University of Warsaw
Faculty of Economic Sciences

Working Papers
No. 9/2020 (315)

SIZE DOES MATTER
A STUDY ON THE REQUIRED WINDOW SIZE FOR OPTIMAL QUALITY MARKET RISK MODELS

Mateusz Buczyński
Marcin Chlebus

Warsaw 2020

WORKING PAPERS 9/2020 (315)

Size does matter. A study on the required window size for optimal quality market risk models

Mateusz Buczyński[a]*, Marcin Chlebus[b]

[a] *Interdisciplinary Doctoral School, University of Warsaw*
[b] *Faculty of Economic Sciences, University of Warsaw*
* *Corresponding author: mp.buczynski2@uw.edu.pl*

Abstract: When it comes to market risk models, should we use full data that we possess or rather find a sufficient subsample? We have conducted a study of different fixed moving window's lengths (from 300 to 2000 observations) for three Value-at-Risk models: historical simulation, GARCH and CAViaR model for three different indexes: WIG20, S&P500 and FTSE100. Testing samples contained 250 observations, each ending with the end of years 2015-2019. We have also addressed the subjectivity of choosing the window's size by testing change points detection algorithms: binary segmentation and Pelt; to find the best matching cut-off point. Results indicate that the size of the training sample greater than 900-1000 observations doesn't increase the quality of the model, while the lengths lower than such cut-off provide unsatisfactory results and decrease model's conservatism. Change point detection methods provide more accurate models. Applying the algorithms with every model's recalculation provides results better by on average 1 exceedance. Our recommendation is to use GARCH or CAViaR model with recalculated window size.

Keywords: Value at Risk; historical simulation; CAViaR; GARCH; forecast comparison; sample size

JEL codes: G32, C52, C53, C58

1 Introduction

According to the econometric modeling strategies, there always exists a minimum number of observations that allows to draw any conclusions from estimated covariates. In an equation for confidence intervals, we notice that the uncertainty decreases as the number of observations increases. In terms of market risk forecasting, it suggests that the more data is fed to any model, the more certain and accurate are its forecasts. But is it true?

VaR is one of the most popular measures of market risk presented in a form of maximum loss over a target horizon that will not be exceeded with a given confidence level, given normal market conditions (Philippe (2006), Dowd (2010)). There are many approaches to the VaR modeling, but mostly it can be understood as a specific quantile of an assumed distribution of predicted rate of return realizations. Our point of interest is to study how many observations should be used to build such a distribution. Too few observations and the training sample's distribution will be very volatile, and vice versa - too many observations will expose the model to unnecessary bias that is not observed at the day of modeling. Therefore the task of determining the size of the training window is not as straightforward as it would seem. In addition to that there is no empirical consensus due to a lack of broad studies in that area.

The document that regulates modeling approach to market risk in financial institutions is Basel III, laying out rules for internal models creation to be followed by any banks and funds (Lee (2014)). According to its newest complement (from 2017, which emerged due to 2010s crises), any market risks should be defined by a measure of expected shortfall (which is conditional VaR) at 2.5% confidence level for at least next 10 trading days. As for the necessary time series length used to build the model, the necessary minimum number of observations is 250 trading days (approximately one year).

Regardless of the predetermined rules, Basel III does not specify any particular approach to VaR modelling. There are three main families of models to be considered: parametric approaches that assume a specific distribution of rate of return realizations and aim to estimate its parameters (eg. GARCH models); non-parametric approaches that do not assume any distributions, with estimates based only on empirical data (eg. historical simulation); semi-parametric approaches that have characteristics of both former families (eg. CAViaR model). Due to such fruitfulness of approaches there exist many comparisons of different approaches in different scenarios, mentioning best scenarios to use particular models (Abad et al. (2014)).

In the field of empirical studies devoted to VaR, many researchers are focused on comparisons of several different approaches. None of these papers state any model to be the best, but they specify characteristics of market conditions in which particular models perform better. Most of the recent

papers are in favor of semi-parametric methods, which are both accurate and flexible (Patton et al. (2019), Wang and Zhao (2016), Abad and Benito (2013), Martins-Filho et al. (2018), Taylor (2019), Abad et al. (2016), Nozari et al. (2010), Şener et al. (2012)). However, there are researchers, who find parametric methods better, claiming that modeling the distribution is more accurate (Buczyński and Chlebus (2018), Buczyński and Chlebus (2019), Ergün and Jun (2010), Berkowitz and O'Brien (2002), Bao et al. (2006), Consigli (2002), Danielsson (2002), Sarma et al. (2003)). In addition to that most of the aforementioned studies compare favored models to the most popular non-parametric method - historical simulation. These studies find that historical simulation might be worse than other studied models.

Unfortunately, most of studies in this field assume the length of training window beforehand. Very long training sample's window is mostly assumed, with sizes of 1000 - 2000 obs. and more (Bao et al. (2006), Danielsson (2002), Sarma et al. (2003), Patton et al. (2019), Martins-Filho et al. (2018), Nozari et al. (2010), Buczyński and Chlebus (2018), Buczyński and Chlebus (2019), Ergün and Jun (2010)). Some of the already mentioned researchers use a small range of different window's lengths, but do not draw any conclusions towards particular model's sensitivity towards these values. Some comments are found in Hendricks et al. (1996), saying that longer window sizes produced forecasts of better quality. Researchers rather rarely try smaller window's lengths (Wang and Zhao (2016), Abad and Benito (2013), Şener et al. (2012), Berkowitz and O'Brien (2002)). Finally, some researchers do not report the length of training sample at all.

