Adjusted Evaluation Measures for Asymmetrically Important Data

George-Jason Siouris♣, Despoina Skilogianni♣,
Alex Karagrigoriou*♣

♣Lab of Statistics and Data Analysis, Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean

Submitted: October 26, 2018 • Accepted: May 13, 2019

ABSTRACT: In this paper we introduce adjustments for standard evaluation measures appropriate for the analysis of data with asymmetrical importance. In risk analysis, it is understood that the returns of an asset do not all provide the same amount of information. This asymmetry of information is crucial for choosing the most appropriate model and evaluating its forecasting ability. In risk analysis, measures like value at risk (VaR) and expected shortfall (ES) concentrate on the left tail of the distribution of returns so that failures in fitting a model on the right tail are not important. Therefore, when we estimate the VaR of an asset, the days of violations are more important than the days of non-violations. The proposed adjustments take into consideration the asymmetry in importance and are filling the gap in the theory of evaluation of percentiles measures. The measures are divided into *fixed partition*, based on prior information or the goal of forecasting, and *non fixed partition*, based on the time proximity of the model failure. The performance of the proposed measures is illustrated with the use of a stock from the industrial metals and minerals index of the American Stock Exchange (NYSE MKT), as well as a warrant, from the Athens Exchange (ATHEX).

JEL classification: C52, C58, C51

Keywords: PVaR, violation ratios, low price effect, low price correction, back-testing, evaluation measures

[*] Corresponding Author. Email: alex.karagrigoriou@aegean.gr

Introduction

Risk measures have been proposed and used, over the years, to quantify overall risk exposure for the purpose of financial supervision, including internal control and banking supervision. Value at risk (VaR) and expected shortfall (ES) are the most popular of such measures, primarily due to their simplicity. Furthermore, VaR has also been popular due to its easy validation and backtesting. The introduction of ES as a new risk measure that possesses the sub-additivity property and measures the loss in the tail, came naturally as a response to the limitations of VaR, specifically its failure to fulfill the sub-additivity property in non-normal cases, and its inability to capture tail risk (Artzner et al., 1997, 1999; Acerbi and Tasche, 2002).

In order to judge the forecasting quality of typical methodologies such as the above, one may rely on a number of popular evaluation measures, such as the mean square error (MSE), the mean absolute error (MAE), and the mean absolute percent error (MAPE). The problem with these measures is that they fail to evaluate the risk measure estimators such as VaR, because these are percentiles. Indeed, the distance between the realization and the percentile does not provide any information about the accuracy of the percentile estimation and as such it must be adjusted. Even though, many evaluation measures and their adjustments have been proposed over the year (see Wang and Bovik, 2009; Singh et al., 2014; Imbens et al., 2005), no one has introduced any appropriate adjustment for percentile evaluation. Therefore, the problem is an open scientific problem with many applications to a variety of fields. Moreover, our proposed adjustments provide a framework for a far greater family of problems as we will show later. For comparison purposes, backtesting methods, such as the violation ratio (VR) for VaR and normalised shortfall (NS) for ES, are commonly used (see for example Broda and Paolella, 2011).

Recently, two published papers (Siouris and Karagrigoriou, 2017; Siouris et al., 2019), have shown that improvement in forecasting based on the price of a stock is possible. The improvement is made in a subset of the dataset with a specific property (e.g., a very low price). For the same subset of very low-priced stocks, in this paper, adjusted evaluation measures are proposed and implemented in order to evaluate forecasting ability and demonstrate the advantages of the method at hand.

The motivation for the present paper, lies in the fact that in risk analysis, it is often understood that the returns of an asset do not all contain the same amount of information, and do not have the same degree of significance (Sokolova and Lapalme, 2009; Asadabadi et al., 2018; Al-Hawamdeh, 2008; Aladag et al., 2010). This is also apparent in other scientific areas, related to medical, climatological, geophysical or meteorological phenomena. The ”asymmetry in the importance” of information is crucial both in choosing the most appropriate model and in evaluating its forecasting ability. In the case of risk analysis, risk measures like

VaR and ES, concentrate on the left tail of the distribution of returns. Hence, failures in fitting a model to the right tail are not considered important. When the VaR of an asset at time t is evaluated, the days of violations are more important than the days of non-violations; the same happens with ES. Similar behavior is also found in other situations, such as health sciences and geosciences. Indeed, in biosurveillance systems for instance, the same phenomenon occurs for epidemic and non epidemic periods, the former being the important ones although it is the latter that are frequently modelled. Based on the above, evaluation measures should take into account this ”asymmetry in the importance” of information borne in the data.

The concept of ”asymmetry in the importance” of information is not rare in science. For example, great earthquakes provide much more information for the estimation of seismic magnitude than smaller ones. Days of extreme losses on the financial markets bear greater information compared to the ”normal” days, as correlations are non-linear within each financial market in the day-by-day returns as well as between financial markets. Epidemic days carry disproportionate information compared to non-epidemic days. This is true in general, whenever the cost function for the difference between the estimation and the realized value is not constant. This non-constant cost function depends on the distance between the two values as in the case of a risk measure when a violation occurs, on the value of the realization and the range that it belongs to as in the case of seismology and epidemiology, or on a pre-specified partition of the dataset.

In this paper, we introduce for the first time, to the best of our knowledge, adjustments of standard evaluation measures appropriate for asymmetrically important data. The proposed adjustments fill a gap in the literature of percentile evaluation both for situations where the cost of error is not equal for all cases and for situations where a clear partition of the dataset exists. The proposed measures will be divided into two general categories based on the method of partitioning the dataset; fixed partition, based on prior information or the goal of the forecasting, and non fixed partition, based on the time proximity of the model failure. The two categories of measures are presented in Section 1 and an evaluation of their performance is given in Section 2. Concluding remarks are provided in Section 3.

1 Methodology

Let Ω be the available dataset, and A_i be a partition of Ω based on our criteria. Any family of non-empty sets is a partition of Ω, if and only if the following conditions hold.

1. $\Omega = \bigcup_{i \in I} A_i$ where A_i are non-empty subsets of Ω, $\forall i \in I \in \mathcal{N}$

2. $A_i \cap A_j = \emptyset$ for $i \neq j$.

The partitioning of the dataset will be done in two main ways, which we will call Type 1 and Type 2. Type 1 is a fixed partition, based on prior information for the dataset or the goal of the forecasting procedure. Type 2 is a dynamically adjusted partition of unit sets, based on their time proximity to the present. Both types, with respective examples, are presented below.

For the purpose of this work, we concentrate on a risk measure for estimating possible losses, and below we introduce the Type 1 adjusted evaluation measure.

The percentage VaR (PVaR), which is a risk measure for estimating the possible percentage losses from trading assets, within a set time period, is defined as follows:

Definition 1: (a) PVaR(p) is the 100pth percentile of the distribution of returns.

(b) PVaR$_t(p)$ is the PVaR(p) at time t, where the distribution of standardized returns with standard deviation σ, (R_t/σ), is denoted by $F(\cdot)$. Observe that the PVaR of an asset takes the form:

$$PVaR_t(p) = -\sigma F^{-1}(p).$$

where $F^{-1}(p)$ is the $100p^{th}$ percentile of the inverse of the distribution of (R_t/σ), namely the standardized returns with standard deviation σ.

