk Spillover between Financial Markets.” *Journal of Econometrics*. Unpublished.

Hopper, Gregory P. 1996. “Value at Risk: A New Methodology for Measuring Portfolio Risk.” *Federal Reserve Bank of Philadelphia Business Review*, July-August.

Hung, Jui-Cheng, Ming-Chih Lee, and Hung-Chun Liu. 2008. “Estimation of Value-at-Risk for Energy Commodities via Fat-Tailed GARCH Models.” *Energy Economics*, 30: 1173-1191.

Jorion, Philippe. 1997. “Approaches to measuring VaR.” In *Value at Risk: The New Benchmark for Controlling Market Risk*. New York: McGraw-Hill.

Kupiec, Paul H. 1995. “Techniques for Verifying the Accuracy of Risk Measurement Models.” *Journal of Derivatives*, 3: 73-84.

Këllezi, Evis and Manfred Gilli. 2000. “Extreme Value Theory for Tail-Related Risk Measures.” University of Geneva Working Paper, Geneva.

Nelson, Daniel B. 1991. "Conditional Heteroskedasticity in Asset Returns: A New Approach." *Econometrica*, 59(2): 347-370.

Nylund, Samppa. 2001. "Value-at-Risk Analysis for Heavy-Tailed Financial Returns." MA Thesis. Helsinki University of Technology.

Roberts, S.W. 1959. "Control Chart Tests Based on Geometric Moving Averages." *Technometrics*, 1(3): 239-250.

Sadeghi, Mehdi and Saeed Shavvalpour. 2006. "Energy Risk Management and Value at Risk Modeling." *Energy Policy*, 34: 3367-3373.

Sadorsky, Perry. 1999. "Oil Price Shocks and Stock Market Activity." *Energy Economics*, 21: 449-469.

Van den Goorbergh, Rob and Peter Vlaar. 1999. "Value-at-Risk Analysis of Stock Returns: Historical Simulation, Variance Techniques or Tail Index Estimation?" *De Nederlandsche Bank, DNB Staff Reports*, 40.

Data Sources

Utah Geological Survey. 2009. WTI and Henry Hub spot prices. Available at: http://geology.utah.gov/emp/energydata/statistics/petroleum3.0/pricedata.xls, accessed May 30, 2009.

Yahoo Finance. 2009. General Volatility Index (VIX) quotes. Available at: http://finance.yahoo.com/q?s=^VIX, accessed May 30, 2009.