Some attention recently has been given to automatic methods of detecting the appropriate sample size. Most popular approach is to find the closest change point in the time series to train the model on homogenic (in terms of expected value or volatility) series, assuming normal (stable) market conditions. For example Smith and Huang (2019) explored two approaches to finding such point: AMOC and binary segmentation. Their results indicate that these methods might be more precise than fixed training sample size. Another researchers (Čížek et al. (2009)) argue that by employing change point techniques one can achieve more accurate and flexible model that works over longer periods of time.

The aim of this paper is to compare different VaR approaches for multiple sample sizes. Primarily, we want to estimate VaR models for window lengths from 50 to 2000 and compare their excess ratios to find out whether there is any level over which increasing sample size does not make any betterment in terms of greater quality. Such analysis can provide a comprehensive overview of sample size selection for a particular model. To create the most unconditional environment for these models we have tested 15 different time series for each approach (five time periods for three different indexes). In addition to that, we specify a non-subjective criterion to find out the best fitting sample size. By that we have selected two change point detection algorithms: binary segmentation

and Pelt algorithm to find the most fitting window size. These algorithms are used both before and during estimation process.

In the next section, methodologies of VaR approaches used in the paper, are presented. In section three, we present the data and experiment setup. In section four we present the results and in the last section main conclusions are presented.

2 Methodology

VaR is defined as a maximum loss over a given time horizon t, at a given level of confidence α and normal market conditions. Importantly, VaR is also a quantile of the empirical distribution of gains and losses over selected time horizon. Mathematical equation to define VaR could be presented as follows (Philippe (2006)):

$$P(r_t < VaR_\alpha(t)|\Omega_{t-1}) = \alpha \tag{1}$$

where r_t is the rate of return of the asset under consideration and Ω_{t-1} is an information set given at time $t-1$.

It is also important to present how VaR models are backtested. One of the simplest approaches is to count the number of occurrences when VaR forecast was lower than the realization of the rate of return. Such measure is called an exceedance I_t and when expressed in terms of relation to whole backtested horizon of length N, we may introduce excess ratio (Philippe (2006)):

$$\hat{\alpha} = 1/N \sum_{t=1}^{N} I_t. \tag{2}$$

2.1 Historical simulation

The simplest non-parametric approach to VaR modelling is historical simulation (Dowd (2010)). It is based on the aforementioned fact that VaR is a quantile of historical returns. VaR is summarized by an α quantile of the empirical distribution of rates of return of the studied asset. In this approach, very much depends on the sample size, due to intentional (or not) inclusion of time periods of heterogenic volatility. The practice shows that the width of the window is fixed and usually ranges from 6 months to 2 years (125 - 500 obs.) (Engle and Manganelli (2001)).

$$VaR_\alpha(t) = q_\alpha \tag{3}$$

2.2 GARCH model

Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) models are one the most common parametric VaR models used in market risk modeling right now. Specifically in this paper the GARCH model under consideration is GARCH(1,1) models with skewed student's t distribution. GARCH(1,1) is one of the simplest of the whole GARCH family models and can be described by the two following equations (Engle (1982), Bollerslev (1986)):

$$\begin{aligned} r_t|\Omega_{t-1} &\sim IID(\mu_t, h_t), \\ h_t &= \beta_0 + \beta_1 \epsilon_{t-1}^2 + \gamma_1 h_{t-1}, \end{aligned} \tag{4}$$

where $IID(\mu_t, h_t)$ is an identical and independent distribution with μ_t conditional mean and h_t conditional variance, which on the other hand is explained by a sum of specific number of lagged squared error terms and conditional variances weighted by two vectors of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

Given these equations, we can define Value-at-Risk as Angelidis et al. (2004):

$$VaR_\alpha(t) = \hat{\mu}_t + q_\alpha \sqrt{\hat{h}_t}, \tag{5}$$

where q_α is an α quantile of the assumed distribution, while $\hat{\mu}_t$ and $\hat{h}_t$ are estimated conditional mean and variance for time t.

Theoretically, only normal distribution should be used as the conditional distribution of the model, however Bollerslev and Wooldridge (1992) have proven that if the model is not conditionally normally distributed, but it specifies the first two conditional moments correctly, the estimates of the quasi likelihood function will be consistent and asymptotically normal. Therefore the usage of distributions of the underlying process other than normal is completely correct and desirable. It is a common characteristic of any time series with financial origin that the distribution of the returns is skewed and has a tendency to have non-zero kurtosis, which drastically lowers the quality of models based on normal distribution. Based on literature, most of the studies finds student's t distribution to be the most fitting, in particular the skewed version (Ergen (2012)).

2.3 CAViaR model

One of the most common semi-parametric VaR models is CAViaR model introduced by Engle and Manganelli (2004). CAViaR model estimates the quantile of the distribution of the data directly instead of trying to model the whole distribution. The model is based on the quantile regression methodology by Koenker and Bassett (1978). The basic formula for CAViaR model (with one lagged Value-at-Risk and one lagged observed value) can be expressed as:

$$VaR_\alpha(t) = \beta_0 + \beta_1 VaR_\alpha(t-1) + l(\beta_2, r_{t-1}, VaR_\alpha(t-1)) \tag{6}$$

where $l(.)$ is a linking function of a lagged value of observables and VaR, while $\boldsymbol{\beta}$ is a vector of parameters.

The point of using linking function within the CAViaR expression is to link the model outcome to the level of rates of return at time $t-1$. In this study we have decided to use one of the CAViaR specifications presented by Engle and Manganelli (2004) - indirect GARCH:

$$VaR_\alpha(t) = \sqrt{\beta_0 + \beta_1 VaR_\alpha(t-1) + \beta_2 (r_{t-1})^2}, \tag{7}$$

The indirect GARCH approach to CAViaR models is very similar to GARCH modeling. In fact, it would be correctly specified model if the underlying data were GARCH(1,1) process with IID distribution.