(c) Let p_t be the asset value at time t and $c(p_t)$ be the minimum price variation (market accuracy) associated with the value of the asset at time t. Then, the minimum possible return of an asset at time t (mpr_t) is the logarithmic return that the asset will produce if its value changes by $c(p_t)$ and is given by:

$$mpr_t = log\left(\frac{p_t + c(p_t)}{p_t}\right)$$

(d) The low price effect area is the range of prices for which the mpr is greater than a pre-specified threshold Θ.

As a simple example consider the case where a stock market operates with a fixed accuracy $c(p_t) = 0.001$. If the threshold is chosen to be $\Theta = 0.001$, then according to Definition 1(c), it is easy to see that for a stock price p_t less than 0.999001 the minimum possible return mpr_t is greater than Θ. Thus, the low price effect area is the set $\{p_t : p_t < 0.999001\}$.

Siouris et al. (2019) proposed the low price correction by rounding the PVaR$_t(p)$ estimate to a legitimate value, namely the next integer multiple of the mpr_t. In particular, the low price correction of the estimation denoted by $\widetilde{PVaR_t}(p)$ is given by:

$$\widetilde{PVaR_t}(p) = \begin{cases} \left(\left\lfloor \frac{PVaR_t(p)}{mpr_t} \right\rfloor + 1 \right) \cdot mpr_t, & \text{if} \quad mpr_t \geq \Theta \\ PVaR_t(p), & \text{if} \quad mpr_t < \Theta \end{cases} \tag{1}$$

where $\lfloor w \rfloor$ is the floor function (integer part) of w.

Observe that under the above low price correction, the market accuracy is passed on to the evaluation of the PVaR, resulting in reasonable estimations and as a result fewer violations. Let $\{y_t\}_{t=1}^{T}$ be a sample of daily logarithmic losses on a trading portfolio, and define the indicator η_t that takes the value 1 if $y_t > PVaR_t(p)$ (or $\widetilde{PVaR_t}(p)$ if the above correction has been enforced) and 0 otherwise. A PVaR violation is said to have occurred if $\eta_t = 1$. Observe that the low price correction is associated with the rationalization of the estimated asset returns as it is the next integer multiplier of the minimum possible return.

In order to measure the accuracy of the above procedure, one cannot rely on popular evaluation measures, such as the MSE, MAE, MAPE, or the heteroskedasticity mean square error (HMSE), because VaR estimations are percentile estimations. The fact that the proximity of the realzsed returns to the estimated percentile does not provide any information creates the need for introducing appropriately adjusted accuracy measures. Note though that this is not the case for backtesting procedures; by concentrating on the underlining risk, only the proximity in days of violation provides information on the quality and accuracy of the proposed methodology. The decrease in the values of the adjusted evaluation measures will show and verify the improved forecasting quality of the low price correction. It is easily seen, that the MSE in the above setting, is defined as follows:

$$MSE = \frac{1}{\sum_{t \leq T} \eta_t} \sum_{t \leq T \& \eta_t = 1} (y_t - PVaR_t(p))^2. \tag{2}$$

Observe that the above is a special weighted MSE expression and can be written as

$$MSE = \sum_{t \leq T} \frac{\eta_t}{v(T)} \cdot (y_t - PVaR_t(p))^2$$

where

$$v(T) = \sum_{t \leq T} \eta_t.$$

All popular evaluation measures used in the literature can be adjusted to fit our proposed methodology by generalizing the special weights used above. The adjusted MSE is defined below.

1.1 Type 1 Evaluation Measures

Definition 2: Adjusted MSE (AMSE) is given by

$$AMSE = \frac{1}{w}\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2,$$

where w_i is a pre-specified weight given to the i-th observation based on its importance, y_i is a sample observation, $\hat{y}_i$ is the observation forecast, n is the sample size, and $w = \sum\limits_{i=1}^{n} w_i$.

We also provide the respective formulas for the adjusted mean absolute error (MAE), the adjusted mean absolute error (AMAE), the adjusted mean absolute percent error (AMAPE), and the adjusted heteroskedasticity mean square error (AHMSE):

$$\begin{aligned}
AMAE &= \frac{1}{w}\sum_{i=1}^{n} w_i|y_i - \hat{y}_i|,\\
AMAPE &= \frac{1}{w}\sum_{i=1}^{n} w_i\left|\frac{100(y_i - \hat{y}_i)}{y_i}\right|,\\
AHMSE &= \frac{1}{w}\sum_{i=1}^{n} w_i\left[\frac{y_i}{\hat{y}_i} - 1\right]^2.
\end{aligned}$$

The adjusted measures presented in this work, provide a more general and flexible framework, for addressing ”asymmetry in importance data”.

As shown in (3), examples of the applicability of these measures can be found in risk analysis; other examples are found in biosurveillance, among other settings. In the case of risk analysis, and specifically in the case of VaR estimations, days in which a violation occurred, are obviously more important than others. The same logic holds for ES estimations. Also, we may want to give different weights to days with positive returns than to days with negative returns, because days with positive returns are not so important from the risk analyst’s point of view. In this case, the partition can be done as follows:

$$\begin{aligned}
A_1 &= \{i \in I | y_i \leq 0\},\\
A_2 &= \{i \in I | 0 < y_i \leq PVaR_i(p)\},\\
A_3 &= \{i \in I | y_i > PVaR_i(p)\},
\end{aligned}$$

where y_i are the daily logarithmic losses of a stock, and $PVaR_i(p)$ is the Percentage Value at Risk, as defined in Siouris et al. (2019).

Under this partition, weights can be given as follows

$$w_i = \begin{cases} g_1, & i \in A_1 \\ g_2, & i \in A_2 \\ g_3, & i \in A_3. \end{cases} \tag{3}$$

As it is clear these weights are chosen arbitrarily, which is the main weakness of Type 1 adjustment of evaluation measures. One possible choice for the weights is $g_1 = 0$, $g_2 = 1$ and $g_3 = 10$, assuming that days of PVaR violation are 10 times more important than days of positive losses that did not violate PVaR and days of negative losses (days of positive returns) are not taken into account.

Analogously, in biosurveillance we have epidemic and non epidemic periods, and the failures of the model should not be evaluated as the same in these two periods.

Remark 1: When equal weights are assigned to all observations, we obtain the standard evaluation measures. As a result, a standard evaluation measure constitutes a special case of the adjusted measures defined above.

Remark 2: Additional functions like the one proposed by Gónzalez-Rivera et al. (2004) could also be used for exploring the robust capabilities of the proposed methodology. Such explorations will be the main theme of a future study.

1.2 Type 2 Evaluation Measures

In the second type of partition, the weights applied are time dependent and exponentially decreasing as we go further back in time. In this case, models that are generally good but fail in the latter part of the time series, should be replaced by more appropriate ones. Weights can be assigned in the same fashion as in the exponentially weighted moving average (EWMA) model, a well-known and widely used model in risk analysis and financial time series:

$$w_{t+1} = \lambda w_t = \lambda^2 w_{t-1} = \ldots = \lambda^{n+1} w_{t-n},$$

where λ is the EWMA constant. Under the assumption that the weights sum to one, namely

$$\sum_{t=1}^{\infty} w_t = w_1 \sum_{t=1}^{\infty} \lambda^t = 1$$

and for $|\lambda| < 1$, we have that $w_1 = 1 - \lambda$. For a finite dataset, which is usually the case, $w_1 = \frac{1-\lambda}{\lambda-\lambda^{n+1}}$, where n is the sample size. In contrast to Type 1 adjustment, the weights in the Type 2 case, are not chosen arbitrarily.

