alfi association of the luxembourg fund industry

guidelines

VAR MODEL BACKTESTING C.S.I.

Practitioners' Thoughts

in association with ALRiM LUXEMBOURG

table of contents

Industry participants have observed, between the end of 2014 and mid-2016, various episodes of market turbulence. The results of VaR models backtesting have often been underwhelming: portfolios of various types, monitored with different models, experienced many exceptions.

This gave birth to some soul searching among practitioners about whether the models they rely on to support decision making and to monitor risk and performance remain valid when the number of exceptions is persistently too high for comfort.

VaR models, being mere representations of the real world, rely on assumptions that are good proxies of real dynamics most of the time. But beyond this "most of the time" or "normal market conditions", dramatic failures may occur, especially when models reach their limits, as it is the case when volatility regimes switch or when liquidity vanishes. As reality changes, its representation must be periodically reviewed.

Following a first effort run jointly by ALFI Risk Management Group and ALRiM to better define what was backtesting and outline some basics in analysis, we decided to further investigate:

- How the industry was coping or could cope with these questions;
- What areas practitioners should further investigate before concluding models need to be adjusted or are simply not working[2].

Hence the CSI (Crime Scene Investigation) themed paper by analogy between the vastly popular television series and the post mortem investigations run by risk managers after the fact when models appear to fail.

This paper does not aim at being (i) a backtesting beginner's guide, (ii) a step by step exhaustive guide of how to run backtesting analysis or (iii) a methodology for validating a risk model. Its purpose is simply to share practitioners' views about (grey) areas of investigation when models fail backtesting procedures.

A poll and discussions among practitioners throughout the industry (21 asset managers, third party ManCos, consultants and independent directors covering 10 VaR model vendors and most relevant VaR models) revealed common ground on concerns about backtesting results, what areas to further investigate and how to go about these investigations. Participants had often been doing extensive work and tests on these topics.

These areas of concern and investigation may be grouped in the following broad categories.

- Embedded assumptions: VaR models rely on a set of assumptions relating to, among others, market data consistency and the absence of intraday trading. These may bias outlier numbers.
- Reference period: VaR models are supposed to generate about as many exceptions as can be expected given their confidence level over a supposedly representative period (e.g. between 0 and 4 per year for a VaR 99% model). If the sample is not representative, then the relationship may not hold and we may observe a number of exceptions that seems incompatible with the model specifications.
- Clustering of exceptions: independence tests run by practitioners examine a specific form of dependence. They check that the occurrence of an exception is not too frequently followed directly by another one. Since dependence may take many forms, we need to qualitatively investigate the occurrence of several exceptions within a short time frame.
- Market swings and exceptional events: some market or company-specific events may generate returns that are beyond what the models are designed to capture. Models are designed to capture systematic and idiosyncratic risks as opposed to event risk. Also, the CESR/ESMA guidelines themselves state that VaR models are built to perform as expected in "normal" market conditions. By this standard, extreme market swings are unlikely to be captured.
- Model reactivity: the model needs to strike a balance between providing risk measures stable enough to be used to manage a portfolio and adapting to changing market conditions.
- Further thoughts on the use of multiple models.

In short, our goal is to help practitioners find answers as to what they could do when faced with a risk model that fails regular backtests and analysis. Hence the analogy to a crime scene investigation (C.S.I.) and the risk manager running post-mortem investigations.

We would like to thank all participants for the quality of their input and their candor during these exchanges which, hopefully, will spur more widespread sharing of ideas about how to deal with this topic.

Although technical and possibly esoteric to some, backtesting may have practical consequences on day to day business. CSSF flags backtesting as one area which Boards of Directors need to be informed about on a regular basis.

This paper draws upon information and ideas provided by practitioners during forums, discussions and various email exchanges.

It is not intended to reflect the views of any specific company but the opinions of the individuals who have agreed to take part in such exchanges of views. The data underlying the analysis disclosed herein is briefly described (see e.g. Annex 1) but will not be made public or shared.

Please note that ALFI-ALRiM, while confident of the broad principles described in this white paper, does not endorse the practical use and specific implementation of these principles or of the related tools and spreadsheets. It is strongly advised to fully review and test this material before using it for professional purposes. ALFI, ALRiM and group members therefore cannot be held liable for any error, imprecision or mistake endured following the use of the provided material.

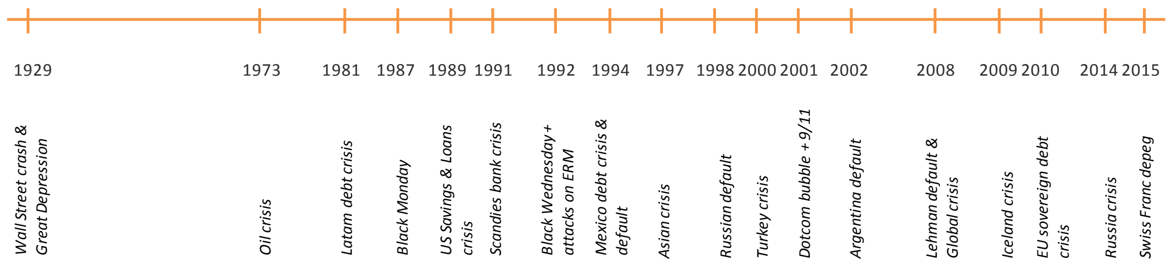
Introduction

This paper is the second issue of a three-paper series based on a joint effort between the ALFI Risk Management working group and ALRiM on the backtesting of VaR models. It follows a previous issue "ABC of VaR Model Backtesting" that described the conceptual and practical basics of VaR model backtesting. Based on industry feedback gathered during a practitioner forum, it discusses how to handle the all-too-common situation where a VaR model fails regular backtests. It leaves out, however, the discussion of governance arrangements around backtesting, which will be the topic of the final paper in this series.

Whenever a VaR model backtest uncovers many exceptions, even if some of them can be explained through regular analysis exercises, their number may be too high for comfort. Risk managers and conducting officers must then decide what to do. Regulations merely require a review of the model and its assumptions, and modifications if need be. Hence the analogy to a crime scene investigation (C.S.I.) and the risk manager running post-mortem investigations. This second paper, then, aims at exploring a variety of venues for investigation to help in this exercise.

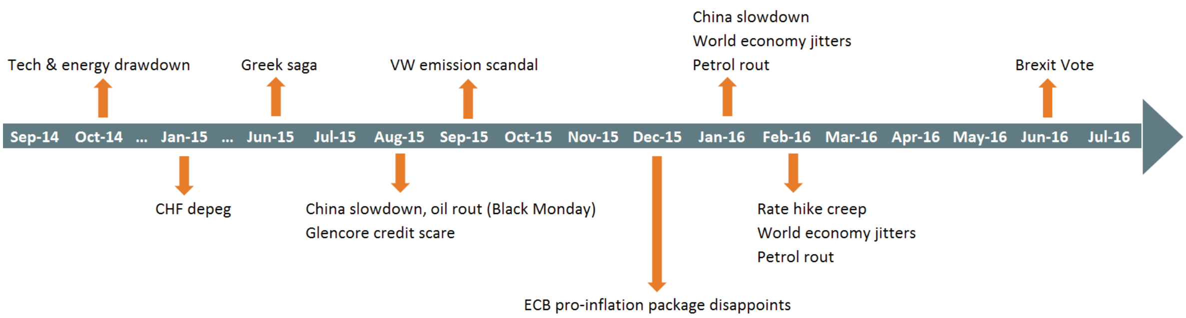
Are market turbulence & failing risk models the new norm?