2.4 Change point detection

Change point detection algorithms struggle to find an observation that determines an influential change in the time series. The main objective of these algorithms is to build a non-overlapping segmentation of the underlying model of time series, based on the detected shifts in its characteristics. The area of research in this topic is very broad, as these techniques are widely used in signal processing and have many applications in finance, bioinformatics, medicine and many more (Aminikhanghahi and Cook (2017)).

To provide a theoretical background, let us consider a non-stationary random process $y = \{y_1, ..., y_t\}$. This process is also assumed to be piece-wise stationary, i.e. there are K unknown instants $t_1 < t_2 < ... < t_K$, at which some characteristics of this process change. The aim of the change point detection algorithms is to find the best possible segmentation τ of the series, according to some general cost function $V(\tau, y) := \sum_{k=1}^{K} c(y_{t_k}, ..., y_{t_{k+1}})$ that is a summation over cost functions for particular segments. In our scenario, we do not determine the number of segments K beforehand, hence general cost function gets an additional penalty for the complexity of segmentation τ. Therefore, following Truong et al. (2020) the optimization problem can be determined as: $\min_{\tau} V(\tau) + pen(\tau)$

The cost function under consideration in this study is based on kernel methods. The original series is mapped onto Hilbert space $\mathcal{H}$. The mapping function $\phi : \mathbb{R}^d \to \mathcal{H}$ is given implicitly by $\phi(y_t) = k(y_t, \cdot) \in \mathcal{H}$. In such setting, the cost function of a particular segment can be defined as:

$$c(y_{t_k}, ..., y_{t_{k+1}}) := \sum_{t=t_k+1}^{t_{k+1}} \|\phi(y_t) - \bar{\mu}_{t_k,...,t_{k+1}}\|^2_{\mathcal{H}} \tag{8}$$

where $\bar{\mu}_{t_k,...,t_{k+1}}$ is the empirical mean of the process over sub-sample from t_k to t_{k+1} and $\|\cdot\|_{\mathcal{H}}$ is a norm in the Hilbert space. Of course the choice of the kernel function is unlimited, whereas the most commonly used kernel for numerical data is either Gaussian or linear. In the Gaussian scenario, we can define:

$$k(x, y) = exp(-\gamma\|x - y\|^2) \tag{9}$$

where $x, y \in \mathbb{R}^d$ and $\gamma > 0$ is called a bandwidth parameter. The cost function is therefore defined by (Truong et al. (2020)):

$$c(y_{t_k}, ..., y_{t_{k+1}}) := (t_{k+1} - t_k) - 1/(t_{k+1} - t_k) \sum_{s,t=t_k+1}^{t_{k+1}} exp(-\gamma\|y_s - y_t\|^2) \tag{10}$$

Regardless of the chosen cost function, there are several search method, which are procedures for discrete optimization processes aimed at minimization of the formulated cost function. We have two of such approaches: an optimal segmentation using Pelt algorithm and approximate by binary segmentation.

To find the optimal number of segments K one could run the optimization for each K and select the minimum. Fortunately, in case of linear penalties for the number of segments, high computational cost can be avoided, by the usage of Pelt algorithm. The algorithm considers the series sequentially and based on the pruning rule, may or may not include it in the set of potential change points. The pruning rule may be determined by:

if $[\min_{\tau} V(\tau, y_{0..t}) + pen(\tau)] + c(y_{t..s}) \geq [\min_{\tau} V(\tau, y_{0..s}) + pen(\tau)]$ then t cannot be the last change point prior to T. In the literature there are several usages of Pelt algorithm for example in DNA sequences and oceanographic data (Killick et al. (2012), Hocking et al. (2013), Maidstone et al. (2017)).

On the other hand, binary segmentation is more greedy sequential algorithm. First change point $\hat{t}_1$ is given by

$$\hat{t}_1 := argmin_{1 \leq t \leq T-1} c(y_0, ..., y_t) + c(y_t, ..., y_T) \tag{11}$$

which means that the algorithm searches for the change point t that minimizes the sum of costs.

The series is then split in two at the instant $\hat{t}_1$ and the same operation is repeated on the resulting sub-samples until no further improvement to the cost function can be done. The solution is only an approximation of a perfect segmentation, since the detection of a change point is not based on a homogeneous sample and every other detection is based on all previous ones. However that doesn't diminish the quality of algorithm as it was used in many applications in finance (Lavielle and Teyssière (2007), Bai (1997), Fryzlewicz (2014)), as well as bio-informatics and DNA sub-sampling (Niu and Zhang (2012), Olshen et al. (2004)).

3 Data and experiment setup

The experiment has been conducted on a set of main stock indexes from three countries: WIG20 (Poland), S&P500 (USA) and FTSE 100 (United Kingdom). There were five distinctive testing samples prepared, each consisting of 250 obs. and set to end with end of years: 2015 – 2019. Therefore, for each studied Value-at-Risk approach (historical simulation, GARCH model with skewed Student t distribution and CAViaR model), we tested three different indexes on five different testing samples, resulting in fifteen different time series. Each of the models were used 250 times to generate a one-day-ahead forecast using a sliding window technique with the length determined as described below.

The training sample was our point of interest for the experiment. In this paper, we compare VaR models in terms of out-of-sample one-step-ahead predictive ability. For each of the tested time series, the 250th observation since the end of the year (very beginning of January) constituted as a constant point and determined beginning of a testing sample. We have played with the length of the training sample, described as number of observations to be included. The numbers that we have tested are from a set $\{50, 100, \ldots, 1250\}$. We have also tested very large samples from a set of $\{1500, 1750, 2000\}$. For each of the models, time series and training sample lengths we have calculated number of exceedances, results of which are presented in tables 1, 2, 3. In addition to that we have calculated mean number of exceedances which is an average over tested years for a particular model and training sample size. All the resulting numbers are compared to the assumed number of exceedances, which for VaR at 2.5% confidence level and 250 testing sample obs. should be equal to 6.25 ± 4 (90% CI).