2 Applications

To apply the proposed methodology, in this section we use, for illustrative purposes, the Northern Dynasty Minerals, Ltd. (NAK) stock from the industrial metals and minerals index of the Americal stock exchange, NYSE MKT, as well as, the National Bank of Greece (NBG) warrant from the Athens stock exchange, ATHEX. These assets, the basic statistical characteristics of which are furnished in Table 1, were chosen for their low prices which will help fully explore the capabilities of the proposed methodology.

2.1 The NBG warrant

Warrants are in many ways similar to options, but with a few key differences distinguishing them. Warrants are generally issued by the company itself, not by a third party and they are traded over-the-counter more often than on an exchange market. Unlike options, warrants cannot be written by investors and they tend to have much longer periods between issue and expiration. They do not pay dividends or come with voting rights. Investors are attracted to warrants as a means of leveraging their positions in a security, hedging against downside risk, or exploiting arbitrage opportunities. The warrants of the National Bank of Greece (NBG) were issued in 2013 as part of the first recapitalization of the Greek banking system, and they were traded on the Athens Exchange (ATHEX). The expiration date of the warrants was 4.5 years from the issue date. They had 9 exercise dates, one every six months. On each exercise date, the owner can buy stocks at a specific price, which in the case of the NBG was 4.6761 euros per stock. Later, when the NBG stock price collapsed, the warrants' price also collapsed.

Below, we first implement the low price correction (lpc) on the PVaR methodology given in (1). Then, the backtesting results for various models considered (EWMA, Normal, historic simulations [HS] and generalized autoregressive conditional heteroskedasticity [GARCH] model with skewed t-student innovations and estimated degrees of freedom) are considered; these appear in Table 2. Note that the primary tool for backtesting is the violation ratio

(VR) which is obtained by comparing observed frequencies and expected number of violations. The predictive ability of all models used is presented in Tables 3-7: the standard evaluation measures (Table 3); the special case of adjusted measures given in (2) (Table 4); Type 1 evaluation measures only in the final days of violations (the days of violations of the corrected model) (Table 5); Type 1 measures given in (3) with $g_1 = 0$, $g_2 = 1$, $g_3 = 10$ (Table 6); and Type 2 evaluation measures, with a typical choice for the parameter λ equal to 0.94 (Table 7).

We must point out the differences in the results from the various evaluation methods in the following tables. Table 2 shows that in all 4 models, the low price correction methodology produces estimations with fewer violations, a result that by itself proves the usefulness of this correction. In contrast, in Table 3 the standard evaluation measures completely fail to demonstrate any improvement of estimations. On the contrary, most of the times the estimations were far worse than previously. This is due to the aforementioned inability of standard evaluation measures to evaluate percentiles. Due to this inconsistency between Tables 2 and 3, we are motivated to provides appropriate adjustments of the standard evaluation measures.

Table 2 gives the VRs for the PVaR estimations for four different models, with and without low price correction implemented, as well as, the PVaR volatilities. Based on the results of the PVaR estimations shown in Table 2 and the fact that an acceptable range of VR (according to the Basel III accords) is 0.8-1.2, the HS and GARCH(1,1) fail the most among all the models considered. The estimations of all the models systematically underestimate the underlying risk. We observe that the proposed low price correction methodology significantly improves the VR in all models. In fact, Table 2 clearly shows that in all 4 models, the low price correction methodology produces PVaR estimations with fewer violations. As a result, in all cases examined, VaR estimations have been improved in the sense that their VR is considerably reduced with the low price correction. Indeed, the VR (2nd vs. 1st column in Table 2) is always improved, in the sense that the frequency of violations is reduced which is associated with a more defensive approach in accordance with Basel III. This is a result that by itself clearly proves the usefulness of the low price methodology methodology. The same conclusions are also evident from Figures 1 and 2. These figures show the improvement arising from the use of the proposed methodology and are provided for a visual understanding of the contribution of this work.

The first adjustment is presented in Table 4, where only the violations are taken into consideration. Even though this table is closer to the truth, it still fails to capture the improvement shown by the VR. This is due to the different denominator in the adjusted evaluation measures for the estimations with and without the low price correction. By far the best adjusted evaluation measure, one, that completely captures the genuine improvement in our estimations with the help of the low price correction, is that in Table 5. In this table, the improvement is genuine for all models and for all measures, as was expected from the definition of the low price correction. The adjusted evaluation measure results shown in Table 5 completely agree with the VR results of Table 2 and also quantify something that was clear from the definition of low price correction, specifically, that the improvement from the risk analysis point of view is genuine. Tables 6 and 7 present the results from other variations of adjusted evaluation models and have to do with the needs of the researchers. Last but not least, Table 5 in almost all cases selects a different model as the best, compared to Table 2.

As we see in Tables 3-7, the chosen model depends on the evaluation measure used and the adjustment chosen. For example, HS is the best model according to the standard MSE evaluation measure, whereas EWMA appears to be superior based on the Type 1 adjustments presented in Tables 4 and 6. GARCH(1,1) is the model selected according to the Type 2 adjusted evaluation measure of Table 7. The most important result, is that, although the standard evaluation measures fail to capture the obvious improvement of the low price correction, the Type 1 adjustment completely succeeds in illustrating this improvement. This is clear from the results for all measures in Table 5 which contains the evaluation measures results for PVaR with and without low price correction in the final days of violations (the days of the violations of the corrected model). Finally, the decrease of HMSE for the lpc GARCH(1,1), in all adjusted cases, is more extreme than before. Note that, various measures based on both quadratic and absolute errors were used for completeness. The investigator is free to choose the most appropriate measure for his/hers purposes. However, caution is needed when AMAE and AMAPE measures are used because both are scale sensitive.

Remark 3: It should be mentioned that any distribution can be used, as in Siouris and Karagrigoriou (2017) where Gaussian and non-Gaussian innovations were examined. In the present work, the best among a series of candidate models was chosen to be presented, namely the one with skewed t-Student innovations. Furthermore, note that alternative models that have recently attracted considerable attention such as the generalized autoregressive score (GAS) models (Creal et al., 2013; Harvey, 2013), will be left for future work.

Remark 4: Note that in addition to the VR, which is equivalent to the well-known Kupiec's test (Kupiec, 1995) other backtesting procedures (see e.g., Christoffersen, 1998) could also be applied.