Industry participants have been faced with crisis at a quickening pace especially since early 2000's as is shown in the following timeline.



In addition to full blown crisis, during the last few years[1] various episodes of market turbulence have also become more common as illustrated by the following graph covering 2014 through Q2 2016.

During these episodes, backtests of VaR models often yielded underwhelming results: portfolios of various types, monitored with different models, experienced too many exceptions given their VaR confidence level.



The fact that risk models may fall short of expectations in turbulent market environments is not new. The latter part of 2008 following the Lehman bankruptcy and ensuing events spreading throughout 2009 is another recent period where risk models often seemed to be failing.

Since then, CSSF has further beefed up its risk management capabilities and further sharpened its approach to risk management and related questions. In the context of RMP reviews, CSSF has provided feedback and touched upon the topic of backtesting exceptions.

The heart of the question and the related soul searching is whether the models remain valid given a number of exceptions too high for comfort.

We therefore decided to further investigate:

- How the industry was coping or could cope with these questions;
- What areas practitioners should further investigate before concluding that models need to be adjusted or are simply not working.

[1] Our analysis window covers Q4 2014 to Q2 2016 included.

What does the industry say?

In order to get feedback from industry participants, the first initiative was to organize a practitioner forum where industry members could exchange and openly discuss such topics, their concerns and how they were trying to manage related issues.

Initial feedbacks were that discussing and sharing experience about the steps taken to try and validate models in this challenging environment is an avenue worth exploring further, and that risk managers were willing to get involved in this process.

To get the best out of such discussions, we drew up a straw poll that participants answered before meeting. The questions answered went along the following lines:

- Model used: name of provider (e.g. RiskMetrics, Barra, etc.)
- Type of model: historical, monte-carlo, parametric, etc.
- Type of portfolios with too many exceptions: equity (vanilla or with derivatives), fixed income (vanilla or with derivatives), derivative based.
- Origin of exceptions: market volatility shifts, mapping, data, low reactivity, etc.
- Additional analysis undertaken or measures set-up: independent validation, use of second model for challenging, analysis of breaks, statistical tests, etc.

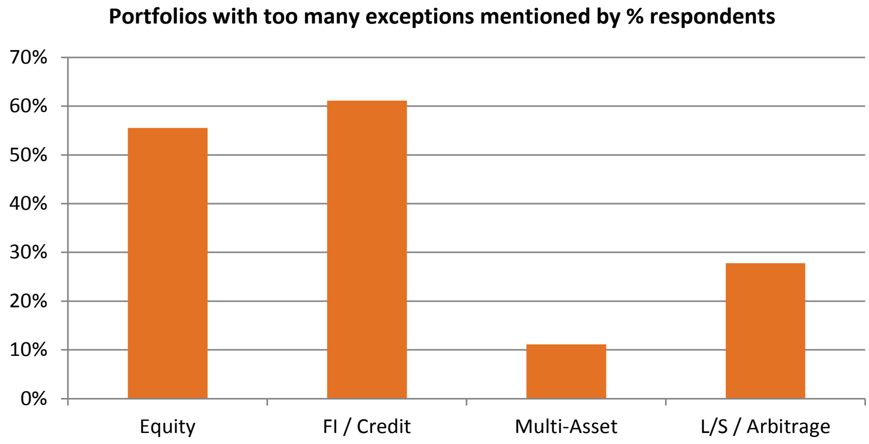
Questionnaires were sent out to 24 asset managers, third party ManCos, consultants and independent directors. Of these, 21 answered. These answers provide a broad coverage:

10 VaR model providers. In alphabetical order:	Various types of model:	9 respondent nationalities. In alphabetical order:
1. Algorithmics 2. APT 3. Barra 4. Gambit 5. Much-Net 6. Point 7. RiskMetrics 8. RMX 9. SAS 10. Statpro	1. Factor-based vs. full revaluation 2. Historical vs. Monte Carlo vs. parametric	1. French 2. German 3. Italian 4. Japanese 5. Lux 6. Nordic 7. Swiss 8. UK 9. US

From our poll, model use and results comes out as[2]:

- 44% of respondents use factor-based models and 56% full revaluation models
- 39% of respondents use historical VaR, 50% use Monte Carlo and 44% use parametric VaR.
- 83% of respondents reported portfolios with more than 4 exceptions. Portfolios with too many exceptions:
  - Equity mentioned by 56% of respondents. These were both plain vanilla equity and portfolios using derivatives.
  - Fixed Income and Credit mentioned by 61% of respondents. Emerging debt, high yield and convertible were often mentioned.
  - Multi-Asset mentioned by 11% of respondents, with exceptions being often due to the equity part of the portfolios.
  - L/S and arbitrage strategies mentioned by 28% of respondents.

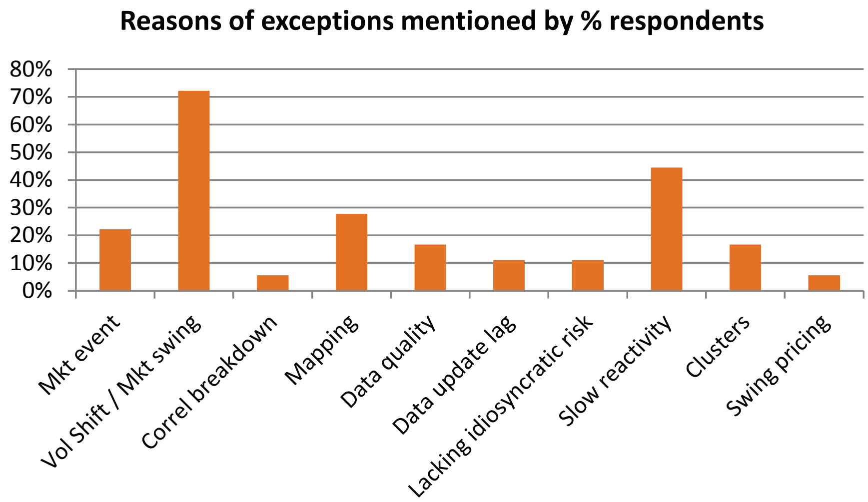
[2] Please note that, for most questions, sums don't add up to 100% as respondents could provide several answers e.g. use more than 1 model.



The following reasons for exceptions were the most often mentioned:

1. Most cited were volatility shifts and market swings (72% of respondents).
2. Slow model reactivity (44% of respondents). In a similar category, the fact that model data is not updated daily (often monthly) was mentioned by 11% of participants, and clusters of exceptions by 17%.