We have also created a mechanism for the automatic training sample's length selection by applying the change point detection techniques mentioned in section 2.4. The first part of the reported results refers to the change point detection applied **beforehand** the model fitting procedure for all forecasts in the testing sample. The second part refers to **recalculation** of the sample size using change point detection algorithm right before training another model for one-day-ahead forecast, hence the algorithm was also applied 250 times. For both methods we have used the aforementioned **Pelt** and **binary segmentation** algorithms. To create even more controlled environment we

propose two approaches to the change point detection algorithms: **liberal** and **conservative** one. Both approaches are based on all the change points detected in range from 500 to 1000, but for the liberal approach we select the change point closest to the 500th obs. and the conservative approach assumes the change point closest to the 1000th obs. Therefore liberal approach takes the shortest sample size and conservative one takes the longest. This results in four different scenarios of training sample lengths in each setup: conservative Pelt and binary segmentation; liberal Pelt and binary segmentation. Results for the automatic change point detection method are presented in tables 4 and 5.

Table 1: Number of exceedances for particular training sample sizes in tested years (WIG 20)

	historical sim.						GARCH (skewed student t)						CAViaR					
	2015	2016	2017	2018	2019	**Mean**	2015	2016	2017	2018	2019	**Mean**	2015	2016	2017	2018	2019	**Mean**
2000	1	1	0	8	5	3.0	8	8	1	8	5	6.0	9	7	0	7	5	5.6
1750	1	4	0	8	5	3.6	8	7	1	8	4	5.6	9	7	0	8	5	5.8
1500	5	7	0	8	5	5.0	8	9	1	8	4	6.0	9	8	1	8	6	6.4
1250	6	8	0	8	5	5.4	8	8	1	8	5	6.0	9	6	1	9	4	5.8
1200	6	8	0	8	5	5.4	8	8	1	8	5	6.0	9	7	1	7	4	5.6
1150	6	7	0	8	5	5.2	8	7	1	8	4	5.6	9	7	1	8	4	5.8
1100	6	9	0	8	5	5.6	8	8	1	8	4	5.8	9	8	1	8	4	6.0
1050	7	9	0	8	5	5.8	8	9	1	8	4	6.0	9	8	1	8	4	6.0
1000	8	9	0	8	5	6.0	8	8	2	8	4	6.0	9	9	1	8	5	6.4
950	10	10	0	8	5	6.6	8	8	2	8	4	6.0	9	6	1	8	6	6.0
900	10	10	0	8	6	6.8	8	8	2	8	5	6.2	10	10	2	8	7	7.4
850	11	9	0	8	6	6.8	8	9	1	8	5	6.2	9	11	1	8	5	6.8
800	12	9	0	8	6	7.0	8	10	1	8	5	6.4	11	9	1	8	6	7.0
750	13	9	0	8	6	7.2	8	10	1	8	5	6.4	11	10	1	11	5	7.6
700	12	9	0	8	6	7.0	8	11	1	8	5	6.6	12	10	1	13	6	8.4
650	11	9	0	9	6	7.0	8	10	0	8	5	6.2	11	10	1	16	6	8.8
600	13	9	0	11	6	7.8	9	10	2	9	-	-	11	10	4	12	6	8.6
550	14	10	0	12	6	8.4	10	11	1	9	-	-	13	8	7	13	4	9.0
500	15	8	0	13	6	8.4	10	10	2	10	-	-	12	9	6	14	6	9.4
450	16	8	1	15	5	9.0	10	8	2	-	-	-	13	8	4	16	5	9.2
400	17	8	2	15	5	9.4	11	9	2	-	-	-	11	6	4	15	5	8.2
350	16	8	2	19	5	10.0	9	6	2	-	-	-	13	6	4	19	6	9.6
300	15	8	2	18	5	9.6	11	6	2	-	-	-	13	5	2	17	5	8.4
250	14	8	2	17	5	9.2	-	6	-	-	-	-	-	-	-	-	-	-
200	14	6	2	15	6	8.6	-	6	-	-	-	-	-	-	-	-	-	-
150	11	6	3	12	5	7.4	-	-	-	-	-	-	-	-	-	-	-	-
100	12	6	7	9	4	7.6	-	-	-	-	-	-	-	-	-	-	-	-
50	16	10	11	9	9	11.0	-	-	-	-	-	-	-	-	-	-	-	-

The rows indicate the number of obs. in fixed size moving training window. In several cases GARCH model couldn't reach convergence, hence the results are not reported. The same applies to CAViaR model, where limit of minimum 300 observations has been suggested by Engle and Manganelli (2004) to avoid lack of convergence. Mean for these cases is not reported as well.

4 Results

4.1 Fixed training sample's size

The results indicate that for each of the tested indexes and models there is a downright tendency of mean exceedances with increasing length of the training sample. For the smallest numbers of obs. in the training sample, we can observe a very differing, often high number of exceedances over the studied years. The stabilization of the number of exceedances, regardless of the underlying time series and the model, starts with circa 900 – 1000 obs. It is a subjectively chosen point, but having studied all the cases separately, we draw a conclusion that since this point, if we increase the number of observations in the sample, we do not see any significant increase in the number of exceedances. There is also a small, but worth noting upright trend of exceedances for the smallest learning sample sizes (excluding the cases where the algorithm didn't converge).

In the study, we have also researched very long spans of the training samples (up to 2000 obs.) to prove that the information that is added to the model with such large lengths is not reflected in an improvement in results. For each of the studied indexes we see very low improvement in the number of exceedances, apart from the historical simulation's results, which tend to score minimums for lengths of 1500 – 2000, and we believe that such a tendency would apply for even larger training samples.