2.2 The NAK stock

The NAK data (2500 observations) extend from January 8, 2008 until December 12, 2017. Details about the behavior of the NAK stock can be found in Siouris et al. (2019). The low price correction methodology is again implemented. Backtesting is done with the same models as before with the addition of the asymmetric power autoregressive conditional heteroskedasticity (APARCH) model with skewed t-student innovations model and degrees of freedom estimated. The results of the backtesting are presented in Table 8 and Figures 1 and 2. Furthermore, the performance of the low price correction methodology according to the accuracy measures, is presented in Tables 9-13: the standard evaluation measures (Table 9); the special case of adjusted measures given in (2) (Table 10); Type 1 evaluation measures only in the final days of violations (the days of violations of the corrected model) (Table 11); Type 1 measures given in (3) with $g_1 = 0$, $g_2 = 1$, $g_3 = 10$ (Table 12); and Type 2 evaluation measures, with a typical choice for the parameter λ equal to 0.94 (Table 13).

Tables 8 and 9 present a comparison of the results of the evaluation methods identical to the ones presented in Tables 4 and 5 and discussed above. These results confirm the necessity of adjusted evaluation measures and provide the motivation for their definition and their implementation.

Table 8, gives the VRs for the PVaR estimations for six different models, with and without low price correction implemented, as well as, the PVaR volatilities. Based on the results shown in Table 8, and the fact that an acceptable range of VR is 0.8-1.2, we have that the PVaR estimation under normality, GARCH(1,1) and GARCH(2,2), are the three models that fail the most. The estimation under normality systematically overestimates the underlying risk, as it misinterprets big positive returns as risk factors, whereas GARCH(1,1) and GARCH(2,2) systematically underestimate it, and EWMA seems to fail. Acceptable models are the ones based on historical simulation and APARCH(1,1). It can be observed that the proposed methodology of the low price correction improves or leaves unchanged the VR in all models.

The first adjustment is presented in Table 10, where only the violations are taken into consideration. Even though this table is closer to the truth, it still fails to capture the improvement shown by the violation ratio. This is due to the different denominator in the adjusted evaluation measures for the estimations with and without the low price corrction. By far the best adjusted evaluation measure, one that completely captures the genuine improvement in our estimations with the help of low price correction, is that in Table 11. In this table the improvement is genuine for all models and all measures, as was expected from the definition of low price correction. The adjusted evaluation measure results shown in Table 11 agree completely with the VR results of Table 8 and also quantify something that was clear from the definition of low price correction, specifically, that the improvement from the risk analysis point of view is genuine. Tables 12 and 13 present results of other variations of the adjusted evaluation models and have to do with the needs of the researchers. Last, but not least, Table 11 in almost all cases selects a different best model as compared to Table 8.

Based on MSE, we observe that for the standard case (Table 9), GARCH(1,1) is the best model for the PVaR estimation. The same is also the case according to Type 1 (Table 12) and Type 2 adjusted measures (Table 13). On the other hand, this is not the case for the results shown in Table 10 which are based on the special Type 1 measure given in (2)) and clearly pick EWMA as the best model. However, such a conclusion is expected if one takes into consideration the fact that EWMA is the most ”nervous” among the competing models, and, additionally, that the focus of the study is solely on violations. It must be noted, that even though GARCH(1,1) was not acceptable based on the VR, it is the one that behaves the best, according to the results in Tables 9, 11 and 13, and is among the best performing models based on the results in Table 10.

Comparing the results of Tables 9 and 13, which are both based on the entire set of observations, we clearly verify the effectiveness and significance of the EWMA weights (i.e., of type 2 measures), as compared to the standard evaluation measures. On the other hand, whereas the standard evaluation measures fail to reveal the failure level of the Normal model, the adjusted evaluation measures do not. Also, the low price correction is evaluated more accurately with the techniques presented in Tables 10-12, which concentrate on the violations and the negative returns, respectively. This is paticularly the case for the Type 1 adjustment, as shown in Table 11 which contains the evaluation measures results for PVaR with and without low price correction in the final days of violations (the days of the violations of the corrected model). Lastly, it must be noted that HMSE is not an appropriate measure for testing normality, and should not be taken into consideration.

3 Conclusions

In this paper, we have presented adjusted evaluation measures, applicable to situations in which not all observations are equally important. The phenomenon is quite common in finance as, for instance, in the case of asset returns which do not all contain the same amount of information. The proposed adjusted evaluation measures are categorized in two general classes, according to the method of partition of the data. Type 1 is a fixed partition, based on prior information or the goal of the forecasting procedure whereas Type 2 is a dynamically adjusted partition of unit sets, based on the model failures, and their time proximity to the present.

For equal weights, Type 1 adjusted evaluation measures are simplified to standard measures. In other words, a standard evaluation measure constitutes a special case of the class of Type 1 adjusted measures. As shown in the application section, adjusted evaluation measures of Type 1, are quite useful, offering additional information about the forecasting ability of models. On the other hand, their main weakness is associated with the fact that weights are arbitrarily chosen. However, if the main goal is to narrow the evaluation to a subset of the available values, as is needed for the correct evaluation of the low price correction, then the choice of weights is straight forward and the results quite clear as it is shown on Tables 4 and 10.

In contrast, the weights in the Type 2 case, are given according to the EWMA models' weights. Such weights typically decrease at an exponential rate as time goes back, which resembles the autocorrelation behavior of assets returns. These time dependent weights provide a framework within which the most appropriate model in the latter part of the time series is chosen. Note that the applicability of the proposed measures is not limited to finance; it can easily be extended to other scientific areas where data do not necessarily carry the same amount of information, such as in physical (earthquakes, floods etc.) or climatological (heat or cold waves) phenomena.

Acknowledgements: This work was completed as part of the research activities of the *Laboratory of Statistics and Data Analysis* of the University of the Aegean. The authors wish to express their appreciation to the Editor and an anonymous reviewer whose comments and suggestions have greatly improved the quality as well as the presentation of the manuscript.

References

Acerbi, C. and Tasche, D. (2002). Expected shortfall: a natural coherent alternative to value at risk. *Economic Notes by Banca Monte dei Paschi di Siena SpA*, 31(2):379–388.

Al-Hawamdeh, S. (2008). *Knowledge Management: Competencies and Professionalism*, volume 7. World Scientific, Singapore.

Aladag, C. H., Egrioglu, E., Gunay, S., and Basaran, M. A. (2010). Improving weighted information criterion by using optimization. *Journal of computational and applied mathematics*, 233(10):2683–2687.

Artzner, P., Heath, D., Delbaen, F., and Eber, J.-M. (1997). Thinking coherentlyg. *Risk*, 10:68–71.

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

Asadabadi, M. R., Saberi, M., and Chang, E. (2018). Targets of Unequal Importance Using the Concept of Stratification in a Big Data Environment. *International Journal of Fuzzy Systems*, pages 1–12.

Broda, S. A. and Paolella, M. S. (2011). Expected shortfall for distributions in finance. In *Statistical tools for finance and insurance*, pages 57–99. Springer, London.

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

Creal, D., Koopman, S. J., and André, L. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28:777–795.

Das, K., Jiang, J., and Rao, J. (2004). Mean squared error of empirical predictor. *Annals of Statistics*, 32(2):818–840.

Gónzalez-Rivera, G., Lee, T.-H., and Mishra, S. (2004). Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *International Journal of Forecasting*, 20:629–645.

Harvey, A. (2013). Generalized autoregressive score models with applications. *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*, 52.