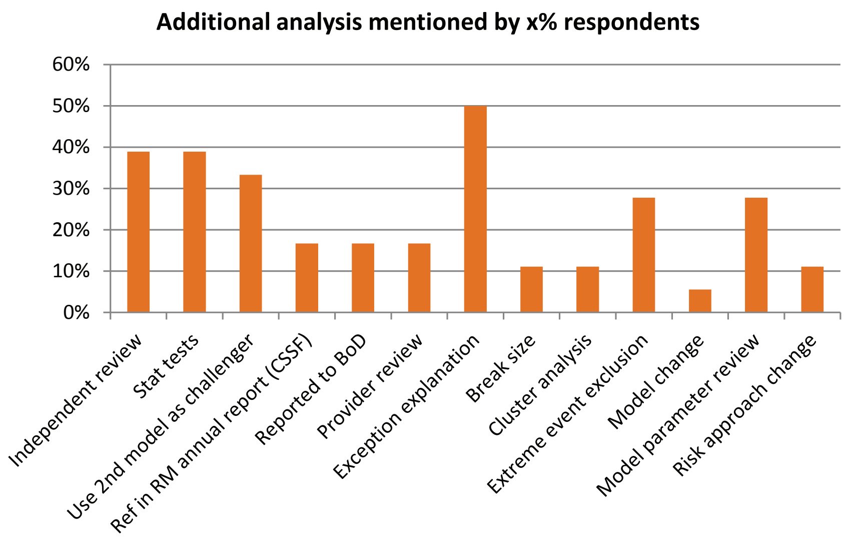
Mapping issues were mentioned by 28% of participants. To this, we can add the lack of idiosyncratic risk in the models (11%) and data quality (often mapping to a USD generic curve / 17%).



When looking at the additional analysis work or measures, the most mentioned ones are:

1. Review and explanation of exceptions: 50% of respondents.
2. Statistical tests aiming to qualify the number of exceptions:
   - Kupiec & Christoffersen: 39% of respondents.
   - Exclude extreme events: 28% of respondents.
   - Cluster/contagion analysis: 11% of respondents.
   - Consider size of the break: 11% of respondents.
3. Independent review and validation of model: 39% of respondents. 17% of respondents also went back to their providers for further validation.
4. Use a 2nd model as challenging (whether from a different provider, or historical vs. MC simulation, or long term vs. short term): 33%.

5. Review of model parameters (decay factor, lookback period, etc.) and fine-tuning them outright or correcting them when volatility spikes is explored by 28% of respondents. To this, we can add that 6% are considering changing model (e.g. go from historical simulation to MC).
6. Reporting to CSSF and/or BoD: 17% each.
7. Changing the risk approach (from absolute to relative VaR or to commitment) is considered by 11% of respondents.



The forum allowed practitioners to discuss these items and bring forward additional areas they had been investigating or were looking to investigate when testing models in this difficult environment. Some of these areas, highlighted and explored in the following sections, draw upon information and ideas provided by practitioners during forums, discussions and various email exchanges. It is not intended to reflect the views of any specific company but the opinions of the individuals who have agreed to take part in such exchanges of views. The data underlying the analysis disclosed herein will not be made public or shared.

Most participants to the practitioner forum felt that, although models could be improved in various respects, they were not faulty or deficient. The prevailing conviction, supported by the results of ongoing analysis, was that the high number of exceptions was due mainly to market swings and event risk beyond what the models are built to capture.

In order to put to rest any suspicion that models were no longer adequate, practitioners believed it was worthwhile to investigate additional testing grounds that may not be fully covered by ongoing analysis.

As mentioned in a previous paper 'ABC of VaR Model Backtesting', areas to investigate as a first step include:

- Reviewing moves of the underlying risk factors;
- Checking for large changes in positions;
- Probing for event risk that might not have been captured (market-impacting events such as FED tapering, Lehman bankruptcy, Fukushima meltdown, etc.);
- Contrasting backtesting results to expected shortfall and stress-test results.

After these first steps, which should be part of ongoing analysis, additional or complementary areas of analysis may need to be explored when models appear to be failing. Following up on the main leads from the practitioner forum and the guidelines provided by CSSF in its 2014 annual report[3], we investigated the following areas:

- Embedded assumptions: VaR models rely on a set of assumptions relating to, among others, market data consistency and the absence of intraday trading. These may bias outlier numbers.
- Reference period: VaR models are supposed to generate about as many exceptions as can be expected given their confidence level over a supposedly representative period (e.g. between 0 and 4 per year for a VaR 99% model). If the sample is not representative, then the relationship may not hold and we may observe a number of exceptions that seems incompatible with the model specifications.
- Clustering of exceptions: independence tests run by practitioners examine a specific form of dependence. They check that the occurrence of an exception is not too frequently followed directly by another one. Since dependence may take many forms, we need to qualitatively investigate the occurrence of several exceptions within a short time frame.
- Market swings and exceptional events: some market or company-specific events may generate returns that are beyond what the models are designed to capture. Models are designed to capture systematic and idiosyncratic risks as opposed to event risk[4]. Also, the CESR guidelines themselves state that VaR models are built to perform as expected in "normal" market conditions. By this standard, extreme market swings are unlikely to be captured.
- Model reactivity: the model needs to strike a balance between providing risk measures stable enough to be used to manage a portfolio and adapting to changing market conditions.
- Further thoughts on the joint use of multiple models.

It is worth noting that, as laid out in a previous paper "ABC of VaR Model Backtesting", too few exceptions should be considered as problematic as too many and investigated accordingly.

Our investigations were based on a database built with risk figures from multiple VaR models and returns from over 70 representative portfolios, covering a wide range of asset classes and strategies (equity, fixed income, multi-asset, emerging markets, derivative based, etc.) over two and a half years (01/01/14 to 30/06/16). The data underlying the analysis disclosed herein will not be made public or shared.

[3] The CSSF also highlights that the backtesting programme requirements described in box 18 of the CESR guidelines constitute, in accordance with point 4 of box 22 of the Guidelines, a minimum framework, which must be supplemented by other validation techniques. In this respect, the CSSF thinks, for example, of additional analyses on the number of overshootings observed over several confi dence intervals, the overshooting concentration or their amplitude or the abnormally low number of overshootings.

[4]ESMA 10-788 defines these notions as:

- Systematic risk = general market risk = risk of loss arising from changes in the general level of market prices.
- Idiosyncratic risk = risk that the value of a financial instrument changes more or less than the market in general (but not in an abrupt or sudden way).
- Event risk = risk that the value of a financial instrument changes in an abrupt or sudden way when compared with the behaviour of the general market and in a way that goes well beyond the normal range of fluctuations in value. Event risk covers, for instance, the migration risk for interest rate products or the risk of significant changes or jumps in equity prices.

Embedded Assumptions

VaR models: model families, strengths and weaknesses

VaR models are generally categorised by reference to two large families of models - analytical and simulation models:

- Analytical VaR models rely explicitly on a workable model. For instance, the delta-normal (DN) method and the (initial) RiskMetrics (RM) method assume that risk factors (mainly returns) are normally distributed. These models are often good at capturing changes in volatilities (heteroscedasticity) but less so at dealing with fat tailed-distributions (leptokurtic distribution).
- Simulation VaR models rely on the generation of series of returns:
  - The historical simulation (HS) method relies on a set of historical data;
  - Monte Carlo (MC) methods call upon stochastic models to randomly simulate many possible future returns.

  Although simulation VaR models overcome some of the limitations of parametric models, they are no panacea either. These models are often good at reflecting actual return distributions, but less so at reacting to market changes[5].

Although the limitations of the models have been thoroughly studied, publicized and are well-known by practitioners, it can be surprising to see how few of them perform reality checks on those implicit assumptions.

It is useful to shortly contrast stylised facts about daily returns with the assumptions underlying widely used models. Daily returns, routinely fed into our VaR models, tend to display the following behaviour:

- Little or no exploitable conditional mean predictability;
- The standard deviation greatly exceeds the mean, rendering the latter nearly irrelevant for VaR modelling;
- The variance appears to be time-varying (heteroscedasticity);
- The variance appears to be negatively correlated with returns (leverage effect);
- Not normally distributed as they tend to be negatively skewed and display (some) excess kurtosis (remember "fat tails");
- The returns are not continuous and exhibit jump components in their behaviour;
- Correlations between assets are time-varying and regime shifting.