Historical simulation's results tend to be in accordance with literature. The lowest score for each of the indexes belongs to historical simulation on a very long span of learning sample (2000 obs.), however such a score is below the assumed number of exceedances. For the shorter training samples we observe much worse results than for the remaining models. In addition to that we do not observe a 'convergence' in the number of exceedances, which is clearly seen in the results of GARCH and CAViaR models. This leads to a conclusion that the assumed significance level of VaR model is not important at all, because it can be easily affected by the number of obs. in the training sample. Nevertheless, the assumed significance level is met for most of the series at 600 – 800 observations, but for some it is never met, and for some it is met even for very small training samples.

The GARCH and CAViaR results are very similar, with CAViaR's tendency to be slightly worse in the number of the exceedances for the years that were more volatile. The results show that they correctly estimate the risk to be more or less the assumed level for at least 500 - 700 obs. in the training sample for GARCH model and 700 – 900 obs. in the training sample for CAViaR model. The higher the number of obs. in training sample for these models, the lower the estimated number of exceedances, of course, but the profit that we obtain from increasing the sample becomes lower. The only flaw of both these models are problems with convergence for very small samples, which

is very natural, but gives an advantage to the historical simulation if one needed to train on a little training sample. In addition, both these models are better in terms of mean exceedances compared at the same level of training sample size for all the tested sizes up to more or less 1000.

The downright tendency in the number of exceedances that we observe can be also subjectively divided into two distinctive ranges of training sample lengths: the liberal one (from 500 – 600 obs. to the threshold of convergence – 900 – 1000 obs.); and the conservative one (the remaining studied training sample sizes above the threshold of convergence). The division is based on the differing results of number of exceedances in each range for particular years. For the liberal sample size lengths we can observe large differences for years 2015 and 2018, which are characterized by several excessive volatility shocks. As we increase the number of obs. in the training sample the differences tend to diminish and the distribution of exceedances for particular year starts being more uniform, which in our opinion is a characteristic of conservative models.

Table 2: Number of exceedances for particular training sample sizes in tested years (S&P 500)

	historical sim.						GARCH (skewed student t)						CAViaR					
	2015	2016	2017	2018	2019	**Mean**	2015	2016	2017	2018	2019	**Mean**	2015	2016	2017	2018	2019	**Mean**
2000	3	1	0	15	4	4.6	7	3	4	8	7	5.8	4	2	3	7	6	4.4
1750	3	3	0	19	5	6.0	7	4	4	9	7	6.2	4	3	3	9	6	5.0
1500	5	5	0	20	5	7.0	7	4	4	9	7	6.2	6	3	3	9	6	5.4
1250	6	5	1	18	4	6.8	7	4	4	8	7	6.0	6	4	3	9	6	5.6
1200	6	5	1	19	4	7.0	7	4	4	8	7	6.0	7	4	3	9	6	5.8
1150	6	6	1	19	4	7.2	7	4	4	8	7	6.0	7	4	3	9	7	6.0
1100	6	6	1	17	4	6.8	8	4	4	8	7	6.2	7	4	3	9	7	6.0
1050	7	8	1	17	4	7.4	8	4	4	8	7	6.2	7	4	3	10	7	6.2
1000	8	8	1	17	4	7.6	9	4	4	8	7	6.4	7	4	3	9	8	6.2
950	8	8	1	16	4	7.4	10	4	4	8	7	6.6	9	4	3	9	8	6.6
900	9	8	1	15	4	7.4	10	4	4	8	7	6.6	9	4	3	9	8	6.6
850	10	7	1	15	4	7.4	10	4	4	8	7	6.6	9	4	3	10	8	6.8
800	10	8	1	17	4	8.0	10	4	4	9	7	6.8	8	4	3	10	7	6.4
750	10	8	1	17	4	8.0	10	4	4	9	6	6.6	9	4	4	9	8	6.8
700	10	7	1	20	4	8.4	10	4	4	9	6	6.6	9	4	3	11	7	6.8
650	10	7	1	19	4	8.2	10	4	4	9	6	6.6	9	4	3	10	6	6.4
600	10	6	1	20	4	8.2	10	4	4	9	6	6.6	9	4	3	11	7	6.8
550	10	5	0	21	4	8.0	10	4	4	9	6	6.6	9	4	4	12	6	7.0
500	10	5	0	24	4	8.6	10	4	4	10	6	6.8	9	4	4	13	6	7.2
450	10	5	0	22	4	8.2	10	-	4	10	6	-	9	4	5	12	6	7.2
400	10	5	3	19	4	8.2	-	-	4	9	6	-	9	5	5	11	7	7.4
350	9	5	3	18	4	7.8	9	-	4	9	6	-	9	5	5	12	6	7.4
300	10	5	3	18	4	8.0	9	-	4	11	6	-	8	5	5	12	6	7.2
250	9	5	6	17	4	8.2	-	-	6	11	6	-	-	-	-	-	-	-
200	8	4	5	15	4	7.2	-	-	7	10	7	-	-	-	-	-	-	-
150	8	5	7	16	3	7.8	-	-	8	-	7	-	-	-	-	-	-	-
100	8	4	8	13	4	7.4	-	-	8	-	-	-	-	-	-	-	-	-
50	13	8	12	11	6	10.0	-	-	-	-	-	-	-	-	-	-	-	-

The rows indicate the number of obs. in fixed size moving training window. In several cases GARCH model couldn't reach convergence, hence the results are not reported. The same applies to CAViaR model, where limit of minimum 300 observations has been suggested by Engle and Manganelli (2004) to avoid lack of convergence. Mean for these cases is not reported as well.