Imbens, G. W., Newey, W. K., and Ridder, G. (2005). Mean-square-error calculations for average treatment effects. https://dx.doi.org/10.2139/ssrn.954748.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3:73–84.

Singh, S., Singh, D. S., and Kumar, S. (2014). Modified Mean Square Error Algorithm with Reduced Cost of Training and Simulation Time for Character Recognition in Backpropagation Neural Network. In *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013*, pages 137–145. Springer, London.

Siouris, G.-J. and Karagrigoriou, A. (2017). A low price correction for improved volatility estimation and forecasting. *Risks*, 5(3):45.

Siouris, G.-J., Skilogianni, D., and Karagrigoriou, A. (2019). Post model correction in value at risk and expected shortfall. *International Journal of Mathematics, Engineering and Management Sciences*, 4(3):542–566.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Wang, Z. and Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117.

Figure 1: Backtesting PVaR without low price correction (NBG warrants)

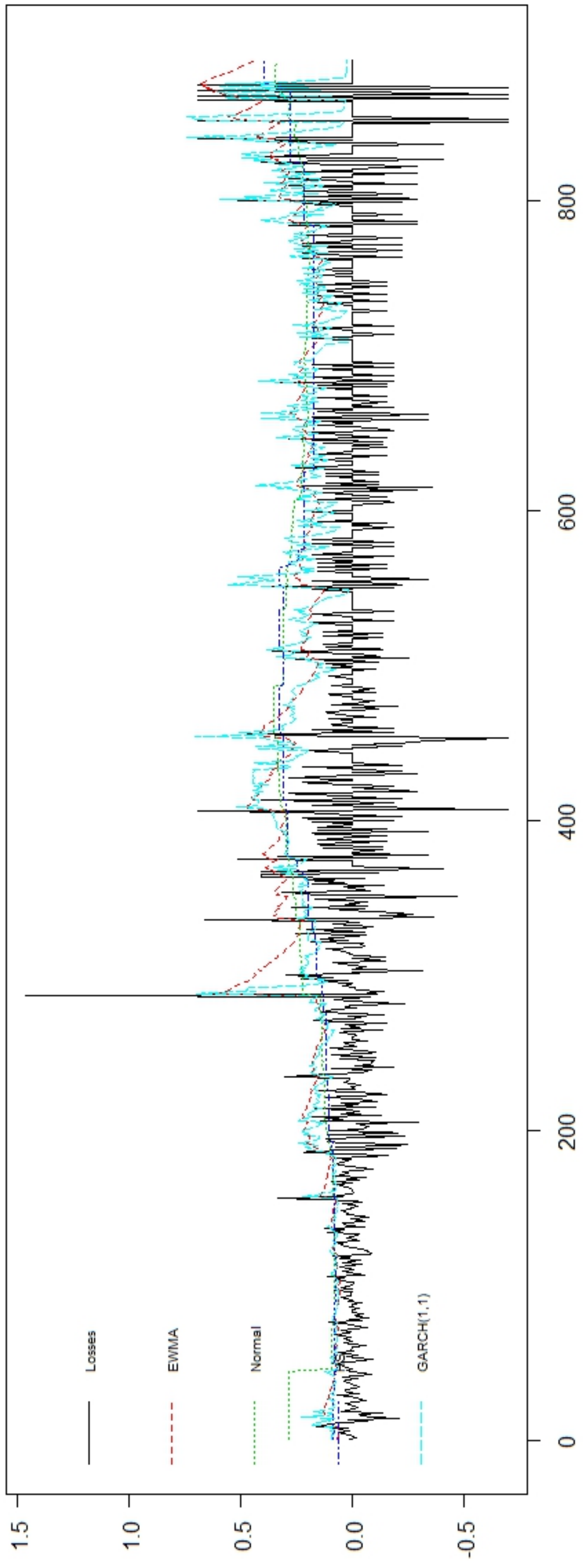

Figure 2: Backtesting PVaR with low price correction (NBG warrants)

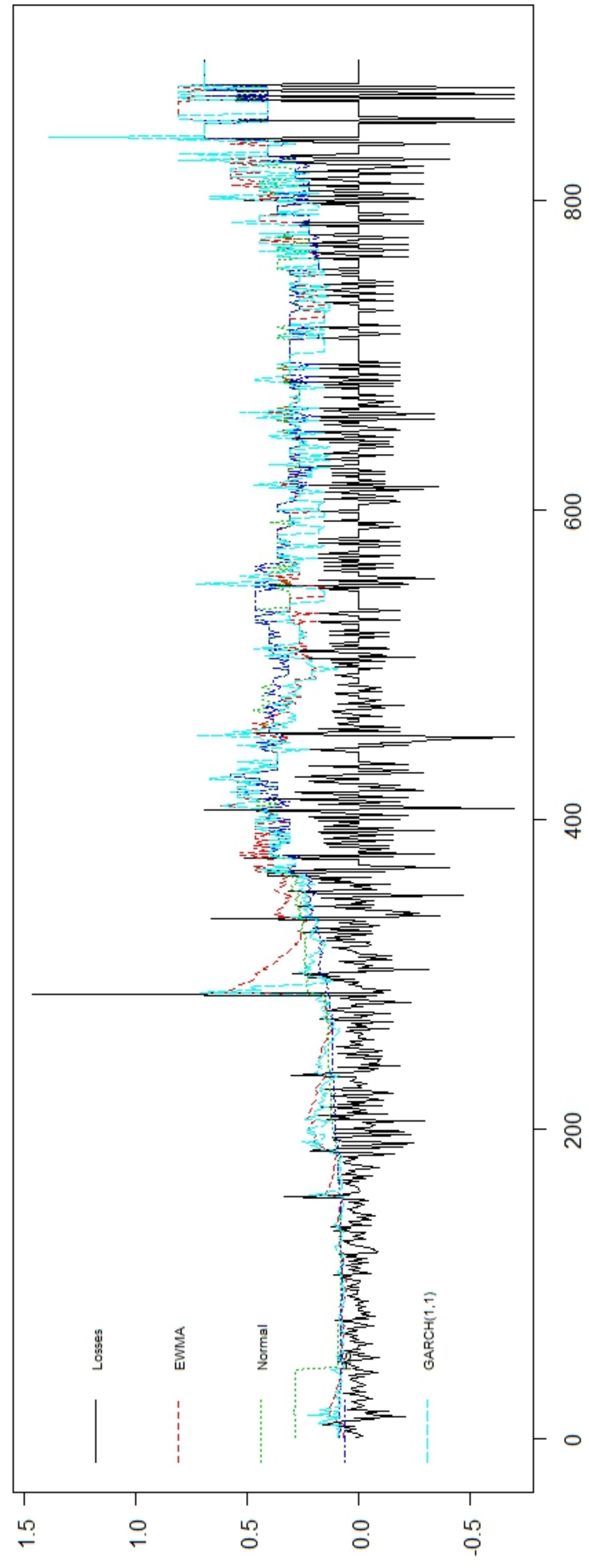

Figure 3: Backtesting PVaR without low price correction (NAK stock)

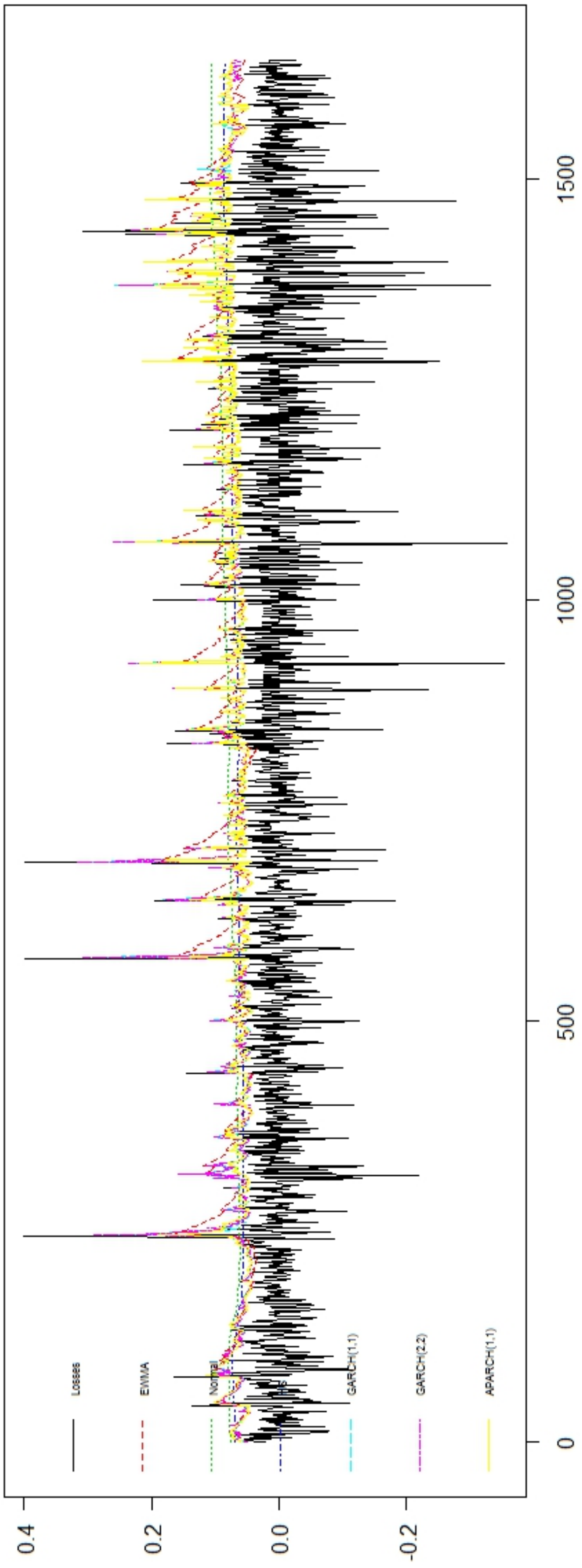

Figure 4: Backtesting PVaR with low price correction (NAK stock)

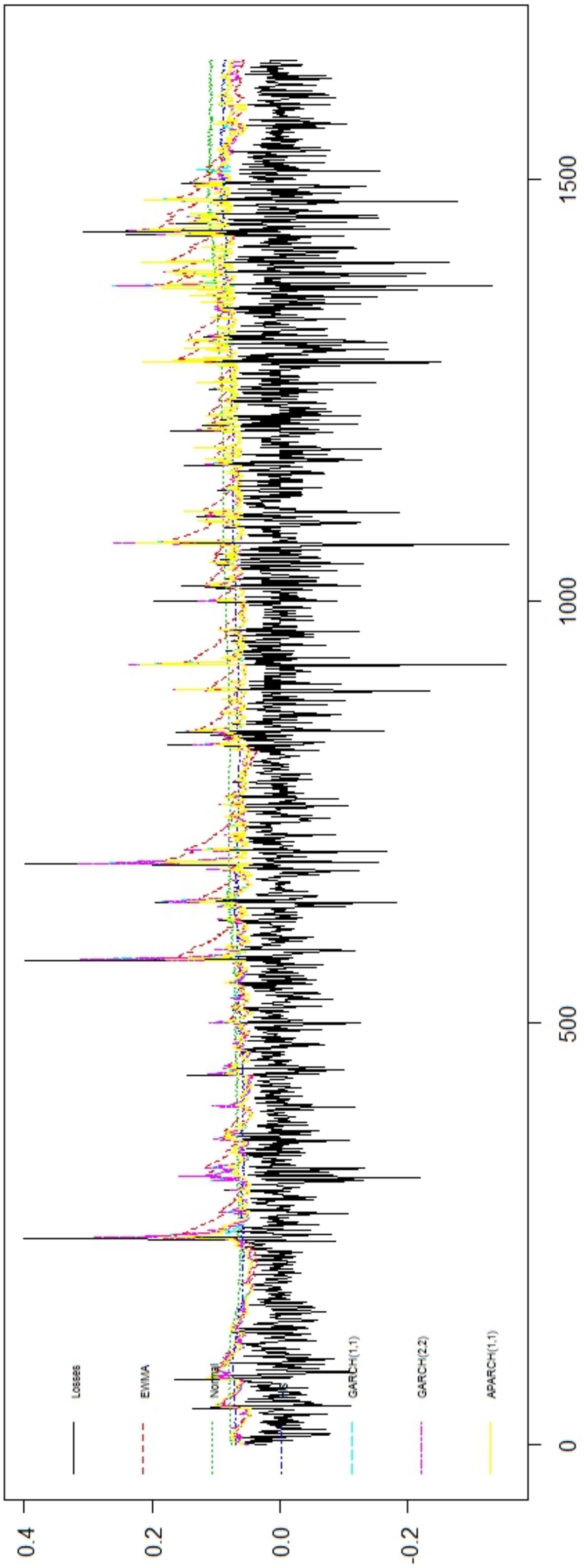

Table 1: Descriptive Statistics for Prices, First Differences of Prices and Logarithmic Returns

	NBG Prices	NBG 1st Dif.	NBG Log(Returns)	NAK Prices	NAK 1st Dif.	NAK Log(Returns)
MEAN	0.3269217	-0.00125592	-0.006584553	4.419676	-0.004597839	-0.0007760009
MEDIAN	0.031	0	0	2.69	0	0
VAR	0.2107202	0.002277542	0.02655832	17.0206	0.0561949	0.002804873
SD	0.4590427	0.0477236	0.1629672	4.125603	0.2370546	0.05296105
MIN	0.001	-0.759	-1.608385	0.21	-2.96	-0.4002432
MAX	1.84	0.82	1.670682	21.08	2.22	0.3566749
SKEWNESS	1.25837	1.214182	-0.5977321	1.083448	-0.3358244	0.1930199
KURTOSIS	3.264962	143.9865	30.42116	3.835573	25.34604	11.83174

Note: Descriptives of Prices, First Differences of Prices and Logarithmic Returns for NBG warrant and NAK.

Table 2: Backtesting With and Without Low Price Correction (lpc) for PVaR - NBG Warrants

	VR without lpc	VR with lpc	PVaR vol without lpc	PVaR vol with lpc
EWMA	1.414141	0.8754209	0.1224841	0.1701248
Normal	1.436588	0.9876543	0.08374533	0.1390613
HS	1.885522	1.234568	0.08965133	0.1495181
GARCH(1,1)	1.818182	1.257015	0.1204021	0.1731923

Note: Violation Ratio (VR) and % Value at Risk (PVaR) volatility are provided with and without lpc for the National Bank of Greece (NBG) warrants for the models EWMA, Normal, Historic Simulations (HS) and GARCH with skewed t-Student errors.

Table 3: Standard Evaluation Measures Results for PVaR With and Without Low Price Correction (lpc) of NBG Warrants

	MSE	MAE	MAPE	HMSE	Number of Violations
EWMA	0.08959510	0.2345721	283.0198	1.414714	891
EWMA (lpc)	0.13321386	0.2882063	296.3516	1.312602	891
Normal	0.07842356	0.2313674	335.4894	1.499397	891
Normal (lpc)	0.11700468	0.2815343	348.1618	1.366298	891
HS	0.06918058	0.2083804	233.1994	1.627136	891
HS (lpc)	0.10728055	0.2591924	245.8914	1.476536	891
GARCH(1,1)	0.07873061	0.2113150	256.1190	2.746523	891
GARCH(1,1) (lpc)	0.12377995	0.2687740	268.0334	1.373017	891

Note: MSE, MAE, MAPE and HMSE *standard* measures are provided with and without lpc for the National Bank of Greece (NBG) warrants for the models EWMA, Normal, Historic Simulations (HS) and GARCH with skewed t-Student errors.

Table 4: Adjusted Type 1 Evaluation Measures Results for PVaR With and Without Low Price Correction (lpc) of NBG Warrants

	MSE	MAE	MAPE	HMSE	Number of Violations
EWMA	0.04454162	0.10670786	27.19767	1.679775	63
EWMA (lpc)	0.05632466	0.09550818	24.05620	2.408766	39
Normal	0.05569806	0.13136742	29.03018	1.924203	64
Normal (lpc)	0.05046753	0.08729689	19.76906	2.075661	44
HS	0.04582400	0.11644768	29.33722	1.929689	84
HS (lpc)	0.04486931	0.09170765	25.06878	2.299259	55
GARCH(1,1)	0.05232076	0.12790622	35.75844	9.692231	81
GARCH(1,1) (lpc)	0.04195689	0.08042647	21.16355	1.881913	56

Note: MSE, MAE, MAPE and HMSE *adjusted* measures given in (2) are provided with and without lpc for the National Bank of Greece (NBG) warrants for the models EWMA, Normal, Historic Simulations (HS) and GARCH with skewed t-Student errors.

Table 5: Adjusted Type 1 Evaluation Measures Results for PVaR With and Without Low Price Correction (lpc) of NBG Warrants

	MSE	MAE	MAPE	HMSE	Number of Violations
EWMA	0.06993500	0.14457923	33.78901	2.658553	39
EWMA (lpc)	0.05632466	0.09550818	24.05620	2.408766	39
Normal	0.07788985	0.16334422	34.04885	2.742979	44
Normal (lpc)	0.05046753	0.08729689	19.76906	2.075661	44
HS	0.06790716	0.15499175	37.27808	2.908765	55
HS (lpc)	0.04486931	0.09170765	25.06878	2.299259	55
GARCH(1,1)	0.07232273	0.15283203	37.89425	10.886520	56
GARCH(1,1) (lpc)	0.04195689	0.08042647	21.16355	1.881913	56

Note: MSE, MAE, MAPE and HMSE *Type 1* measures for the final days of violations are provided with and without lpc for the National Bank of Greece (NBG) warrants for the models EWMA, Normal, Historic Simulations (HS) and GARCH with skewed t-Student errors.

Table 6: Adjusted Type 1 Evaluation Measures Results for PVaR With and Without Low Price Correction (lpc) of NBG Warrants

	MSE	MAE	MAPE	HMSE	Number of Violations	
	MSE	MAE	MAPE	HMSE	Number of Violations	M(*)
EWMA	0.03656765	0.10480884	133.24073	1.302309	63	254
EWMA (lpc)	0.04329688	0.10755786	174.29886	1.564897	39	278
Normal	0.04488074	0.12469699	172.85350	1.488760	64	253
Normal (lpc)	0.04054776	0.10345640	211.37578	1.434692	44	273
HS	0.03782291	0.10781418	97.02337	1.586415	84	233
HS (lpc)	0.03547745	0.09325940	122.11400	1.671224	55	262
GARCH(1,1)	0.04341466	0.11944931	110.73903	7.589229	81	236
GARCH(1,1) (lpc)	0.03491252	0.08956792	127.48493	1.403276	56	261

Note: MSE, MAE, MAPE and HMSE *Type 1* measures given in (3) are provided with and without lpc for the National Bank of Greece (NBG) warrants for the models EWMA, Normal, Historic Simulations (HS) and GARCH with skewed t-Student errors.

Table 7: Adjusted Type 2 Evaluation Measures Results for PVaR With and Without Low Price Correction (lpc) of NBG Warrants

	MSE	MAE	MAPE	HMSE	Number of Violations
EWMA	0.3535559	0.5649415	14.27348	1.295894	891
EWMA (lpc)	0.5957803	0.7379845	15.31295	1.163027	891
Normal	0.1762310	0.3972225	14.83337	1.755109	891
Normal (lpc)	0.4567260	0.6492885	11.36114	1.397210	891
HS	0.2024337	0.4331419	14.92063	1.793435	891
HS (lpc)	0.4558232	0.6481503	11.20965	1.400882	891
GARCH(1,1)	0.1512406	0.2297988	17.07272	10.058642	891
GARCH(1,1) (lpc)	0.5364373	0.6970300	13.72804	1.253154	891

Note: MSE, MAE, MAPE and HMSE *Type 2* measures are provided with and without lpc for the National Bank of Greece (NBG) warrants for the models EWMA, Normal, Historic Simulations (HS) and GARCH with skewed t-Student errors.

Table 8: Backtesting With and Without Low Price Correction (lpc) for PVaR - NAK

	VR without lpc	VR with lpc	PVaR vol without lpc	PVaR vol with lpc
EWMA	0.8038977	0.7917174	0.03178796	0.03203627
Normal	0.7551766	0.7308161	0.01371889	0.01399949
HS	1.242387	1.205847	0.009581244	0.009917231
GARCH(1,1)	1.144945	1.084044	0.02573922	0.02595559
GARCH(2,2)	1.096224	1.010962	0.02578427	0.02595401
APARCH(1,1)	1.181486	1.010962	0.02228844	0.02245804

Note: Violation Ratio (VR) and % Value at Risk (PVaR) volatility are provided with and without lpc for Northern Dynasty Minerals, Ltd. (NAK) for the models EWMA, Normal, Historic Simulations (HS), GARCH with skewed t-Student errors and APARCH with skewed t-Student errors.

Table 9: Standard Evaluation Measures Results for PVaR With and Without Low Price Correction (lpc) of NAK

	MSE	MAE	MAPE	HMSE	Number of Violations
EWMA	0.011649290	0.08958197	377.9544	1.437325	1642
EWMA (lpc)	0.011871865	0.09067956	384.4612	1.422141	1642
Normal	0.009965702	0.08464938	371.0182	1.437617	1642
Normal (lpc)	0.010171222	0.08578432	378.1745	1.422154	1642
HS	0.007967333	0.07329502	318.6766	1.611337	1642
HS (lpc)	0.008140035	0.07439704	325.7300	1.589516	1642
GARCH(1,1)	0.008671210	0.07447498	318.9933	1.619391	1642
GARCH(1,1) (lpc)	0.008847690	0.07557095	325.9550	1.595655	1642
GARCH(2,2)	0.008645398	0.07438822	317.6607	1.620310	1642
GARCH(2,2) (lpc)	0.008821949	0.07549002	324.6830	1.596413	1642
APARCH(1,1)	0.008435548	0.07389520	315.4472	1.640718	1642
APARCH(1,1) (lpc)	0.008608753	0.07496037	322.5007	1.615285	1642

Note: MSE, MAE, MAPE and HMSE *standard* measures are provided with and without lpc for Northern Dynasty Minerals, Ltd. (NAK) for the models EWMA, Normal, Historic Simulations (HS), GARCH with skewed t-Student errors and APARCH with skewed t-Student errors.

Table 10: Adjusted Type 1 Evaluation Measures Results for PVaR With and Without Low Price Correction (lpc) of NAK

	MSE	MAE	MAPE	HMSE	Number of Violations
EWMA	0.007639176	0.04453162	26.02664	1.935885	66
EWMA (lpc)	0.007639877	0.04418522	25.19443	1.837000	65
Normal	0.008402171	0.05068587	26.36000	1.625744	62
Normal (lpc)	0.008505468	0.05119064	26.16090	1.530473	60
HS	0.005996688	0.04143981	26.46755	1.428645	102
HS (lpc)	0.006074069	0.04175005	26.25985	1.363650	99
GARCH(1,1)	0.006066455	0.04011085	26.98915	1.832928	94
GARCH(1,1) (lpc)	0.006277397	0.04128416	27.30712	1.779717	89
GARCH(2,2)	0.006327416	0.04170298	28.05656	1.902314	90
GARCH(2,2) (lpc)	0.006719502	0.04402593	29.09760	1.900888	83
APARCH(1,1)	0.006006253	0.04017759	26.93264	1.952717	97
APARCH(1,1) (lpc)	0.006370651	0.04217833	27.69914	1.958707	90

Note: MSE, MAE, MAPE and HMSE *adjusted* measures given in (2) are provided with and without lpc for Northern Dynasty Minerals, Ltd. (NAK) for the models EWMA, Normal, Historic Simulations (HS), GARCH with skewed t-Student errors and APARCH with skewed t-Student errors.

Table 11: Adjusted Type 1 Evaluation Measures Results for PVaR With and Without Low Price Correction (lpc) of NAK

	MSE	MAE	MAPE	HMSE	Number of Violations
EWMA	0.007756672	0.04519538	26.38935	1.965658	65
EWMA (lpc)	0.007639877	0.04418522	25.19443	1.837000	65
Normal	0.008682149	0.05233320	27.17981	1.679915	60
Normal (lpc)	0.008505468	0.05119064	26.16090	1.530473	60
HS	0.006178286	0.04263564	27.17236	1.471903	99
HS (lpc)	0.006074069	0.04175005	26.25985	1.363650	99
GARCH(1,1)	0.006406549	0.04219355	28.24221	1.935671	89
GARCH(1,1) (lpc)	0.006277397	0.04128416	27.30712	1.779717	89
GARCH(2,2)	0.006860086	0.04500551	30.09218	2.062461	83
GARCH(2,2) (lpc)	0.006719502	0.04402593	29.09760	1.900888	83
APARCH(1,1)	0.006472701	0.04313437	28.76565	2.104373	90
APARCH(1,1) (lpc)	0.006370651	0.04217833	27.69914	1.958707	90

Note: MSE, MAE, MAPE and HMSE *Type 1* measures for the final days of violations are provided with and without lpc for Northern Dynasty Minerals, Ltd. (NAK) for the models EWMA, Normal, Historic Simulations (HS), GARCH with skewed t-Student errors and APARCH with skewed t-Student errors.

Table 12: Adjusted Type 1 Evaluation Measures Results for PVaR With and Without Low Price Correction (lpc) of NAK

	MSE	MAE	MAPE	HMSE	Number of Violations	M (*)
EWMA	0.005783502	0.05026074	203.1109	1.1478504	66	732
EWMA (lpc)	0.005855648	0.05078779	208.2615	1.0977242	65	733
Normal	0.005466878	0.05030973	204.9854	0.9769964	62	736
Normal (lpc)	0.005546999	0.05119376	212.3019	0.9261190	60	738
HS	0.004351346	0.04089672	140.4530	1.0091305	102	696
HS (lpc)	0.004413035	0.04156739	145.8659	0.9645801	99	699
GARCH(1,1)	0.004474295	0.04057048	143.7540	1.2120823	94	704
GARCH(1,1) (lpc)	0.004587856	0.04173164	151.1883	1.1624685	89	709
GARCH(2,2)	0.004567731	0.04135123	145.4559	1.2322970	90	708
GARCH(2,2) (lpc)	0.004738480	0.04304300	154.9257	1.1989760	83	715
APARCH(1,1)	0.004407723	0.04039633	140.5289	1.2949761	97	701
APARCH(1,1) (lpc)	0.004577134	0.04194031	149.2077	1.2671618	90	708

Note: MSE, MAE, MAPE and HMSE *Type 1* measures given in (3) are provided with and without lpc for Northern Dynasty Minerals, Ltd. (NAK) for the models EWMA, Normal, Historic Simulations (HS), GARCH with skewed t-Student errors and APARCH with skewed t-Student errors. (*) M is the number of losses that did not violate the PVaR.

Table 13: Adjusted Type 2 Evaluation Measures Results for PVaR With and Without Low Price Correction (lpc) of NAK

	MSE	MAE	MAPE	HMSE	Number of Violations
EWMA	0.004548872	0.06121097	437.8729	1.305457	1642
EWMA (lpc)	0.004881161	0.06386919	456.6321	1.276753	1642
Normal	0.012498704	0.11032323	787.3810	1.117259	1642
Normal (lpc)	0.013002047	0.11266620	803.0433	1.112968	1642
HS	0.008485877	0.08887018	632.3044	1.156358	1642
HS (lpc)	0.008890938	0.09126820	650.0581	1.151683	1642
GARCH(1,1)	0.006948708	0.07920440	564.2632	1.192386	1642
GARCH(1,1) (lpc)	0.007330618	0.08192787	582.4148	1.187543	1642
GARCH(2,2)	0.006356866	0.07447842	529.4152	1.211453	1642
GARCH(2,2) (lpc)	0.006767026	0.07746784	551.9782	1.203366	1642
APARCH(1,1)	0.006909158	0.07894514	562.4500	1.196158	1642
APARCH(1,1) (lpc)	0.007282727	0.08161779	580.4357	1.190488	1642

Note: MSE, MAE, MAPE and HMSE *Type 2* measures are provided with and without lpc for Northern Dynasty Minerals, Ltd. (NAK) for the models EWMA, Normal, Historic Simulations (HS), GARCH with skewed t-Student errors and APARCH with skewed t-Student errors.