Let's illustrate our point with relatively typical backtesting outcomes:

- Assume that the backtesting highlighted potential problems with the unconditional coverage of a VaR model that relies on a normality assumption of the returns. Here, further analysis should include checking whether the actual distribution is consistent with this assumption. This can be achieved in several ways, the simplest being to contrast on a same graph the actual distribution of returns with that predicted by the normal law. This quick but useful reality check ought to be further enhanced by statistical measures and tests. These could include comparing the higher order moments of the distribution (scaled 3rd and 4th order moments i.e. skewness and kurtosis), performing a visual inspection of the return's distribution (q-q plot) or implementing statistical goodness-of-fit tests (e.g. Pearsons Chi-square, Lilliefors, etc.).
- Assume that you are using a model based on historical simulation and the problem identified is related to independence of outliers. Basic (equal-weighted) forms of historical simulations respond slowly to changes in market regimes (volatilities and/or correlations). A sudden upward shift in the volatility regime will not be adequately reflected in the VaR derived from such a model. This could lead to clusters of outliers.

In such occurrences, it is widely acknowledged that the models will fail to capture important features of the observed returns and could potentially understate market risk.

[5] Beyond these two large family of models, other models have been built with the aim of addressing shortcomings of traditional (simulation and parametric) models. As example, let's mention the semi-parametric family of models which includes models that e.g. use historical returns but scale these by the ratio of current volatility to current volatility. This aims to go beyond the return normal distribution assumption and to make the model more reactive to market changes.

Tweaking the model

Among the possibilities to capture those changes and counteract a model's acknowledged weaknesses, tweaking the model is the most obvious. For instance:

- For an analytical model as referenced infra, when users identify the emergence of excess kurtosis and negative skewness, they could adjust the model by using parametric distributions with fatter and asymmetric tails (e.g. Normal Vs Student-t), approximations aiming to integrate higher moments of the distribution (e.g. Cornish-Fischer, see Zangari (1996)) or non-parametric empirical forms (see e.g. Barone-Adesi, Giannopoulos and Vosper, 1999).
- For a historical model as referenced infra, when users identify the emergence of a regime change such as shifting volatilities, they could try to make the model more reactive by e.g. shortening the historical series used or using some exponential weighting of returns (see Boudoukh et al. (1998)).

The challenge in such a temporary model modification is not so much the switch per se as the timing of it (and even more so, when to revert to 'normal'). We do not know of any recognized method to determine it. Indicators need to be developed and could include follow-up of market factors, correlations and rolling statistical tests on the properties that need to be captured (e.g. moments of distribution vs. normal distribution or changes in variance, etc.)

Valuation models (VaR vs. NAV), risk mapping and inputs

Another obvious step is to control the mechanics that lead to the generation of the data used for backtesting. As backtesting usually confronts actual P&L with VaR levels, it is important to either make sure that VaR & P&L are generated using similar assumptions, or at least to be aware of the differences. This straightforward insight could lead risk managers to compare and analyse the following:

- Valuation engines;
- Contractual and market data;
- Mapping to valuation/risk drivers.

Discrepancies in any of these could lead to bias in backtesting results. These need to be acknowledged and eventually controlled for.

As an example, let's take the case of a portfolio composed of a basket of equities and a Total Return Swap (TRS) swapping the return of this basket for that of an index plus a spread. Let's assume that the accounting department values the TRS based on an accrued valuation for NAV purposes, whereas the risk manager uses a full mark-to-market of the TRS for VaR computation. Given those valuation conventions, the NAV won't capture the offsetting relationship between the basket and the TRS whereas this relationship will be captured in risk figures. This discrepancy will probably lead to the NAV series being more volatile than that of the VaR, probably generating undue breaches/hits during the backtesting procedures. In such a case, the backtesting should be re-run after regeneration of one of the series based on congruent assumptions (i.e. rebuild VaR series by using accrued valuation for the TRS). Eventually, valuation models should be aligned. Such outliers could also result from differences in the market data used by accounting and risk management, or from different mapping processes. In order to identify candidates for investigation, risk managers need to look at risk & returns broken down at instrument (category) level.

Neutralising model assumptions

If reviewing embedded assumptions, mapping, valuation models and input data doesn't turn up material shortcomings, then practitioners might want to turn the question around by trying to prove that their models are not causing the high number of outliers.

Highlighting unusual market swings, although a first step, is not enough to prove that these were indeed the cause of the outlier(s). To firm up the causality assumption, a possible approach would be to neutralize the impacts relating to concerns arising from (i) the consistency between valuation and risk models and methods, (ii) model assumptions such as normality of risk factor returns, independence of returns, etc. and (iii) the impact of trading, as explained below.

Neutralising trading returns

VaR backtesting (and simple outlier counting) usually is "dirty": it compares yesterday's (forward looking) VaR estimate to the return realized today. This implicitly assumes no returns from trading. To neutralise this assumption, we may want to turn to clean backtesting which will measure the return of the portfolio had it not traded. By keeping the portfolio as was at the end of the previous day and measuring the returns of this hypothetical portfolio, the impact of trading on returns can be isolated.

Another approach that may be useful a posteriori and less resource demanding is to design a dummy portfolio representative of the investment strategy and management style of the analyzed portfolio, by combining the reference risk factors represented by key indices (volatility, equity, credit, interest rates, etc.). We then generate the returns and daily VaR measures of this portfolio over the analysis horizon (including rebalancing to align with the investment strategy).

Given that this portfolio does not depend on multiple and potentially complex valuation models (and related market data) and that trading can be neutralised, series can be expected to be clean from those biases.

Backtesting the returns of such portfolios against recomputed historical and parametric VaR series will yield a pattern of exceptions that can then be compared to that of the original portfolio (visually and/or with statistical tests).

On our sample of portfolios, the outlier profiles of the initial portfolios and the stylized ones were very close: the number of exceptions was similar, and these generally occurred on the same days.

Such an outcome tends to support the assumption that outliers are not caused by model idiosyncrasies such as assumptions, market data or trading. This, in turn, gives further credit to our initial result that outliers are due to exceptional market circumstances and events.

Reference Period

As highlighted in the first paper in our series, 'ABC of VaR model backtesting', VaR are point estimates which like any other statistics carry a confidence level and are based on a partial set of information. Accordingly, working with a one year reference period and a 99% VaR might have some drawbacks since the scarcity of observed exceptions (2-3 points expected in the tail) might not allow to draw a line between noise and inadequate VaR models To enhance the meaningfulness of tests, we may also test other cut off rates (e.g. 95%) and/or use a longer historical window where suitable.

This section explores the empirical impact of lengthening the historical window from one year to 2.5 years on our testing sample covering multiple VaR models and returns from over 70 representative portfolios, covering a wide range of asset classes and strategies (equity, fixed income, multi-asset, emerging markets, derivative based, etc.).

To test whether the periods under review are representative, we look at the number of portfolios having more than 4 exceptions (need analysis) and 6 exceptions (deemed at issue) per year. We do this for 2014, 2015, and H1 2016. When using our thresholds of 10 and 12 hits for the whole two-and-a-half-year period (645 observations) rescaled using the binomial distribution, we get the following results.

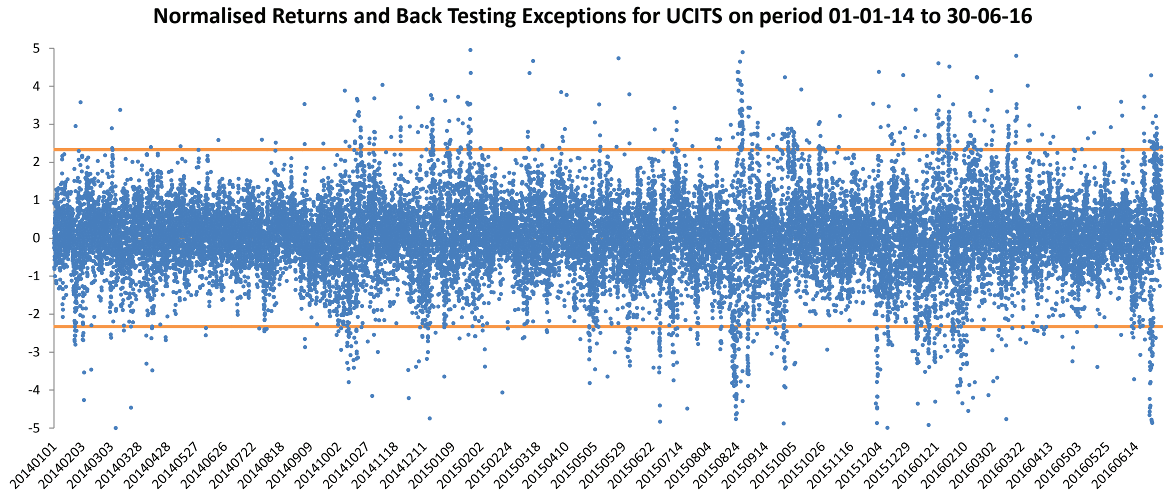
Nb Ptf	2014	2015	H1 2016	2014-2016
Ok	91%	48%	55%	62%
Need analysis	5%	26%	28%	19%
Deemed at issue	3%	26%	17%	19%
Nb ptf	100%	100%	100%	100%

From the table, it is obvious that 2014 was relatively calm whereas 2015 and especially H1 2016 saw more events and market conditions that have not been captured adequately by risk models.

Clustering of exceptions

The following graph plots the returns normalized by volatility (as deducted from the VaR assuming normally distributed returns) for all portfolios in our sample during the past two and a half years.

Any normalized return beyond the -2.33 line is a backtesting exception (for easy reading, we have truncated the graph to +/- 5 standard deviations).



First observations from this are that:

- Most exceptions are concentrated on or around specific dates;
- The days with high numbers of exceptions also feature rather extreme returns.

On this graph, we have also pulled in the information from the timeline, making the link between the severe market disruptions and the backtesting exceptions. Digging further into the events around the time where some of those extreme market events occurred, we can consider several consecutive days as constituting a single event. For instance:

- The fallout from the VW emission scandal hitting the news lasted 4 consecutive business days, from 25/09/2015 to 30/09/2015.
- The Chinese slowdown which was accompanied by heightened fears relating to the world economy and a rout in oil prices extends from 07/01/2016 until 22/01/2016 and hits 7 days with backtesting exceptions.

The list of these events and key market data at or around these are summed up in Appendix.

We may want to consider each of these episodes as single events instead of multiple exceptions. When re-running our analysis accounting for these clusters on the whole period, we construct the following table.

% Ptf	Initial	Cluster Impact	Post Cluster
Ok	68%	19%	87%
Need analysis	9%	-6%	3%
Deemed at issue	24%	-13%	10%
% ptf	100%	0%	100%

Instead of having 33% of portfolios beyond thresholds, we now have only 3% requiring further analysis and 10% deemed at issue. They may be various ways of interpreting the outcome of this clustering analysis, one of which further lends credit to our initial finding by showing that the higher than expected numbers of backtesting exceptions are attributable to specific market conditions and events which models have struggled and largely failed to capture. When performing this type of analysis, it is important to establish and document as clearly as possible the link between clustered exceptions, and to show that they are all attributable to the same common event. Causality can be established, for instance, by keeping copies of the daily market commentary published by financial news outlets. In this case, the same sources must consistently be used to avoid any suspicion of cherry-picking. It is just as important to take a conservative/cautious stance when running the analysis that may lead to grouping or discarding observations. If the exceptions at the back end of a cluster are arguably attributable to another market or company event than the first exceptions in a cluster, then they should not be grouped together.

Market swings & exceptional events

Models are not built to deal with extreme events. Such events should be considered as part of stress testing, rather than ongoing risk measurement. Therefore, when such episodes multiply beyond expectations, counting such items as exceptions may distort the backtesting results. In this analysis, we neutralize the impact of market swings / extreme events resulting in conditions beyond model parameters.

To define such occurrences in a conservative way, we control for the size of swings normalized by volatility and derive the probability of such events from a Gaussian distribution.

We isolate the 7 clusters previously identified (in red) and add 3 additional extreme market moves: 1-day events that may (e.g. ECB disappointment) or may not be attached to a specific announcement or event. Quite interestingly, 7% of days in the analysis account for 65% of exceptions.

Period	Event Name	nb days	nb exceptions	Avg size of move (StDev)	Probability	Max size of move (StDev)	Probability
09.10.14 to 17.10.14	Tech & energy drawdown	6	5%	2.96	0.16%	3.14	0.08%
29/04/2015	EUR yield curve steepening and eq drawdown	1	2%	2.79	0.26%	2.79	0.26%
02.06.15 to 03.06.15	Greek saga	2	3%	2.79	0.26%	2.85	0.22%
29/06/2015	Eq mkts down 2 to 4% with vols up 4%	1	4%	3.01	0.13%	3.01	0.13%
19.08.15 to 01.09.15	China/Black Monday/Glencore	8	19%	3.30	0.05%	4.53	0.00%
25.09.15 to 30.09.15	VW emission scandal	4	4%	3.01	0.13%	3.47	0.03%
03/12/2015	ECB pro inflation package disappoints	1	3%	3.52	0.02%	3.52	0.02%
07.01.16 to 22.01.16	China slowdown/world economy jitters/petrol rout	14	10%	2.86	0.21%	3.26	0.06%
02.02.16 to 12.02.16	Rate hike creep/world economy jitters/petrol rout	9	8%	3.31	0.05%	5.48	<0.0000%
24.06.16 to 27.06.16	Brexit vote	2	7%	6.64	<0.0000%	7.71	<0.0000%
	Cluster	7%	65%	3.2	0.07%	/	/

The magnitude of market movements on any of those days is already very large (often > 3 standard deviations and up to nearly 8) and hence the probability of those very low. But what is even more impressive is that those clusters are characterized by such events occurring day after day for the whole duration of the cluster. This pushes the probability of such events to exceptionally low levels (further undermining the assumption of i.i.d. normally distributed returns). Take, for instance, the impact of the VW emission scandal: a move of 3 standard deviations or more is supposed to take place 0.22% of the time; the probability of such events taking place 3 days in a row is about a millionth of a percent.

When re-running our analysis accounting for these clusters and other events on the whole period, we construct the following table.

% Ptf	Initial	Cluster Impact	Event Impact	Final
Ok	68%	+19%	+7%	94%
Need analysis	9%	-6%	-1%	1%
Deemed at issue	24%	-13%	-6%	4%
% ptf	100%	0%	0%	100%

The total number of exceptions drops from 1.20% of observations initially to 0.9% after the clustering analysis and to 0.47% after accounting for other events. This is to be compared to expected exceptions of 1%.

Instead of having 29% of portfolios beyond thresholds, we now have only 1% requiring further analysis and 3% deemed at issue (again taking 10 and 15 as our thresholds for the two and a half year-period).

The event analysis therefore could be seen as lending further credit to the initial finding, showing that the higher than expected number of backtesting exceptions could be attributable to specific market conditions and events.

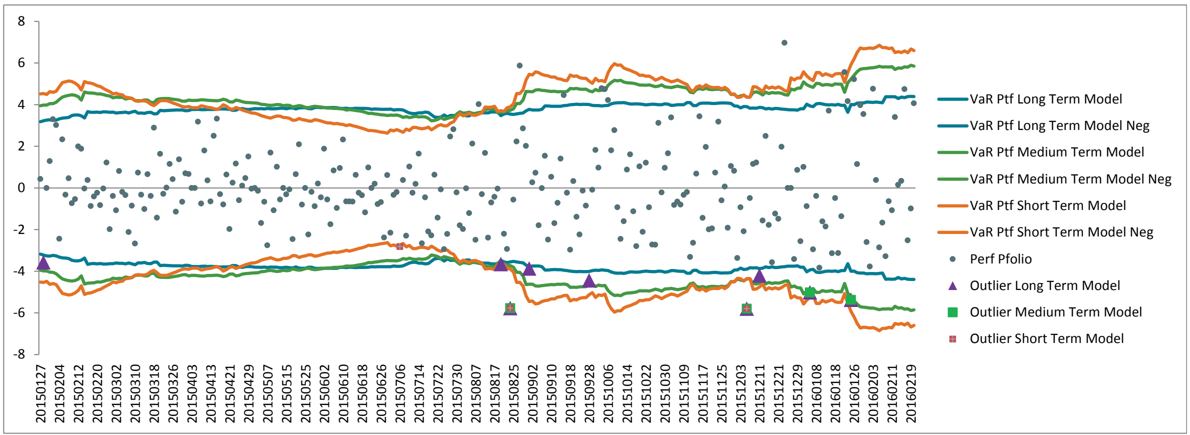
One word of caution: CESR guidelines 10-788 define the VaR as "the maximum potential loss at a given confidence level (probability) over a specific time period under normal market conditions". But the burden of proof is on the VaR user here: any claim that a specific backtesting exception is attributable to an event beyond normal market conditions must be substantiated with facts and numbers:

- To establish causality, similar documentation as when dealing with an exception cluster (as described above) needs to be gathered and

- To establish amplitude, it is advisable to set up a dashboard with key risk factor metrics that can easily be updated on the fly and post facto for any given day or period. This monitor could include e.g. (i) key risk factor returns, (ii) a comparison with historical returns and/or volatilities and (iii) implicit probabilities of such moves, given the volatility measured by the model just before the event and assuming normally distributed returns.

Model Reactivity

A "good" risk model needs to strike a balance between providing risk measures stable enough to be incorporated (or at least considered) in the portfolio management process, and reacting to market developments. To illustrate the difficulty in doing so, the following graph compares different VaR measures with the returns from one of the portfolios in our sample, from January 27, 2015 to June 29, 2016.



We notice that:

- Exceptions appear in clusters;
- Every VaR model represented here reacts to new data, though to varying degrees;
- These reactions might be deemed too gradual to account for fast-changing market volatility.

The means to achieve higher reactivity depend on the type of VaR model:

- For parametric models and Monte-Carlo simulations, the decay factor used to estimate volatility using EWMA or GARCH can be altered to lower the half-life of the measure;
- For historical models, the observation period can be shortened or the weighting can be tilted towards giving more relevance to more recent observations.

When testing models with higher reactivity, we should look for models that would improve the backtesting situation and still strike an acceptable balance between

- Stability: the VaR should not overestimate risk due to non-persistent market shocks (jumps) in order to remain within regulatory and contractual limits such as the 20% limit for the 20-day VaR 99% required (with possibility of rescaling) for UCITS monitored using absolute VaR;
- Reactivity; on this note we tested whether shorter term models would allow to better (i) capture changing market volatility, (ii) reduce the amount of exceptions in clusters and (iii) materially reduce the amount of exceptions overall.

Taking our representative portfolio as an illustration, we see that shorter-term models:

- React faster to changes in volatility whether in upward or downward trends;
- Capture more outliers when volatility increases rapidly but generate outliers when volatility moves down;
- May generate VaR markedly higher than in the standard model (and eventually breach limits).

Looking at our sample, we reach the following observations:

- Overall, the amount of exceptions decreases by 15% to 25%, depending on which short-term model is used.
- Short-term models trade off substantially more volatility in the risk measure against an outlier capture improved by 15-25%, through a better reactivity to market conditions;
- Improvement is not across the board: some portfolios are worse off as better reactivity also creates previously non-existent outliers.

Further Thoughts

Other routes were explored, such as running in parallel:

- Long terms and short term models;
- Models using different methodologies (historical vs. parametric vs. Monte Carlo, or factor-based vs. full revaluation) models;

and combining those different measures into a single measure or switching to an alternative better suited model when conditions warrant it.

Such proposals raised various issues such as:

- How to define the reference in what case?
- How to switch back and forth between models?

while managing governance, risks arising from moral hazard, and the temptation of regulatory arbitrage.

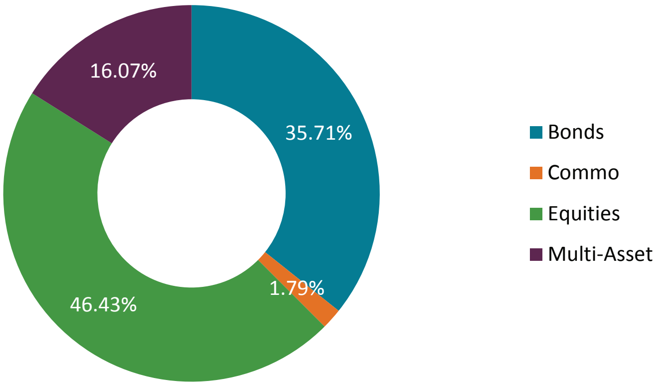
Practitioners rely on risk models that are often at the heart of decision-making and are instrumental to monitoring and balancing risk and performance. Faced with a risk model that seems to be failing regular backtests and analysis, they should proceed with caution: before implementing changes that may have drastic effects on risk metrics, portfolio management and potentially fund distribution, thorough analysis needs to be undertaken in order to document whether the model is effectively failing or is being used in conditions that it simply wasn't designed to accommodate. We hope this paper, in conjunction with the "ABC of VaR Model Backtesting" paper it builds upon, will help practitioners in this task.

CSSF flags backtesting as one area which Boards of Directors need to be informed about on a regular basis. It is easy to see why: over the last decades, VaR and other risk metrics have transitioned from mere statistical footnotes to a critical input in daily portfolio management, strategy definition, and even incentive structures. The more embedded they are in a fund's management structure, the more important their credibility becomes. This makes their backtesting not only a Board-level issue, but one where governance arrangements more generally need to be well thought out. The closing paper in our series will examine this topic.

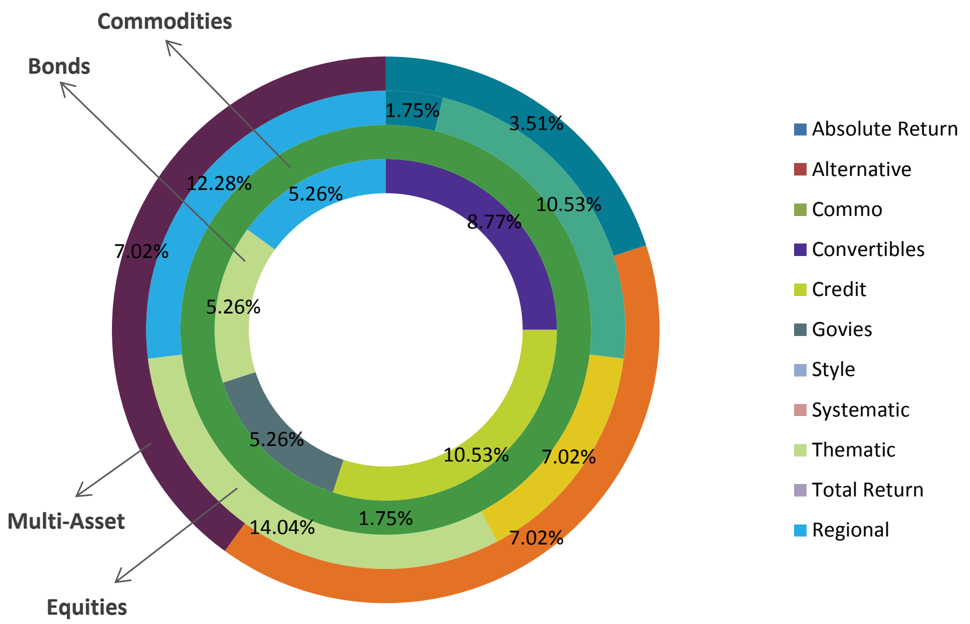
To conclude, we would like to thank all participants for the quality of their input and their candor during the exchanges leading up to this paper. We hope it will help overcome the "too technical" stigma attached to backtesting, and spur more widespread sharing of ideas about how to deal with issues in this area.

Annex 1 - Description of sample data set

	Bonds	Commo	Equities	Multi-Asset	Grand Total
Global	17.86%	1.79%	12.50%	16.07%	48.21%
Regional	17.86%	0.00%	33.93%	0.00%	51.79%
Grand Total	35.71%	1.79%	46.43%	16.07%	100.00%



	Bonds	Commo	Equities	Multi-Asset	Grand Total
Absolute Return			1.75%	3.51%	5.26%
Alternative				7.02%	7.02%
Commo		1.75%			1.75%
Convertibles	8.77%				8.77%
Credit	10.53%				8.77%
Govies	5.26%				5.26%
Style			10.53%		10.53%
Systematic			7.02%		7.02%
Thematic	5.26%		14.04%		19.30%
Total Return				7.02%	7.02%
Regional	5.26%		12.28%		17.54%
Grand Total	35.09%	1.75%	45.61%	17.54%	100.00%



Annex 2 - Market Impact of Extreme Market Swings & Events (incl. clusters)

	09/10/2014 – 17/10/2014 Tech & energy drawdown	29/04/2015 – 29/04/2015 EUR yield curve steepening and eq drawdown	02/06/2015 – 03/06/2015 Greek saga	29/06/2015 – 29/06/2015 Eq mkts down 2 to 4% with vols up 4%	19/08/2015 – 01/09/2015 China / Black Monday / Glencore	22/09/2015 – 30/09/2015 VW emission scandal	03/12/2015 – 03/12/2015 ECB pro inflation package disappoints	07/01/2016 – 22/01/2016 China / world economy jitters / petrol rout	02/02/2016 – 12/02/2016 Rate hike creep / world economy jitters / petrol rout	24/06/2016 – 27/06/2016 Brexit vote
S&P 500 INDEX	-4.17%	-0.37%	0.11%	-2.09%	-8.73%	-2.39%	-1.44%	-4.19%	-3.85%	-5.34%
NASDAQ COMPOSITE INDEX	-4.70%	-0.63%	0.32%	-2.40%	-8.37%	-4.32%	-1.67%	-5.06%	-6.12%	-6.43%
NYSE Arca Oil	-6.63%	0.21%	0.01%	-2.35%	-8.68%	-2.55%	-1.55%	-5.60%	-4.17%	-7.46%
Euro Stoxx 50 Pr	-2.98%	-2.65%	0.25%	-4.21%	-8.77%	-2.64%	-3.61%	-3.70%	-8.77%	-11.21%
SPI SWISS PERFORMANCE IX	-3.05%	-1.50%	-0.11%	-1.50%	-7.71%	-3.12%	-1.64%	-3.97%	-7.72%	-5.52%
MSCI EM	-2.20%	-0.70%	-0.71%	-2.16%	-5.54%	-2.89%	-0.31%	-6.47%	-4.24%	-4.89%
NIKKEI 225	-6.82%	-2.69%	-0.47%	-2.88%	-11.62%	-3.77%	0.01%	-6.78%	-16.30%	-5.72%
HANG SENG INDEX	-1.03%	-0.15%	0.22%	-2.61%	-9.75%	-4.19%	-0.28%	-9.06%	-6.51%	-3.07%
CBOE SPX VOLATILITY INDX	6.88	0.98	-0.31	4.83	17.61	4.36	2.20	1.75	5.42	6.60
VSTOXX Index	4.45	1.93	-0.51	3.93	11.76	5.03	-0.16	2.03	8.35	2.94
US ULTRA BOND CBT Jun17	-0.10	0.07	0.15	-0.14	0.07	-0.09	0.17	-0.12	-0.22	-0.29
US 10YR NOTE (CBT)Jun17	-0.13	0.04	0.13	-0.12	-0.04	-0.07	0.11	-0.15	-0.18	-0.24
US 2YR NOTE (CBT) Mar17	-0.09	0.01	0.04	-0.06	-0.03	-0.09	0.01	-0.16	-0.12	-0.17
EURO BUXL 30Y BND Jun17	-0.05	0.13	0.32	-0.11	0.14	-0.09	0.11	-0.10	-0.14	-0.30
EURO-BUND FUTURE Jun17	-0.04	0.12	0.32	-0.13	0.15	-0.10	0.18	-0.12	-0.10	-0.20
EURO-SCHATZ FUT Jun17	0.01	0.03	0.03	-0.03	0.04	-0.02	0.13	-0.07	-0.03	-0.08
LONG GILT FUTURE Mar17	-0.06	0.13	0.21	0.04	-0.07	-0.03	0.11	-0.08	-0.22	-0.36
SHORT GILT FUTURE Mar17	-0.05	0.04	0.08	-0.32	-0.05	0.28	0.05	-0.07	-0.02	-0.28
MARKIT CDX.NA.IG.28 06/22	3.45	1.53	1.44	4.31	6.66	12.35	1.27	11.39	17.94	14.93
MARKIT CDX.NA.HY.28 06/22	-1.26	-0.17	-0.30	-0.98	-0.85	-4.93	-0.37	-1.18	-2.46	-2.44
MARKIT ITRX EUROPE 06/22	3.57	1.50	-2.38	10.04	7.26	14.64	2.67	12.02	25.79	22.96
MARKIT ITRX EUR XOVER 06/22	24.98	9.41	-6.09	45.14	27.22	66.74	10.78	36.64	90.86	92.27
MARKIT CDX.EM.27 06/22	0.13	-0.08	0.01	-0.33	-0.46	-1.73	-0.54	-0.81	-0.88	-1.18
BRENT CRUDE FUTR May17	-4.27%	1.12%	-0.49%	-1.30%	0.21%	-1.46%	2.05%	-8.95%	-3.07%	-7.44%
Gold Spot $/Oz	1.41%	-0.62%	-0.34%	0.37%	2.02%	-1.62%	0.80%	0.39%	9.71%	5.39%
EUR-USD X-RATE	0.21%	1.34%	3.18%	0.62%	2.64%	-0.12%	3.06%	0.14%	3.38%	-3.16%
GBP-USD X-RATE	-0.47%	0.65%	0.92%	-0.05%	-2.28%	-2.44%	1.29%	-2.49%	0.48%	-11.10%
Japanese Yen Spot	-1.11%	0.13%	-0.42%	-1.06%	-4.05%	-0.56%	-0.51%	0.26%	-6.40%	-3.92%

If and when backtesting turn out results that could call into question the appropriateness of VaR models used, risk managers and conducting persons might want to run further analysis before jumping to conclusions. The following table may be used as a sanity check to ensure relevant analysis has been carried out.

Investigation Area	Tool	Explanation
Clusters & Events	Graph of normalized returns	Normalize returns by standard deviations and plot these against the 2.33 standard deviation (99% confidence level) will allow to visually identify clusters and events. Such graphs are useful both at portfolio level and at fund range level (see e.g. 'Clustering of Exceptions' in Section II).
Clusters & Events	Event Timeline	Keep a timeline of events that have materially impacted markets. This may be fed by the ongoing backtesting analysis (daily and/or monthly).
Events	Market Monitor	Keep a timeline of events that have materially impacted markets. This may be fed by the ongoing backtesting analysis (daily and/or monthly).
Cluster & Events	Clustering Analysis, Event Risk Analysis	The impact of clusters and beyond model events ought to be neutralized in backtest analysis. Clusters of exceptions that clearly and non-ambiguously tie back to a unique market event may be considered as a unique hit instead of a string of hits. Similarly, events that are beyond model parameters may perhaps be excluded from analysis and analysis re-run with the adapted number of exceptions and of total observations. One important caveat: analysis and documentation need to establish as clearly as possible the link between exceptions in the case of a cluster and the causality between the event and the exception in the case of event risk. It is important to take a conservative/cautious stance when running the analysis that may lead to grouping or discarding observations.
Model Parameter – Reference Period & Confidence Level	Control for reference period & confidence level	Using the regulatory standard of 1Y reference period and a 99% confidence level make for poorly specified distribution tails containing 2 to 3 observations. Any slightly unusual period may rapidly show more than 4 (need analysis) or even 6 (deemed at issue) hits, casting doubt on the validity of the VaR model. Maybe unduly so. Testing longer reference periods and/or lower confidence levels allows for better specification of distribution tails as these will contain more observations. 2 caveats: (i) reference period(s) used needs to be representative of current conditions/set-up e.g. model can't have been materially amended and (ii) confidence level shouldn't be lower than 95% (derives from regulatory requirement). Conversely, in the face of a period of increased volatility, the risk manager might want to use a shorter reference period which would better represent current market conditions and allow improved VaR reactivity. Possibly in conjunction with a lower confidence level in order to maintain robust distribution tail specifications.

Investigation Area	Tool	Explanation
Model Assumption – Distribution	Control for distribution	For models relying on specific distribution assumptions (e.g. normal distribution of risk factor returns), controlling that (i) the assumption remains broadly valid (e.g. graphically or statistical tests) and (ii) backtest results are not unduly dependent on the assumption (e.g. contrast with historical or Monte Carlo VaR or, where possible, use alternative distributions such as Student) are useful checks & balances.
Model Assumption – Reactivity	Control for increased reactivity	Control graphically that models are not reacting too slowly to market changes: this might be the case if VaR doesn't adapt to market changes with each event followed by several outliers while the model adapts. Testing more reactive models and contrasting backtesting results may provide useful insights. Increased reactivity may be achieved in various ways according to the kind of model used. Without being exhaustive, possibilities include shortening the reference period, using exponential weighting, shortening half-lives / increasing decay factor, using shorter holding periods, using shorter term models, allowing for overlapping observation periods, etc.
Model Assumption – No Trading	Clean(er) backtesting	'Dirty' backtesting compares yesterday's (forward looking) VaR estimate to the return realized today which implicitly assumes no returns from trading. Trading may generate P&L that can either add or hide outliers. To neutralise this assumption, we may want to turn to 'clean' backtesting which compares returns had the portfolio not traded to risk figures. 'Clean' backtesting may not always be an option given its computational complexity. Hence the need for 'cleaner' or 'less dirty' approaches such as designing a dummy portfolio representative of the investment strategy and management style of the analyzed portfolio (e.g. by combining the reference risk factors represented by key indices and rebalancing features). The backtest results of such pro-forma portfolios can then be used as a control variable to the original portfolio results with similar results pointing to the lack of impact from trading.
Mapping – market & static data, instrument modelling, P&L engines	Risk & return breakdown by asset class, by instrument	For portfolios with unexpected levels of outliers, it may be important to take a more granular look at the numbers and drill down to asset class or even instrument level. By doing this, it is easier to identify whether a specific (kind of) instrument is generating the outlier. If it is the case, having a closer look at mapping, modelling and alignment of P&L engines (NAV or performance Vs VaR).

luxembourg fund industry

The Association of the Luxembourg Fund Industry (ALFI), the representative body for the Luxembourg investment fund community, was founded in 1988. Today it represents more than 1,500 Luxembourg-domiciled investment funds, asset management companies and a wide v ariety of service providers including depositary banks, fund administrators, transfer agents, distributors, law firms, consultants, tax advisers, auditors and accountants, specialist IT providers and communications agencies.

Luxembourg is the largest fund domicile in Europe and its investment fund industry is a worldwide leader in cross-border fund distribution. Luxembourg-domiciled investment structures are distributed in more than 70 countries around the globe, with a particular focus on Europe, Asia, Latin America and the Middle East.

ALFI defines its mission as to "Lead industry efforts to make Luxembourg the most attractive international investment fund centre".

Its main objectives are to:

Help members capitalise on industry trends
ALFI's many technical committees and working groups constantly review and analyse developments worldwide, as well as legal and regulatory changes in Luxembourg, the EU and beyond to identify threats and opportunities for the Luxembourg fund industry.

Shape regulation
An up-to-date, innovative legal and fiscal environment is critical to defend and improve Luxembourg's competitive position as a centre for the domiciliation, administration and distribution of investment funds. Strong relationships with regulatory authorities, the government and the legislative body enable ALFI to make an effective contribution to decision-making through relevant input for changes to the regulatory framework, the implementation of European directives and the regulation of new products or services.

Foster dedication to professional standards, integrity and quality
Investor trust is essential for success in collective investment services and ALFI thus does all it can to promote high professional standards, quality products and services, and integrity. Action in this area includes organizing training at all levels, defining codes of conduct, transparency and good corporate governance and supporting initiatives to combat money laundering.

Promote the Luxembourg investment fund industry
ALFI actively promotes the Luxembourg investment fund industry, its products and services. It represents the sector in financial and economic missions organised by the Luxembourg government around the world and takes an active part in meetings of the global fund industry.

ALFI is an active member of the European Fund and Asset Management Association, of the International Investment Funds Association, of Pensions Europe, of the International Association of Pension Funds Administrators and of the Global Impact Investing Network.

For more information, visit our website at www.alfi.lu and follow ALFI on

May 2017