Table 3: Number of exceedances for particular training sample sizes in tested years (FTSE 100)

	historical sim.						GARCH (skewed student t)						CAViaR					
	2015	2016	2017	2018	2019	**Mean**	2015	2016	2017	2018	2019	**Mean**	2015	2016	2017	2018	2019	**Mean**
2000	4	3	1	3	4	3.0	8	4	4	7	7	6.0	8	4	2	6	5	5.0
1750	4	5	1	4	5	3.8	8	4	4	7	7	6.0	9	4	2	6	6	5.4
1500	9	6	1	4	4	4.8	9	4	4	6	7	6.0	9	4	2	6	5	5.2
1250	10	6	1	4	4	5.0	9	4	4	6	6	5.8	10	4	2	5	6	5.4
1200	10	6	1	4	4	5.0	9	4	4	6	6	5.8	10	4	2	5	5	5.2
1150	10	7	1	4	4	5.2	9	4	4	6	7	6.0	10	4	2	6	6	5.6
1100	10	7	1	4	4	5.2	9	4	4	6	7	6.0	10	4	2	6	6	5.6
1050	10	8	1	3	5	5.4	9	4	4	6	7	6.0	10	4	2	6	6	5.6
1000	11	8	1	3	5	5.6	9	4	4	6	7	6.0	10	4	3	6	7	6.0
950	13	7	1	4	5	6.0	9	4	4	6	7	6.0	10	4	2	7	7	6.0
900	13	9	1	4	5	6.4	9	4	4	6	7	6.0	10	4	3	6	6	5.8
850	13	9	1	4	6	6.6	11	5	4	6	7	6.6	10	4	3	8	7	6.4
800	14	7	1	4	6	6.4	11	4	4	6	7	6.4	10	4	4	9	7	6.8
750	16	7	1	4	6	6.8	11	4	4	7	7	6.6	10	5	4	9	7	7.0
700	15	6	1	5	6	6.6	11	4	4	6	7	6.4	10	3	3	10	7	6.6
650	13	7	1	6	6	6.6	11	4	4	9	7	7.0	10	2	4	12	8	7.2
600	14	6	1	8	6	7.0	11	4	4	9	7	7.0	10	3	4	15	8	8.0
550	15	6	1	10	6	7.6	11	4	4	11	7	7.4	9	3	6	15	9	8.4
500	14	6	1	13	6	8.0	11	5	4	10	7	7.4	10	3	6	17	9	9.0
450	13	5	1	14	6	7.8	-	5	4	11	7	-	11	3	6	15	8	8.6
400	12	4	1	15	6	7.6	11	5	4	11	7	7.6	10	3	4	-	9	-
350	11	4	2	14	6	7.4	11	4	4	12	7	7.6	12	3	4	-	8	-
300	11	4	2	14	6	7.4	12	-	4	11	7	-	9	4	5	-	8	-
250	8	3	2	14	6	6.6	12	-	5	11	7	-	-	-	-	-	-	-
200	7	4	5	13	6	7.0	10	-	3	13	7	-	-	-	-	-	-	-
150	8	3	8	12	6	7.4	-	-	-	-	7	-	-	-	-	-	-	-
100	9	6	9	14	6	8.8	-	-	-	-	-	-	-	-	-	-	-	-
50	13	12	12	15	7	11.8	-	-	-	-	-	-	-	-	-	-	-	-

The rows indicate the number of obs. in fixed size moving training window. In several cases GARCH model couldn't reach convergence, hence the results are not reported. The same applies to CAViaR model, where limit of minimum 300 observations has been suggested by Engle and Manganelli (2004) to avoid lack of convergence. Mean for these cases is not reported as well.

4.2 Automatic training sample's length selection

The results of the automatically chosen lengths of the training sample's for each studied index are presented in tables 4, 5. In addition to that, we report calculated sizes of windows for beforehand application of change point detection algorithms in 6. The results indicate that the automatically chosen span is in accordance with the previous conclusions about liberal and conservative samples. Our method to determine the best point is more or less stable and fluctuates around 500 – 600 obs. for the liberal method and around 900 – 1000 obs. for the more conservative one. It also needs to be noted that there is not much difference in the training sample's size determined by both of these

models. We cannot determine, which one of them is better, given that it was not the aim of this research.

The results for the automatically chosen lengths of the training sample's indicate that the length can be predetermined using an objective method and produce estimates of good quality and in accordance with previous conclusions. Mean number of exceedances for both methods fall very close to the assumed number of exceedances for CAViaR and GARCH models, whereas for the historical simulation the number of exceedances is slightly higher (7 to 8 on average). Such difference should be attributed to the previously drawn conclusion that historical simulation needs longer training sample to produce estimates of the same quality that other compared models.

Much better results are provided if the length of the training window is not predetermined, but recalculated with every other model fitting. Mean number of exceedances is lower for almost each of the methods in each setting (liberal, conservative). Even though the training sample did not exceed 1000 obs. mean number of exceedances lowered by 1 – 2 exceedances. For GARCH model mean number of exceedances is around 4 – 5 exceedances, with maximum of 9 exceedances. CAViaR model obtained slightly worse results, with around 5 – 6 exceedances, but mostly in point with assumed excess level. Maximum number of exceedances for CAViaR was 11. Historical simulation's results are still worse than for these two models, with 6 – 7 mean exceedances.

Table 4: Number of VaR exceedances using automatic change point detection sample size for beforehand sizes for WIG 20, S&P 500 and FTSE 100

	historical sim.						GARCH (skewed student t)						CAViaR					
	2015	2016	2017	2018	2019	**Mean**	2015	2016	2017	2018	2019	**Mean**	2015	2016	2017	2018	2019	**Mean**
WIG20																		
Pelt																		
liberal	14	9	0	13	6	8.4	10	10	1	9	5	7.0	11	11	2	12	6	8.4
conservative	11	9	0	8	5	6.6	8	10	2	8	4	6.4	8	9	2	8	5	6.4
BinSeg																		
liberal	13	9	0	10	6	7.6	9	11	0	8	5	6.6	11	10	1	11	3	7.2
conservative	12	9	0	8	5	6.8	8	8	0	8	4	5.6	9	10	2	8	6	7.0
S&P 500																		
Pelt																		
liberal	10	7	1	24	4	9.2	10	0	4	10	7	6.2	9	4	3	12	6	6.8
conservative	8	8	1	18	4	7.8	10	4	4	8	7	6.6	8	4	3	9	8	6.4
BinSeg																		
liberal	10	7	0	24	4	9.0	10	0	4	10	7	6.2	9	4	3	12	6	6.8
conservative	8	8	1	18	4	7.8	10	4	4	8	7	6.6	8	4	3	9	7	6.2
FTSE 100																		
Pelt																		
liberal	15	6	1	11	6	7.8	11	4	4	10	7	7.2	10	3	3	14	9	7.8
conservative	13	7	1	4	5	6.0	9	4	4	6	7	6.0	10	5	3	9	7	6.8
BinSeg																		
liberal	15	6	1	14	6	8.4	11	4	4	10	7	7.2	10	3	3	14	9	7.8
conservative	13	7	1	4	5	6.0	9	4	4	6	7	6.0	10	5	3	8	6	6.4

BinSeg applies to binary segmentation algorithm, while liberal and conservative approaches are further explained in the text.

Table 5: Number of VaR exceedances using automatic change point detection sample size for beforehand and recalculated sizes for WIG 20, S&P 500 and FTSE 100

	historical sim.						GARCH (skewed student t)						CAViaR					
	2015	2016	2017	2018	2019	**Mean**	2015	2016	2017	2018	2019	**Mean**	2015	2016	2017	2018	2019	**Mean**
WIG20																		
Pelt																		
liberal	12	9	0	8	6	7.0	8	8	2	8	5	6.2	11	7	2	8	7	7.0
conservative	6	8	0	8	5	5.4	8	8	1	8	5	6.0	9	8	1	9	4	6.2
BinSeg																		
liberal	9	9	0	8	5	6.2	8	8	2	8	4	6.0	10	6	2	9	5	6.4
conservative	6	8	0	8	5	5.4	8	7	1	8	5	5.8	9	7	1	8	5	6.0
S&P 500																		
Pelt																		
liberal	9	8	1	15	4	7.4	10	4	4	9	7	6.8	9	4	3	10	7	6.6
conservative	6	5	1	19	4	7.0	7	4	4	8	7	6.0	8	4	3	9	6	6.0
BinSeg																		
liberal	10	8	1	15	4	7.6	10	4	4	9	7	6.8	9	4	3	10	8	6.8
conservative	6	5	1	19	4	7.0	7	4	4	8	7	6.0	7	4	3	9	6	5.8
FTSE 100																		
Pelt																		
liberal	15	9	1	4	5	6.8	11	4	4	6	7	6.4	10	5	3	9	7	6.8
conservative	10	6	1	4	4	5.0	9	4	4	6	6	5.8	10	4	2	6	6	5.6
BinSeg																		
liberal	14	8	1	4	6	6.6	10	4	4	6	7	6.2	10	5	3	9	7	6.8
conservative	10	6	1	4	4	5.0	9	4	4	6	7	6.0	10	4	2	6	7	5.8

BinSeg applies to binary segmentation algorithm, while liberal and conservative approaches are further explained in the text.

5 Conclusions

In this study we have researched the impact of the training sample's size on the results of several Value-at-Risk models: historical simulation, GARCH model with skewed Student's t distribution and CAViaR model at 2.5% confidence level. Each of these approaches was tested 250 times on the span of five distinctive years: 2015 – 2019. The training sample sizes that we have tested are in a range from 50 to 1250, increasing by 50 (50, 100, …, 1250) and three additional sample sizes: 1500, 1750, 2000. In addition to that we have created a setup for an automatic detection of necessary sample size to provide sufficient results by utilization of change point detection algorithms: Pelt and binary segmentation in liberal and conservative setting. These methods were applied to determine the number of observations in training sample either before the whole fitting process or with each one-day-ahead forecast.

Table 6: Calculated number of observations using change point detection algorithms for WIG 20, S&P 500 and FTSE 100

	WIG 20				S&P 500				FTSE 100			
	Pelt		BinSeg		Pelt		BinSeg		Pelt		BinSeg	
	liberal	conservative	liberal	conservative	liberal	conservative	liberal	conservative	liberal	conservative	liberal	conservative
2015	550	850	585	855	605	975	605	975	535	965	535	965
2016	680	800	695	835	625	945	625	955	590	955	590	955
2017	685	930	675	675	575	975	520	950	580	915	580	915
2018	530	945	635	945	525	995	525	995	520	825	510	835
2019	725	980	615	980	625	950	625	985	535	895	540	915
Mean	634	901	641	858	591	968	580	972	552	911	551	917

BinSeg applies to binary segmentation algorithm, while liberal and conservative approaches are further explained in the text.

Based on the results that we have obtained, our recommendation would be to use the GARCH or CAViaR models with the proposed automatic training sample's length selection that is recalculated with every model's refitting. Using this approach, one can get much better results than while using a predefined window's length, whilst still being in the range of 500 – 1000 obs. We would like to emphasize that the mean number of exceedances for this method is below the assumed level, hence for shorted training samples it would reach the assumed level. We recommend to reject model's based on the historical simulation approach, as their results are much worse compared to other models trained on the same number of observations. In addition to that, it is easy to play with the assumed excess level, just by increasing the number of observations, which in our opinion is not in accordance with Basel rules.

In case when the automatically chosen lengths cannot be used or determined, we recommend to use at least 900 – 1000 observations, as we have proven that since that number the number of exceedances converges to the assumed excess level. In addition we recommend to use the training sample's size from the subjective division that we have created – in case the risk management tilts towards liberal solutions, the number of exceedances should be lower than the threshold we have set, and if the more conservative models are preferred the threshold should be risen to the levels above the threshold.

In the end, we would like to highlight that the model's accuracy is all about the information it gets while in training. It is obvious that the more shocks the model 'observes', while in training, the more biased towards conservatism it will be, and the other way around. Therefore, each decision

about the length of the training sample, should be based on the thorough study of the underlying time series, if applicable.

References

Abad, P. and Benito, S. (2013). A detailed comparison of value at risk estimates. *Mathematics and Computers in Simulation*, 94:258–276.

Abad, P., Benito, S., and López, C. (2014). A comprehensive review of value at risk methodologies. *The Spanish Review of Financial Economics*, 12(1):15–32.

Abad, P., Muela, S., Lopez, C., and Sánchez-Granero, M. (2016). Evaluating the performance of the skewed distributions to forecast value-at-risk in the global financial crisis. *The Journal of Risk*, 18.

Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367.

Angelidis, T., Benos, A., and Degiannakis, S. (2004). The use of garch models in var estimation. *Statistical Methodology*, 1(1):105–128.

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory*, 13(3):315–352.

Bao, Y., Lee, T.-H., and Saltoglu, B. (2006). Evaluating predictive performance of value-at-risk models in emerging markets: a reality check. *Journal of Forecasting*, 25(2):101–128.

Berkowitz, J. and O'Brien, J. (2002). How accurate are value-at-risk models at commercial banks? *The Journal of Finance*, 57(3):1093–1111.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bollerslev, T. and Wooldridge, J. M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews*, 11(2):143–172.

Buczyński, M. and Chlebus, M. (2018). Comparison of semi-parametric and benchmark value-at-risk models in several time periods with different volatility levels. *e-Finanse*, 14:67–82.

Buczyński, M. and Chlebus, M. (2019). Old-fashioned parametric models are still the best. a comparison of value-at-risk approaches in several volatility states. *Faculty of Economic Sciences, University of Warsaw Working Papers*, 12(297).

Consigli, G. (2002). Tail estimation and mean–var portfolio selection in markets subject to financial instability. *Journal of Banking & Finance*, 26(7):1355–1382.

Danielsson, J. (2002). The emperor has no clothes: Limits to risk modelling. *Journal of Banking & Finance*, 26(7):1273–1296.

Dowd, K. (2010). Value-at-risk. *Encyclopedia of Quantitative Finance*.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Engle, R. F. and Manganelli, S. (2001). Value at risk models in finance. Report, European Central Bank.

Engle, R. F. and Manganelli, S. (2004). Caviar. *Journal of Business & Economic Statistics*, 22(4):367–381.

Ergen, I. (2012). Var prediction for emerging stock markets: Garch filtered skewed t distribution and garch filtered evt method. *Federal Reserve Bank of Richmond working paper*.

Ergün, A. T. and Jun, J. (2010). Time-varying higher-order conditional moments and forecasting intraday var and expected shortfall. *The Quarterly Review of Economics and Finance*, 50(3):264–272.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, 42(6):2243–2281.

Hendricks, S. A., Wassell, J. T., Collins, J. W., and Sedlak, S. L. (1996). Power determination for geographically clustered data using generalized estimating equations. *Statistics in Medicine*, 15(18):1951–1960.

Hocking, T. D., Schleiermacher, G., Janoueix-Lerosey, I., Boeva, V., Cappo, J., Delattre, O., Bach, F., and Vert, J.-P. (2013). Learning smoothing models of copy number profiles using breakpoint annotations. *BMC Bioinformatics*, 14(1):164.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Koenker, R. W. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

Lavielle, M. and Teyssière, G. (2007). *Adaptive Detection of Multiple Change-Points in Asset Price Volatility*, pages 129–156. Springer Berlin Heidelberg, Berlin, Heidelberg.

Lee, E. (2014). Basel iii and its new capital requirements, as distinguished from basel ii. *The Banking Law Journal*, 131(1):27–69.

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2):519–533.

Martins-Filho, C., Yao, F., and Torero, M. (2018). Nonparametric estimation of conditional value-at-risk and expected shortfall based on extreme value theory. *Econometric Theory*, 34(1):23–67.

Niu, Y. S. and Zhang, H. (2012). The screening and ranking algorithm to detect dna copy number variations. *The annals of applied statistics*, 6(3):1306–1326.

Nozari, M., Raei, S., Jahangiry, P., and Bahramgiri, M. (2010). A comparison of heavy-tailed estimates and filtered historical simulation: Evidence from emerging markets. 6:347–359.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572.

Patton, A., Ziegel, J., and Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, 211.

Philippe, J. (2006). *Value at Risk, 3rd Ed*. McGraw-Hill.

Sarma, M., Thomas, S., and Shah, A. (2003). Selection of value-at-risk models. *Journal of Forecasting*, 22(4):337–358.

Smith, A. and Huang, C.-K. (2019). A study on window-size selection for threshold and bootstrap value-at-risk models. *Journal of Risk Model Validation*.

Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric laplace distribution. *Journal of Business & Economic Statistics*, 37(1):121–133.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Wang, C.-S. and Zhao, Z. (2016). Conditional value-at-risk: Semiparametric estimation and inference. *Journal of Econometrics*, 195(1):86–103.

Čížek, P., Härdle, W. K., and Spokoiny, V. (2009). Adaptive pointwise estimation in time-inhomogeneous conditional heteroscedasticity models. *Econometrics Journal*, 12:248–271.

Şener, E., Baronyan, S., and Ali Mengütürk, L. (2012). Ranking the predictive performances of value-at-risk estimation methods. *International Journal of Forecasting*, 28(4):849–873.

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl