INSTITUTO
UNIVERSITÁRIO
DE LISBOA

UNIVERSIDADE
DE LISBOA

Backtesting Expected Shortfall: Historical Simulation Based Analysis of a US Diversified Portfolio

Patrícia Alexandra Sobreda da Cruz

Master in Financial Mathematics

Supervisor:
PhD, Paulo Viegas de Carvalho, Assistant Professor, ISCTE-IUL

September, 2025

BUSINESS
SCHOOL

Ciências ULisboa | Departamento de Matemática

Department of Finance

Department of Mathematics

Backtesting Expected Shortfall: Historical Simulation Based Analysis of a US Diversified Portfolio

Patrícia Alexandra Sobreda da Cruz

Master in Financial Mathematics

Supervisor:
PhD, Paulo Viegas de Carvalho, Assistant Professor, ISCTE-IUL

September, 2025

Acknowledgment

First and foremost, I would like to express my gratitude to my supervisor, Professor Paulo Viegas de Carvalho, for his guidance and support throughout the development of this thesis.

On a personal note, I am profoundly thankful to my family for always believing in me and giving me the motivation to continue, in particular, to my parents for their patience and constant support.

Finally, I would also like to extend my gratitude to my friends and colleagues, whose encouragement fueled my determination throughout this academic journey, and to all those who contributed to the completion of this thesis.

Resumo

A transição de *Value-at-Risk* para *Expected Shortfall*, conforme estabelecido pelo Comité da Basileia, no *Fundamental Review of the Trading Book*, gerou um debate significativo relativamente à possibilidade de fazer *backtesting* à medida recém-adotada. Como tal, esta dissertação avalia ambas as métricas comparativamente, testando a primeira através do *Traffic Light Test* e do teste de Kupiec, e a segunda através dos testes de Pearson e de Nass. A *Expected Shortfall* é calculada frequentemente através de Simulação Histórica, uma abordagem não paramétrica simples que, contudo, depende fortemente do tamanho da amostra. Para avaliar a sua precisão, são aplicados dois modelos baseados nessa estrutura: a Simulação Histórica padrão e um método ponderado através da idade de cada observação, denominado Simulação Histórica Ponderada. Ademais, esta tese fornece uma análise aprofundada do desempenho metodológico durante períodos de stresse financeiro, particularmente o vivido nos Estados Unidos da América durante o primeiro semestre de 2025, devido às tarifas aplicadas pelo Presidente Donald Trump. Para esse propósito, utiliza-se uma carteira diversificada composta por ações do mercado financeiro americano. Os resultados mostram que, embora ambos enfrentem desafios em períodos de instabilidade financeira como o supramencionado, a Simulação Histórica Ponderada supera a abordagem padrão. Este estudo conclui que, entre as metodologias testadas, o teste de Nass com 8 níveis é o mais coerente para a *Expected Shortfall*, equilibrando a precisão estatística com a prudência, fornecendo evidências que validam a possibilidade de fazer *backtesting* a essa medida e destacam a importância de testá-la diretamente.

Palavras-Chave: Expected Shortfall, Simulação Histórica, Simulação Histórica Ponderada, Value-at-Risk, Backtesting.

Classificação JEL: C52, G17.

Abstract

The transition from Value-at-Risk to Expected Shortfall, under the Basel Committee's Fundamental Review of the Trading Book, sparked significant debate regarding the backtestability of the newly adopted measure. In response, this dissertation assesses both measures comparatively, backtesting the former through the Traffic Light and the Kupiec tests and the latter through the Pearson and Nass tests. Expected Shortfall is often computed using Historical Simulation, a user-friendly non-parametric approach that, however, strongly depends on the sample size. To assess its accuracy, two models based on this framework are examined: the standard Historical Simulation method and an age-weighted method, called Weighed Historical Simulation. Moreover, this thesis provides a deeper analysis on methodological performance during financial stress periods, particularly the one lived in the United States of America during the first semester of 2025, due to the tariffs applied by President Donald Trump. Accordingly, a diversified portfolio composed of equities from the American financial market is used. The results show that, even though both models face significant challenges during periods of financial instability such as the one mentioned previously, the Weighted Historical Simulation outperforms the standard approach. This study concludes that, among the tested methodologies, the Nass test with 8 levels is the most coherent backtest for Expected Shortfall, balancing statistical accuracy with prudentiality, therefore providing evidence supporting the backtestability of this measure and highlighting the importance of backtesting it directly.

Keywords: Expected Shortfall, Historical Simulation, Weighted Historical Simulation, Value-at-Risk, Backtesting.

JEL Classification: C52, G17.

Contents

List of Tables

List of Figures

List of Acronyms

A - Amber Zone
B - Basel Traffic Light test
BCBS - Basel Committee on Banking Supervision
CVaR - Conditional Value-at-Risk
c.d.f. - Cumulative Distribution Function
ES - Expected Shortfall
EWMA - Exponentially Weighted Moving Average
FHS - Filtered Historical Simulation
FRTB - Fundamental Review of the Trading Book
G - Green Zone
GARCH - Generalized Autoregressive Conditional Heteroskedasticity
HS - Historical Simulation
i.i.d. - Independent and Identically Distributed
K - Kupiec Unconditional Coverage Test
LRT - Likelihood Ratio Test
MN - Multinomial Distribution
$\mathbf{N}_N$ - Nass test for N levels
P&L - Profit and Loss
p.d.f. - Probability Density Function
$\mathbf{P}_N$ - Pearson test for N levels
POF - Proportion of Failures
R - Red Zone
US - United States
USD - United States Dollar
VaR - Value-at-Risk
WHS - Weighted Historical Simulation

CHAPTER 1

Introduction

Measuring and assessing financial risk effectively is crucial for both financial institutions and regulatory authorities. Consequently, relying on an appropriate metric is considered essential.

Over the past few decades, Value-at-Risk (VaR) has been established as the standard risk measure to quantify market risk. By definition, it estimates the maximum potential loss within a specific time period and confidence level. However, despite its intuitive appeal and acceptance under previous frameworks, VaR has been heavily criticized, mainly because it does not capture the extent of losses that exceed the defined threshold.

When faced with extreme events such as financial crises, relying on a measure that does not provide the entire scope of the major losses can be especially devastating. For that reason, in 2016, the Basel Committee on Banking Supervision (BCBS) published the Fundamental Review of the Trading Book (FRTB), proposing the replacement of VaR with Expected Shortfall (ES) to determine regulatory capital requirements.

On the one hand, ES provides a more comprehensive assessment of tail risk by averaging the returns of the distribution that are worse than the VaR cutoff point, thereby capturing extreme losses that other measures may overlook. On the other hand, backtesting this measure is less straightforward, which led to the FRTB's recommendation of still performing backtests based on the VaR estimate. As a result of this regulatory inconsistency between risk measurement and model validation, a new field of research emerged.

As there remains no universally accepted method for backtesting ES, a simple, model-free and often used forecasting procedure is the Historical Simulation (HS). As its name indicates, this approach uses historical data to predict future events, thereby assuming that history repeats itself, both when stable and when affected by stressful events.

Because HS does not impose distributional assumptions, this model is more apt to capture features of market stress periods, which is when the performance of parametric models suffers the most. This is a relevant advantage, particularly when dealing with the uncertainty of the United States (US) market lived in the first semester of 2025, due to the tariffs applied by President Donald Trump[1].

Given these circumstances, this study aims to address several topics. By comparing VaR and ES, one of the main goals is to address the relevance and backtestability of the latter over the former. Furthermore, this thesis considers and evaluates two HS estimation

[1]Donald Trump was elected, for the second time, as President of the United States of America in November, 2024, but only started his second term on January 20, 2025, time when the commitments to impose tariffs and taxes on foreign countries resurfaced.

approaches, with the primary objective of finding another simple yet more accurate model to estimate the risk measures in use. Finally, another goal is to provide regulators and practitioners with practical insights on how the main methodological approaches used to model and backtest ES perform under the chosen framework. This analysis focuses on the periods of instability lived in the US since 2013, especially the one lived during the first semester of 2025.

In pursuit of those objectives, I estimate VaR and ES using two estimation models over the same diversified US equity portfolio. The methods applied are the standard HS and an extension of if, called Weighted Historical Simulation (WHS), which is computed using two decay parameters for further refinement of the comparison study. Following this, I conduct a quality assessment of these risk models by backtesting each measure using practical and accessible approaches: the FRTB's method and the Unconditional Coverage test (Kupiec, 1995) for VaR and the Pearson and the Nass tests for ES.

The results of this process emphasize the importance of backtesting ES and highlight the performance of a simple model that demonstrates a better performance under stress conditions, WHS. Under the chosen framework, the multinomial backtests are, overall, less prudent than the FRTB Traffic Light test. Nevertheless, the Nass test with 8 levels seems to provide the most accurate results, while combining model-free implementation with rigorous statistical foundations. Furthermore, the outcomes of this test align with those produced by the Basel test, both when market conditions are stable and when they vary.

Considering that, this dissertation is structured as follows. Chapter 3 introduces the risk measures and their properties. Chapter 4 discusses the portfolio in use, as well as the proposed backtesting procedures taken into consideration in this study, particularly in the HS framework. Chapter 5 presents the empirical results and provides a comparative analysis. Finally, Chapter 6 provides the conclusions of the study, presenting a discussion of the main findings and implications, as well as suggestions for future research.

CHAPTER 2

Literature Review

Financial institutions need to have robust capital reserves in order to protect their overall stability. However, finding the right capital amount to hold is a challenging task because neither the underestimation nor the overestimation of risk results in a positive outcome: insolvency might be the consequence of dealing with the former, and even though it is not harmful to society, if the case is the latter, it leads to the costly hold of extra capital.

For the reasons specified before, an extensive body of research has been devoted to the study of different risk measures.

2.1. The limitations of Value-at-Risk

Since the mid-1990s, Value-at-Risk has become the standard measure for market risk (Jorion, 2006). That being said, it has also received extensive criticism.

As a risk measure, VaR fails to satisfy a key property of coherent risk measures, called subadditivity (Artzner et al., 1999). This leads to limitations when dealing with diversification as well as aggregation problems, meaning that, even if independent, the risk of a portfolio can be larger than the sum of its components' risks, when measured by VaR (Barone-Adesi and Giannopoulos, 2001).

Addressing these limitations, an alternative measure, called Expected Shortfall, was proposed and then found to be superior to the former, not only for its sensitivity to extreme events but also for meeting the required coherence properties, as shown by Acerbi and Tasche (2002).

2.2. The regulatory shift from VaR to ES

As noted previously, VaR was an unreliable measure, predisposed to fail when faced with periods of high uncertainty. Precisely, that is what happened during the financial crisis of 2007-2008.

To address the shortcomings of the framework in place at the time, the Basel Committee introduced, and then published, the Fundamental Review of the Trading Book, replacing VaR, with a confidence level of 99%, for ES, with a confidence level of 97.5% (BCBS, 2013, 2019).

By using ES as the primary risk measure for internal models of capital adequacy, banking institutions would account for tail risk, thereby ensuring more conservative financial reserves. However, despite its advantages, there were obstacles that led to the decision of relying on a different measure, VaR, for backtesting procedures, therefore undermining the consistency of the regulatory framework (Acerbi and Szekely, 2014; Costanzino and Curran, 2018).

As a result of this inconsistency, the development of a simple, yet effective, backtesting procedure for ES emerged as a central topic of investigation within the risk management literature.

2.3. The evolution in backtesting Expected Shortfall

Regarding risk model validation, backtestability is not a negotiable property, as this process consists of comparing the forecasts with the observed returns, providing a way of analysing the accuracy of risk measures. Because VaR is a quantile, backtesting it is not a problem, consisting of a straightforward process of counting the number of exceptions, that is, exceedances of the VaR threshold (Kupiec, 1995). In contrast, for ES that is not the case, as it is the average of the values in that quantile, instead of the quantile itself.

That being said, one of the major difficulties in backtesting ES arises from the fact that it lacks a mathematical property called elicitability, which means that there is no scoring function that can validate the forecasts on its own (Gneiting, 2011). However, this makes backtesting harder, not impossible. In fact, researchers even argue that backesting it is not more difficult than backtesting VaR (Acerbi and Szekely, 2014; Du and Escanciano, 2017).

Following this paradigm shift, recent academic contributions have made backtesting ES a viable task. Notably, Acerbi and Szekely (2014) introduced three model-independent, non-parametric ES backtests, shown to be more powerful than VaR backtests, yet they rely on simulation techniques.

Stemming from the concept of joint elicitability between VaR and ES, first introduced by the authors mentioned before and later explored by Fissler and Ziegel (2016), researchers were able to uncover new backtesting methodologies to jointly evaluate the forecasts of these two measures. Using this insight, some authors explored comparative backtesting approaches (Fissler et al., 2016; Nolde and Ziegel, 2017; Patton, Ziegel and Chen, 2019). Still, these methods are suitable for comparison purposes, that is, model selection and not model validation.

Another approach that led to several backtesting adaptations was the approximation of ES using VaR. Acerbi and Tasche (2002) first described ES as an average of VaRs at different confidence levels, and later, Emmer et al. (2015) proposed approximating the integral with a Riemann sum at four significance levels. Based on the second idea, Kratz et al. (2018) proposed a multinomial coverage test that extended it to any number of quantiles; however, the results indicated that four or eight levels improved performance.

Additional developments to the backtesting framework include a regression-based approach (Bayer and Dimitriadis, 2022), conditional and unconditional coverage tests, based on cumulative violations (Du and Escanciano, 2017) and a non-parametric method that consists of counting the worst realizations for the secured position (estimated capital reserve and realized cash-flow) that add up to a negative total (Moldenhauer and Pitera, 2019). Nevertheless, although extensive research has been done on this topic, there is no universally accepted backtesting procedure.

2.4. The role of Historical Simulation

Historical Simulation has been known to uphold attractive features like its simplicity and non-parametric nature. For that reason, the study of this topic goes back to a couple of decades ago, with researchers like Hull and White (1998) and Barone-Adesi and Giannopoulos (2001). Since then, other studies have been conducted regarding this approach (Du et al., 2024) and recent contributions have been made both in favor and against this framework.

García-Risueño (2025) has pointed out serious drawbacks in the use of HS, most noticeably its tendency to systematically underestimate ES. This measure heavily depends on the sample size, as it relies on past events to predict future ones, and with fewer datapoints, extreme events are less likely to be captured. For that reason, increased sample sizes and fitting the data to a continuous and fat-tailed probability density function are recommended to improve ES's accuracy.

Although the use of HS presents some challenges, it is favored not only by some regulators but also by some researchers. Mozumder et al. (2024) presented a different perspective that addressed a lack of research in user-friendly risk models in emerging markets, where banks operate under an underdeveloped regulatory environment and might face heightened risks and vulnerabilities. Given the circumstances, it was concluded that HS, a simple, model-free and user-friendly method, outperforms the other tests studied.

2.5. The missing link in research

Several studies have compared the performance of VaR and ES backtests, such as the ones from Almudhaf (2018), Novales and Garcia-Jorcano (2019) and Del Brio, Mora-Valencia and Perote (2020). However, few have directly compared these risk measures using the same portfolio. Furthermore, while backtesting methods for ES have been significantly explored, specifically within the HS framework, it is not the case.

A recent contribution made by Catarino (2023) explored several backtests, with special regard to the regulatory settings in place at the time. In this study, the strategy taken consists of estimating VaR and ES using a parametric and a non-parametric approach. The latter, that is, an HS model that incorporates volatility-adjusted realized returns, assumed to follow an elaborated specification of a generalized autoregressive conditional heteroskedasticity (GARCH) process, proved to be more adequate, however, bound by the methodology used.

Following these findings, the study of these measures within the HS framework, particularly using user-friendly approaches and diversified portfolios, is an interesting research topic.

By adopting this strategy, my aim is to fill this research gap, contributing to a better understanding of risk model validation under the evolving Basel regulatory framework.

CHAPTER 3

Theoretical Framework

3.1. Desirable properties of Risk Measures

Monitoring risk is critical for financial institutions. However, it can only be as effective as the metric taken to provide a quantitative summary of the potential losses. In light of this, researchers established a set of desirable properties for risk measures.

3.1.1. Coherence

Artzner et al. (1999) introduced four properties that any risk measure should satisfy to be considered coherent. For a risk measure, ρ, these are defined as it follows:

(i) Monotocity: The lower valued position bears a higher risk than the higher valued position.

$$\forall X, Y \in \mathcal{G}, X \leq Y \implies \rho(Y) \leq \rho(X) \tag{3.1}$$

(ii) Positive homogeneity: The risk of a scaled position is equal to the scaled risk of the position, when scaled by the same positive factor.

$$\forall \lambda \geq 0, X \in \mathcal{G}, \rho(\lambda X) = \lambda \rho(X) \tag{3.2}$$

(iii) Translation invariance: Adding a risk-free asset reduces risk by the same amount.

$$\forall X \in \mathcal{G}, \alpha \in \mathbb{R}, \rho(X + \alpha) = \rho(X) - \alpha \tag{3.3}$$

(iv) Subadditivity: The total risk of a portfolio must not exceed the sum of the individual risks, reflecting the benefits of diversification.

$$\forall X, Y \in \mathcal{G}, \rho(X + Y) \leq \rho(X) + \rho(Y) \tag{3.4}$$

When comparing VaR and ES with respect to this notion, ES has the advantage because it satisfies the four properties (Acerbi and Tasche, 2002). The same does not apply to VaR, as it generally does not satisfy the last one.

3.1.2. Elicitability

The realized mean of the score function, called mean score, allows the evaluation of forecasting performance, that is, the evaluation of predictive models using only a sample of forecasts and the realized outcomes. For that reason and because elicitability requires the existence of a scoring function, a connection was established with backtesting.

Coined by Lambert et el. (2008), this concept states that a functional ρ is elicitable if there exists a scoring function $S(x, y)$ that minimizes its expected value:

$$\rho = \arg\min_{x} \mathbb{E}[S(x, Y)] \tag{3.5}$$

Since VaR is essentially a quantile, the above is true through the scoring function[2]:

$$S(x, y) = (\mathbb{1}_{\{x \geq y\}} - \alpha)(x - y) \tag{3.6}$$

On the other hand, Gneiting (2011) showed that a scoring function whose expected value is minimized by ES does not exist and, therefore, ES is not elicitable. However, even though ES has this disadvantage, the lack of this property does not mean it cannot be backtested at all. Rather, it means that direct comparison across models using scoring functions applied only to ES is not allowed.

Following this notion, a key breakthrough came from the realization that VaR and ES are jointly elicitable (Fissler and Ziegel, 2016). Thus, although ES lacks individual elicitability, it can still be evaluated when combined with VaR.

3.1.3. Backtestability

Regarding the debate over the backtestability of ES, Acerbi and Szekely (2017) proposed a formal definition of a backtestable statistic. According to it, a statistic ρ is backtestable if there exists a function $Z_\rho(x, y)$ with a null expected value

$$\mathbb{E}[Z_\rho(x, Y)] = 0 \tag{3.7}$$

and that is strictly monotonic in x

$$\mathbb{E}[Z_\rho(x_1, Y)] < \mathbb{E}[Z_\rho(x_2, Y)], \; if \; x_1 < x_2. \tag{3.8}$$

In this context, if the test function is, in fact, equal to zero, then the prediction is correct; however, the further it is from zero, the worse the prediction is[3]. Based on this premise, elicitability is a necessary condition for backtestability. Furthermore, any backtestable statistic is elicited by the convex integral of the test function Z_ρ:

$$S_\rho(x, y) = \int^x Z_\rho(t, y) \; dt \tag{3.9}$$

This confirmed that ES, not being elicitable, cannot be backtested individually. Nonetheless, since it is jointly elicitable with VaR, it is possible to develop a backtesting procedure using the pair as an auxiliary statistic.

As previously discussed, backtest functions evaluate how accurate a model's prediction is, but that is not enough for absolute model validation. To that end, the distribution P_Z of Z_ρ, under some null hypothesis, needs to be known. For most risk measures, it is required access to the full daily predictive distribution of returns to then approximate P_Z via simulation or resampling. VaR, however, constitutes an exception to that premise. Being a quantile, its backtest is based on outcomes that follow a Bernoulli distribution, allowing the calculation of the p-value to be done directly, without needing the full predictive distribution.

[2]The notation α refers to the α-quantile.

[3]When the test function's value is positive it indicates overestimation, while negative values indicate underestimation.

3.2. Value-at-Risk

Often described as the maximum potential loss over a determined time period, at a certain confidence level, VaR is one of the most prominent risk measures. In other words, let X_t represent the daily uncertain profit and loss (P&L) of a risky financial position at time t ant let $F_t(x)$ denote its unknown cumulative distribution function (c.d.f.). For a given significance level $\alpha \in (0, 1)$, the 1-day VaR at level α, denoted by $\text{VaR}_{\alpha,t}$, is given by the following formula:

$$\text{VaR}_{\alpha,t} = -\inf\{x : F_t(x) \geq \alpha\} \tag{3.10}$$

In practice, the distribution $F_t(x)$ is unknown, therefore it has to be estimated. For that purpose, there are two main approaches, parametric and non-parametric ones.

Parametric methods assume that X_t follows an assigned distribution, such as a normal distribution, with mean μ_t and volatility σ_t, from which VaR is computed by

$$\text{VaR}_{\alpha,t} = -\left(\mu_t + \sigma_t \Phi^{-1}(\alpha)\right) = -\mu_t + \sigma_t \Phi^{-1}(1-\alpha) \tag{3.11}$$

where Φ is the c.d.f. of the standard normal distribution. For a one-day horizon, the value of μ_t can be neglected due to its proximity to zero. However, it is still necessary to estimate the value of σ_t. This can be achieved through simply using the standard deviation of past returns or through other approaches such as the Exponentially Weighted Moving Average (EWMA) method, developed by JPMorgan (1996).

Although computationally efficient, these methods strongly rely on the distributional assumption, which might not hold when faced with extreme market conditions. Financial data often exhibit fat-tails and skewness, resulting in the underestimation of extreme losses, under the normality assumption.

For that reason, non-parametric approaches are most frequently used. Among these, the most common is the Historical Simulation method, which avoids model misspecification by using past returns to directly estimate VaR. However, in spite of its non-parametric nature and ability to capture fat-tails, it still poses disadvantages like the sensitivity to the data range and the equal weighting of past events. To address these limitations, several extensions to HS have been proposed.

It is reasonable to assume that more recent data is more indicative of current risk conditions. For that reason, Weighted Historical Simulation assigns declining weights on past data (Boudoukh et al., 1998). To account for volatility changes, Hull and White (1998) proposed adjusting historical changes to reflect the ratio between the current and the historical daily volatility, which made it feasible to combine HS with econometric modeling. This breakthrough was then explored by another proposed method called Filtered Historical Simulation (FHS) (Barone-Adesi and Giannopoulos, 1998).

As mentioned previously, several methodologies have been developed to compute VaR due to its conceptual and computational simplicity. Still, its shortcomings motivate the use of other measures, especially ES.

3.3. Expected Shortfall

Also known as Conditional Value-at-Risk (CVaR), ES quantifies the average of losses that exceed the VaR threshold. Formally, for a given significance level $\alpha \in (0, 1)$, the Expected Shortfall is defined as the expected loss of a VaR_α exception:

$$\text{ES}_{\alpha,t} = -\mathbb{E}[X_t | X_t \leq -\text{VaR}_{\alpha,t}] \tag{3.12}$$

As proposed by Acerbi and Tasche (2002), this measure can also be represented as an integral of VaR

$$\text{ES}_{\alpha,t} = \frac{1}{\alpha} \int_0^\alpha \text{VaR}(u)\ du \tag{3.13}$$

where $\text{VaR}(u) \equiv \text{VaR}_u$. Through the integration of VaR at different levels, from 0 to α, this definition shows that ES contains information about the entire tail, unlike VaR itself, which only contains information about the one quantile level α.

Analogously to VaR, the estimation of ES depends on the approach used to estimate the underlying distribution of returns. Under the normality assumption, its parametric estimation is computed as

$$\text{ES}_{\alpha,t} = \int_0^{\Phi^{-1}(\alpha)} x\phi_t(x)\ dx \tag{3.14}$$

where $\phi_t(x)$ is the standard normal probability density function (p.d.f.). Furthermore, it is also possible to simplify this expression, under this distribution, by ignoring the drift adjustment $-\mu_t$, as previously discussed for the estimation of VaR. In doing so, the resulting expression is defined as:

$$\text{ES}_{\alpha,t} = \frac{\sigma_t}{\alpha} \phi(\Phi^{-1}(\alpha)) \tag{3.15}$$

As a way to discard the drawbacks of parametric assumptions, it is also possible to compute ES through non-parametric approaches. Under the standard HS approach, ES is estimated by retrieving the average of the worst returns, at a significance level α, that is, the average of the α-tail of the sample. Building on the significant advantages already identified, the description of this method is in direct alignment with the definition of ES, favoring its application. Nevertheless, the drawbacks stated for VaR still apply for ES, as do the possible adjustments to mitigate them.

Alongside the decision to replace VaR with ES, a change to the significance level was made. The replacement of $\text{VaR}_{1\%}$ with $\text{ES}_{2.5\%}$ is due to the fact that, for Gaussian returns, their values are identical, as under the standard normal distribution, where $\text{VaR}_{1\%} \approx 2.33$ and $\text{ES}_{2.5\%} \approx 2.34$. On the other hand, for distributions with heavier tails, ES at a significance level of 2.5% tends to exceed VaR at 1%, acting as a penalization mechanism in the form of increased capital requirements. As shown in Table 3.1, under the assumption that returns follow a t-student distribution, with $\mu = 0$ and $\sigma = 1$, the lower the degrees of freedom, ν, are, the more pronounced the penalization is.

TABLE 3.1. Values of $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$, for a Student-t distribution with different degrees of freedom.

Student-t	$\nu = 15$	$\nu = 10$	$\nu = 5$	$\nu = 2.5$
$\text{VaR}_{1\%}$	2.60	2.76	3.36	5.25
$\text{ES}_{2.5\%}$	2.64	2.82	3.52	6.21

In summary, ES offers information on the tail distribution, and therefore extreme losses that VaR does not take into account. This and other advantages, characteristic of this metric, justify its preference over the previous one, for a more robust framework that captures the real-world changes. However, backtesting this measure is not as straightforward as it is for VaR and, for that reason, it is relevant to address different, simple and achievable ways to do so.

CHAPTER 4

Data and Methodology

4.1. Sample Description

During the first semester of 2025, the United States of America faced a heightened period of instability. This was due not only to the geopolitical tensions between Russia and Ukraine and Israel and Iran, but also to the tariffs applied by President Donald Trump.

Given the relevance of understanding how the main ES modeling and backtesting approaches behave under stress events similar to those mentioned, particularly, the most recent one, this study focuses on US equities and takes into account the financial data downloaded from Reuters between January 2nd, 2012 and May 30th, 2025.

To ensure diversification, the portfolio consists of eleven equities, one from each sector of the S&P500 index (sectorial diversification), and three more equities, from the three major sectors of the index as of 30 May 2025. The majority of the chosen equities refer to the most prominent ones in their representative sector.

To ensure simplicity, for individual equities by sector, the weight of each equity is the weight of the representative sector, while for sectors with two equities, their weights are proportionally calculated from their market cap to fit the sector's weight in the index.

TABLE 4.1. Composition of the portfolio in analysis.

Sector	Weitgh	Equity	Weight
Information Technology	31.6%	Microsoft Corp (MSFT)	16.1%
		Nvidia Corp (NVDA)	15.5%
Financials	14.3%	Berkshire Hathaway B (BRK.B)	6.9%
		JP Morgan Chase & Co (JPM)	7.4%
Consumer Discretionary	10.6%	Home Depot Inc (HD)	6.6%
		McDonald's Corp (MCD)	4.0%
Communication Services	9.6%	Alphabet Inc (GOOGL)	9.6%
Health Care	9.6%	Johnson & Johnson (JNJ)	9.6%
Industrials	8.7%	GE Aerospace (GE)	8.7%
Consumer Staples	5.9%	Walmart Inc (WWT)	5.9%
Energy	3.0%	Exxon Mobil Corp (XOM)	3.0%
Utilities	2.5%	NextEra Energy Inc (NEE)	2.5%
Real Estate	2.1%	American Tower Corp (AMT)	2.1%
Materials	1.9%	Linde Plc (LIN)	1.9%

Through the retrieved information, namely the closing prices per day and the weights of each equity, the P&L of the portfolio is determined so that VaR and ES can be computed.

Recent studies on backtesting ES brought attention back to a simple and user-friendly,

non-parametric method called Historical Simulation. Following this path, this thesis compares the standard HS method with Weighted Historical Simulation, an extension of HS that assigns higher weights to more recent data.

4.2. Historical Simulation

The standard HS model relies directly on the empirical distribution of past returns to estimate VaR and ES. In the context of VaR, this consists of estimating the α-quantile of that distribution, which for a look-back window of n past returns, X_{t-1}, X_{t-2}, ..., X_{t-n}, arranged in ascending order, is denoted by:

$$\text{VaR}_{\alpha,t}^{\text{HS}} = -\ percentile\{\{X_{t-1}, X_{t-2}, ..., X_{t-n}\}, 100\alpha\} \tag{4.1}$$

After the estimation of VaR, it is possible to estimate ES, using the formula bellow:

$$\text{ES}_{\alpha,t}^{\text{HS}} = -\frac{\sum_{i=1}^{n} \mathbb{1}_{\{X_{t-i} \leq -\text{VaR}_{\alpha,t}^{\text{HS}}\}} X_{t-i}}{\sum_{i=1}^{n} \mathbb{1}_{\{X_{t-i} \leq -\text{VaR}_{\alpha,t}^{\text{HS}}\}}} \tag{4.2}$$

This approach is favored by many for being easy to implement and not requiring distributional assumptions or additional estimations. Even so, some of its major critiques are that it heavily depends on the sample size and that it gives equal weighting to all observations, regardless of how old they are.

To address these limitations, this dissertation follows two recommendations. García-Risueño (2025) noted the importance of having larger sample sizes. For that reason and to better reflect real-world behavior, this study uses a rolling window method, consisting of 500 days, to estimate the daily values of the measures. On the other hand, Mozumder et al. (2024) addressed the relevance of using user-friendly approaches like HS. Given that and to overcome the equal weighting criticism, another extension of the HS approach, WHS, is also applied.

4.2.1. Weighted Historical Simulation

A simple method that enjoys the advantages of HS while overcoming some of its criticism is the Weighted Historical Simulation. This approach is an extension of standard HS that assigns heavier weights to more recent events, as they better reflect current economic conditions and are more likely to influence future market behaviors.

Boudoukh et al. (1998) computed VaR under this framework, also known as the hybrid approach, through the following steps:

Step 1: Apply the weighting scheme to the returns, where λ is the weight factor.

$$w_{t-i} = \frac{\lambda^{n-(t-i)}(1-\lambda)}{1-\lambda^n},\ i = 1, ..., n,\ \lambda \in [0, 1] \tag{4.3}$$

Step 2: Order the returns in ascending order.

Step 3: Accumulate the weights, starting from the lowest return, until α is reached.

Step 4: Use linear interpolation between adjacent points to achieve its exact value.

After the estimation of VaR, it is possible to compute ES as the weighted average of the losses that exceed VaR.

$$\mathrm{ES}_{\alpha,t}^{\mathrm{WHS}} = -\frac{\sum_{i=1}^{n} \mathbb{1}_{\{X_{t-i} \leq -\mathrm{VaR}_{\alpha,t}^{\mathrm{WHS}}\}} w_{t-i} X_{t-i}}{\sum_{i=1}^{n} \mathbb{1}_{\{X_{t-i} \leq -\mathrm{VaR}_{\alpha,t}^{\mathrm{WHS}}\}} w_{t-i}} \tag{4.4}$$

This method offers a good balance between simplicity and adaptability, as it is a direct extension of HS that, by using the weighting scheme, is more adaptive to changing market conditions. Furthermore, compared to other more complex models, such as FHS or GARCH-based approaches, WHS is easier to implement and does not require any distributional assumptions. However, its accuracy relies on the choice of the decay parameter, which can be negatively affected if it is poorly calibrated.

This approach has been explored for VaR using two decay parameters, 0.97 and 0.99 (Boudoukh et al., 1998). Conclusions indicated that the higher decay produced better results.

Building on their findings and on their illustrative example that used $\lambda = 0.98$ to demonstrate the method's adaptability over time, it becomes particularly interesting to evaluate WHS using a decay factor of 0.98, as a mid-ground between the two already explored. Additionally, a slower decay, closer to HS, is also explored to understand if the findings also apply to these parameters. Given that Hull (2023) used a decay factor of 0.995, this study computes the age-weighed VaR and ES with $\lambda = 0.98$ and a $\lambda = 0.995$, for a rolling window of 500 observations.

4.3. Backtesting Value-at-Risk

In order to validate risk models and ensure regulatory compliance, it is essential to backtest the metric used. As previously addressed, the process of assessing the model's accuracy in predicting the potential losses using VaR is rather simple, consisting of comparing the estimates with the actual returns and counting the number of times the predicted value was exceeded (VaR breaches).

In light of this, one can define the breach indicator function at time t, at the significance level α, as

$$I_{\alpha,t} = \mathbb{1}_{\{X_t < -\mathrm{VaR}_{\alpha,t}\}} = \begin{cases} 1, \ X_t < -\mathrm{VaR}_{\alpha,t} \\ 0, \ otherwise \end{cases} \tag{4.5}$$

where $I_{\alpha,t}$ is equal to one if there is VaR breach or zero otherwise.

Therefore, for a look-back window of n past returns, the observed breach rate is calculated by:

$$\hat{\alpha} = \frac{1}{n} \sum_{t=1}^{n} I_{\alpha,t} \tag{4.6}$$

4.3.1. Backtesting VaR under the FRTB Framework

The backtesting framework established by the Basel Committee is based on the principle that, if correctly specified, the VaR breach indicator should follow an independent and identically distributed (i.i.d.) Bernoulli process, with probability α. Extending this notion

to n trading days, this means that the total number of VaR exceedences should follow a Binomial(n, α) distribution, resulting in the following hypothesis test:

HYPOTHESIS H_0. $\mathbb{E}\left[\sum_{t=1}^{n} I_{\alpha,t}\right] = n\alpha$

HYPOTHESIS H_1. $\mathbb{E}\left[\sum_{t=1}^{n} I_{\alpha,t}\right] \neq n\alpha$

This statistical framework was then formalized into a Traffic Light system based on the cumulative probabilities of the binomial distribution of observed VaR breaches.

As its name indicates, this system is divided into three colour zones, the green zone, when the cumulative probability is below 95%, the amber zone, when the cumulative probability is between 95% and 99.99%, and the red zone, when the cumulative probability surpasses 99.99%. The probability of outcomes falling into the red category is very low, indicating that there is almost certainly a problem with the model. For outcomes that fall into the amber category, they are generally associated with inaccurate models. For that reason, higher capital requirements are imposed in the form of a backtesting add-on that is based on the number of exceptions.

TABLE 4.2. Basel Traffic Light backtesting zones, for $\alpha = 1\%$ and $n = 250$.

Backtesting Zone	Number of Exceptions	Backtesting Multiplier	Comulative Probability
Green	0	1.50	8.11%
	1	1.50	28.58%
	2	1.50	54.32%
	3	1.50	75.81%
	4	1.50	89.22%
Amber	5	1.70	95.88%
	6	1.76	98.63%
	7	1.83	99.60%
	8	1.88	99.89%
	9	1.92	99.97%
Red	≥ 10	2.00	99.99%

4.3.2. Unconditional Coverage Test (Kupiec, 1995)

Among the most established backtesting procedures is the Kupiec's Unconditional Coverage test, also known as the Proportion of Failures (POF) test, for relying on the comparison between the observed and the predicted failure rate. Similarly to the FRTB approach, the null hypothesis relies on the assumption that exceptions follow an i.i.d. Bernoulli distribution, which, using the previously notation, is defined as:

HYPOTHESIS H_0. $\hat{\alpha} = \alpha$

HYPOTHESIS H_1. $\hat{\alpha} \neq \alpha$

Denoting the total number of VaR exceedences as $n_e = \sum_{t=1}^{n} I_{\alpha,t}$, the likelihood ratio test statistic, $LR \sim \chi_1^2$, used in this hypothesis test is defined as:

$$LR = -2 \ln \left[\left(\frac{\alpha}{\hat{\alpha}} \right)^{n_e} \left(\frac{1-\alpha}{1-\hat{\alpha}} \right)^{n-n_e} \right] \tag{4.7}$$

This test is widely used, both in practice and in research. In particular, Deng and Qiu (2021) concluded that unconditional coverage tests are often more reliable than conditional ones and that these should be used with caution in some cases. For these reasons and to aim for both simplicity and practicability, this thesis focuses on unconditional coverage tests.

4.4. Backtesting Expected Shortfall

Although ES's superiority is recognized, the shift from VaR to it as the primary risk measure is not reflected in the backtesting procedure. This regulatory stance arises from the difficulties associated with backtesting ES, due to its non-elicitable nature. Nevertheless, relying exclusively on VaR-based backtests to validate ES models can be viewed as an inconsistent baseline framework to work with, which favors the importance of doing research on this specific topic.

4.4.1. Backtesting ES under the FRTB framework

As stated previously, the Basel framework requires the backtesting of ES to be based on VaR, through the process of

Step 1: Estimating ES with a confidence level of 97.5%.

Step 2: Backtesting VaR at a confidence level of 99%.

thereby, taking advantage of the relation between $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$.

Due to the definition of ES and the shortcomings associated with backtesting it, the approach of approximating this measure via VaR stems as a simple and direct way of assessing it. For that reason, other methods have also further explored this notion.

4.4.2. The Pearson and Nass Tests

Rooted in the representation of ES as an integrated VaR (Acerbi and Tasche, 2002), Emmer et al. (2015) proposed approximating ES_α[4] as an average of VaR at four confidence levels, where $q_\beta(L) = \text{VaR}_{1-\beta}(L)$ and $L = (X)^-$ is the portfolio losses distribution:

$$\text{ES}_{1-\beta} = \frac{1}{\beta} \int_{\beta}^{1} q_u(L) \approx \frac{1}{4} \left[q_\beta(L) + q_{0.75\beta+0.25}(L) + q_{0.5\beta+0.5}(L) + q_{0.25\beta+0.75}(L) \right] \tag{4.8}$$

Using these findings, Kratz et al. (2018) proposed expanding this approximation to any number of quantiles, considering the VaR probability levels $\beta_1, ..., \beta_N$, defined as

$$\beta_j = \beta + \frac{j-1}{N}(1-\beta), \ j = 1, ..., N, \ N \in \mathbb{N} \tag{4.9}$$

[4]In the authors notation α represents the confidence level however, because the present document refers to α as the significance level, $\beta = 1 - \alpha$ is used to represent the confidence level.

where $\beta_0 = 0$ and $\beta_{N+1} = 1$. Following the recommendations of the Basel Committee, let $\beta_1 = \beta = 0.975$ be the confidence level for ES.

Defining $I_{j,t} = I_{1-\beta_{j,t}}$, from equation (4.5), and denoting the sequence that counts the number of VaR exceedances at each level as

$$V_t = \sum_{j=1}^{N} I_{j,t}, \tag{4.10}$$

then, if specified correctly, $(V_t)_{t=1,...,T}$ should satisfy the following:

$$P(V_t \leq j) = \beta_{j+1} \tag{4.11}$$

$$\forall s \neq t, V_t, V_s \; independent \tag{4.12}$$

Under these conditions, the random vector $\vec{O} = (O_0, ..., O_N)$ defined as

$$O_j = \sum_{t=1}^{T} \mathbb{1}_{\{V_t = j\}}, \; j = 0, ..., N \tag{4.13}$$

should follow a multinomial distribution with T trials and $N + 1$ possible outcomes:

$$\vec{O} \sim \mathbf{MN}(T, (\beta_1 - \beta_0, ..., \beta_{N+1} - \beta_N)) \tag{4.14}$$

Based on the notion that $0 = \theta_0 < \theta_1 < ... < \theta_N < \theta_{N+1} = 1$ is an arbitrary sequence of parameters for the model $\vec{O} \sim \mathbf{MN}(T, (\theta_1 - \theta_0, ..., \theta_{N+1} - \theta_N))$, the hypothesis test is formalized accordingly:

Hypothesis H_0. $\beta_j = \theta_j, \; for \; j \in \{1, ..., N\}$

Hypothesis H_1. $\beta_j \neq \theta_j, \; for \; at \; least \; one \; j \in \{1, ..., N\}$

To evaluate this hypothesis test, the authors explore three tests: the Pearson Chi-Squared test, the Nass test and the Likelihood Ratio test (LRT). The first two tests are relatively straightforward and naturally fit with the empirical nature of the non-parametric methods used. In contrast, the latter is a complex procedure that requires estimating the model parameters via maximum likelihood methods which, in addition to not being standard practice in the backtesting framework, requires out-of-model assumptions that do not align with the nature of HS and WHS. For these reasons, the LRT is not considered in the present analysis.

(i) Pearson Chi-Squared Test (Pearson, 1900): standard goodness-of-fit procedure used to infer whether the observed losses beyond the VaR threshold are consistent with the model predictions. Essentially, the test statistic used is defined as the weighted sum of squared deviations between the observed and the expected cell violations:

$$S_N = \sum_{j=0}^{N} \frac{(O_{j+1} - T(\beta_{j+1} - \beta_j))^2}{T(\beta_{j+1} - \beta_j)} \underset{H_0}{\overset{d}{\sim}} \chi^2_N \tag{4.15}$$

This simple test does not require additional assumptions, which is consistent with the methods of HS and WHS used, but its accuracy decreases as N increases.

(ii) Nass Test (Nass, 1959): improved version of the Pearson test, designed to perform better when the cell probabilities are low (higher value of N). In this test, the correction factor

$$c = \frac{2\mathbb{E}[S_N]}{\text{var}[S_N]} \tag{4.16}$$

where $\mathbb{E}[S_N] = N$ and $\text{var}[S_N] = 2N - \frac{N^2+4N+1}{T} + \frac{1}{T}\sum_{j=0}^{N} \frac{1}{\beta_{j+1}-\beta_j}$, is used to introduce an improved approximation to the distribution of S_N:

$$cS_N \underset{H_0}{\overset{d}{\sim}} \chi^2_\nu, \text{ where } \nu = c\mathbb{E}[S_N] \tag{4.17}$$

Consequently, this test maintains the benefits of the Pearson test, while being more accurate for smaller samples but having slightly less power.

In the context of backtesting ES, both tests provide practical methods to assess the accuracy of the extreme losses predictions, being particularly compatible with HS and WHS, for relying on the observed data, instead of distributional assumptions.

Notably, the Nass test has the advantage of performing well with limited data, making it a reliable and user-friendly approach that is applicable to both regulatory settings and real-world situations. Also, Kratz et al. (2018) state that this test with $N = 8$ should be preferable to the Pearson test with $N = 4$.

Considering this and their conclusion that, regardless of the chosen test, $N = 4$ and $N = 8$ are the two levels that should be considered, these are the values of N used in this dissertation.

Overall, it is important to note that this methodology addresses the backtesting inconsistency of the Basel framework by directly evaluating the adequacy of ES models. By splitting the tail into several quantile-based intervals, instead of using a single quantile, this approach replicates the structure of the tail to then test the full distribution of losses within it.

CHAPTER 5

Empirical Results

This chapter presents empirical findings, integrating the methodology previously described with the evaluation of market risk models. That being said, the analysis is designed to provide a thorough evaluation of VaR and ES within the HS and WHS frameworks.

In the case of VaR, the backtesting procedures employed were the Basel Committee FRTB Traffic Light test (B) and the Kupiec's Unconditional Coverage test (K). For ES, the applied backtests were the Pearson (P_N) and Nass (N_N) tests, for levels $N = 4$ and $N = 8$. All backtests were performed using a rolling window of 250 observations, as recommended by the Basel framework.

In view of the above, I estimated $VaR_{1\%}$ and $ES_{2.5\%}$ by implementing the following models:

(i) HS model: standard Historical Simulation with a rolling window of 500 observations.
(ii) WHS model: Weighted Historical Simulation with a decay factor of 0.995 and a rolling window of 500 observations.
(iii) WHS model: Weighted Historical Simulation with a decay factor of 0.98 and a rolling window of 500 observations.

Additionally, it was also necessary to compute $VaR_{2.5\%}$ for all three models, along with the auxiliary quantiles required for the Pearson and Nass tests, at each level. The respective confidence and significance levels are presented in Table 5.1.

TABLE 5.1. Auxiliary quantiles for the Pearson and Nass tests: confidence and significance levels, for $\beta = 97.5\%$.

$N = 4$			$N = 8$		
j	β_j	α_j	j	β_j	α_j
1	0.975000	0.025000	1	0.975000	0.025000
			2	0.978125	0.021875
2	0.981250	0.018750	3	0.981250	0.018750
			4	0.984375	0.015625
3	0.987500	0.012500	5	0.987500	0.012500
			6	0.990625	0.009375
4	0.993750	0.006250	7	0.993750	0.006250
			8	0.996875	0.003125

The VaR and ES estimates produced by the HS and WHS models were evaluated on a daily basis against the realized P&L outcomes. Subsequently, the resulting exceptions were classified into one of the three regulatory zones, defined by the FRTB Traffic Light backtesting procedure, as a way to reflect the adequacy of the risk measure at that point

in time.

In Table 5.2, the results computed between November 19$^{\text{th}}$, 2014 and May 30$^{\text{th}}$, 2025, are summarized into the percentage of Green (G), Amber (A) and Red (R) outcomes, across the different tests.

TABLE 5.2. Rolling Sample $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$ backtesting results.

	Historical Simulation			Weighted Historical Simulation					
				$\lambda = 0.995$			$\lambda = 0.98$		
	G (%)	A (%)	R (%)	G (%)	A (%)	R (%)	G (%)	A (%)	R (%)
$\text{VaR}^{\text{B}}_{1\%}$	46.87	37.55	15.57	77.58	22.42	0.00	71.98	28.02	0.00
$\text{VaR}^{\text{K}}_{1\%}$	47.63	43.05	9.32	73.91	26.09	0.00	89.74	10.26	0.00
$\text{ES}^{\text{P}_4}_{2.5\%}$	69.47	17.03	13.50	82.64	13.21	4.15	87.30	12.70	0.00
$\text{ES}^{\text{P}_8}_{2.5\%}$	59.35	21.36	19.29	74.67	21.18	4.15	81.19	13.97	4.84
$\text{ES}^{\text{N}_4}_{2.5\%}$	69.47	19.40	11.14	82.79	13.65	3.57	87.30	12.70	0.00
$\text{ES}^{\text{N}_8}_{2.5\%}$	63.54	25.18	11.28	75.84	24.16	0.00	85.19	14.81	0.00

The empirical analysis summarized in Table 5.2 indicates that, as expected, the WHS model outperforms the standard HS framework.

Remarkably, in all the backtests conducted, a slight adjustment, such as replacing the standard HS, which consists of using a decay factor of $\lambda = 1$, for a WHS model with a decay parameter of $\lambda = 0.995$, significantly improved performance. Extending this evaluation to a WHS model with a lower decay parameter, that is, placing slightly greater weight in more recent data compared to the previously considered WHS specification, further emphasized the decay parameter's relevance. In this case, the results show that the lower decay WHS model contributed to an overall statistically improved performance, effectively eliminating nearly all instances of red zone classifications.

5.1. Empirical Evaluation of VaR: A Model-Based Comparative Backtesting Analysis

An initial assessment of the performance of the HS model in estimating VaR revealed clear limitations. The green zone allocations represent less than half of the total observations, which means that a substantial proportion of the results fall within the amber and red categories. Furthermore, both backtesting procedures applied to this model yielded similar results, reinforcing the conclusion that HS performed poorly under the tested conditions.

A more detailed analysis of the daily HS results of the Basel test indicates that this model suffers the most in periods of financial stress. In fact, the amber categorizations align with documented episodes of high volatility.

Specifically anticipating and following red zones, the amber outcomes appear in periods featuring the trade war between the US and China and the early pandemic shocks

(2018-2020), the Covid-19 crisis (2020-2021), the macroeconomic inflationary trends, restrictive monetary police and geopolitical tensions (2022-2023) and the uncertainty lived due to the tariffs applied by President Donald Trump (2025). Figure 5.1 presents a visualization of these findings, that is, the daily Basel categorizations of the HS estimates of VaR.

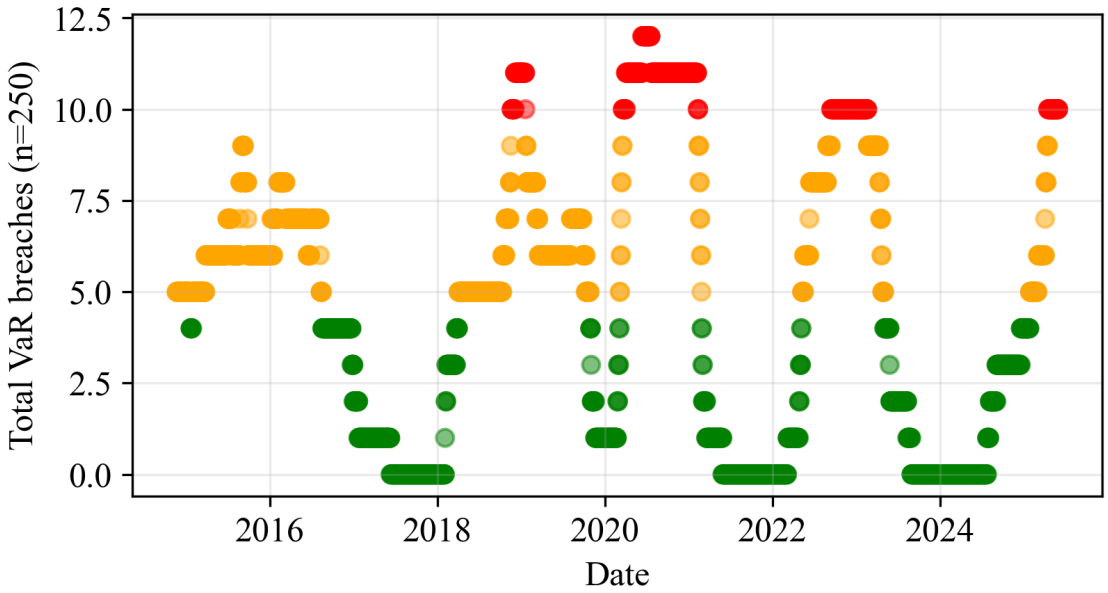

FIGURE 5.1. Daily green, amber and red classifications in the FRTB Traffic Light test for $\text{VaR}_{1\%}$ estimated through HS.

In contrast, the WHS model performed much better, indicating that this procedure responds faster to changing market conditions.

Both approaches to the weighted model effectively eliminate severe regulatory breaches (0% red allocations). However, the green and amber zone percentages progress differently within the backtests used. When the decay parameter is 0.995, the difference between both the proportions of green and amber outcomes in the Basel and Kupiec tests is minor, with both improving from the values associated with HS. Regardless, the case is not the same for WHS with a decay parameter of 0.98, where performance diverges more noticeably. Although the improvement from HS is still prominent in both backtests, when comparing the results of WHS with $\lambda = 0.98$ to those of $\lambda = 0.995$, the Basel test to the former returns around 5% less while Kupiec's returns around 15% more, regarding green outcomes. These results reveal a clear distinction between the regulatory and statistical dimensions of model evaluation, a topic that is further explored in the following section.

That said, it is relevant to understand, in greater detail, the model that exhibits the strongest performance under the Basel framework. In corroboration with the evidence provided by Boudoukh et al. (1998), this is the one with the highest decay factor, $\lambda = 0.995$.

As discussed previously, this approach achieved better results by fully eliminating the red outcomes observed under HS and significantly reducing the proportion of amber classifications. Nonetheless, periods of pronounced uncertainty still reveal concentrations

of amber outcomes.

Overall, the improvements seen on the WHS reflect the model's ability to incorporate recent information with greater weight, allowing it to adapt faster to the volatility changes in the market compared to the equal-weighted procedure. Still, this model is not immune to fragilities associated with stress events.

From a supervisory standpoint, the concentration of amber outcomes carries significant diagnostic value, as their persistence over extended periods may signal an underestimation of risk or model instability. Moreover, in practice, such patterns would trigger additional regulatory attention and lead to changes in the model.

Figure 5.2 presents a visualization of the daily Basel categorizations of the WHS estimates of VaR, with $\lambda = 0.995$.

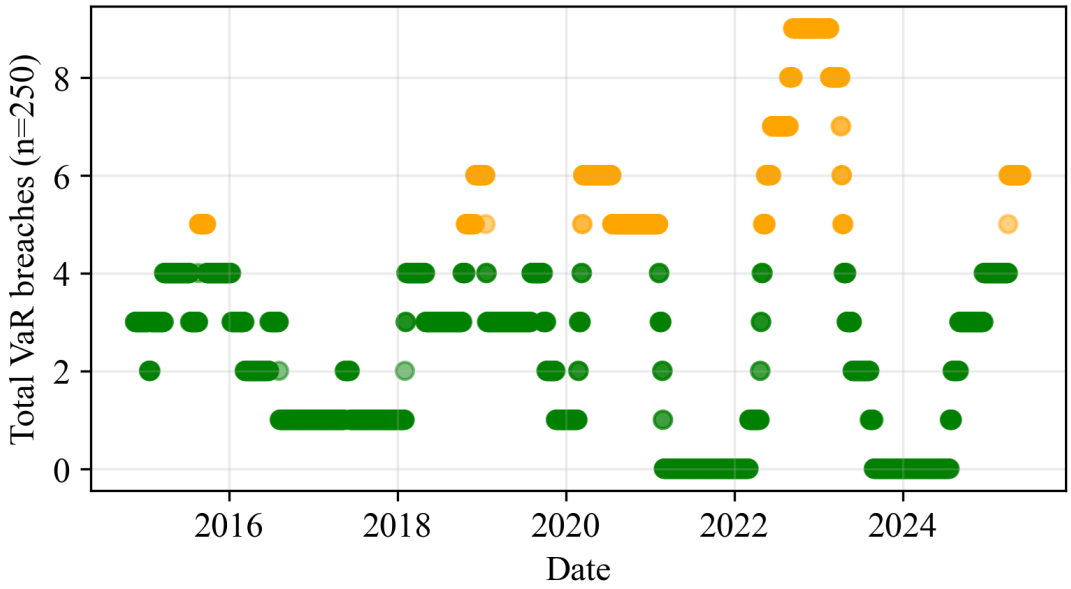

FIGURE 5.2. Daily green, amber and red classifications in the FRTB Traffic Light test for $\text{VaR}_{1\%}$ estimated through WHS, where $\lambda = 0.995$.

In a concluding remark, the comparison between the presented figures demonstrates that the WHS approach preserves the fundamental structure of the HS model while enhancing its performance, which favors its preference.

5.2. Empirical Evaluation of ES: A Model-Based Comparative Backtesting Analysis

Extending the analysis from VaR to ES, the limitations of standard HS are confirmed, as WHS once again yields better results.

In the context of HS, it becomes clear that, compared to the FRTB Traffic Light test, all models result in fewer 95% level rejections, that is, more green classifications. Even so, red allocations are still prominent, particularly in the Pearson test.

Considering more thoroughly the results of this backtest, it is evident that the percentage of HS results assigned to the red zone is higher in the Pearson test with 8 levels

than in the Basel test. This pattern is identified across the tested WHS models as well, culminating in a noticeable absence of red categorizations for all backtests except this one, when the decay parameter is 0.98. Another interesting finding is that, for all approaches, the discrepancy of green outcomes across the levels used is greater in the Pearson test than in Nass. This is a consequence of Pearson's decrease in effectiveness when N increases, as mentioned in Subsection 4.4.2, and is in line with the evaluation that Catarino (2023) did, confirming that its rejections are unreliable and, most likely, incorrect.

From another perspective, the Nass backtest is designed to perform better with higher values of N, working as an improved version of Pearson's, but it still has some drawbacks. In the case of standard HS, it returns around 4% less red zone allocations than the Basel test. This suggests that replacing the regulatory adopted test with it would be less conservative, since poorly performing models would face decreased capital add-ons. Remarkably, this tendency is not replicated in all tested models, as the Nass test with 8 levels yields results very similar to the evidence produced by the Basel test, when ES is estimated through WHS, with $\lambda = 0.995$. In addition, these particular outcomes also show more clearly the correction properties of this test with higher N, for both the improvement of red classifications and its closeness to the Basel test.

To be precise, in the context that the Pearson test with 8 levels is unreliable and that Nass with 4 levels provides less accurate evaluations, the power of the Pearson test for $N = 4$ is also compromised. This is a straightforward deduction due to the proximity of its values with the results of Nass with 4 levels and the conclusion that the Nass test with 8 levels is preferable to the Pearson test with 4 levels, as stated by Kratz et al. (2018).

Given the overall circumstances, it is evident that the multinomial tests diverge significantly from the FRTB backtest. Catarino (2023) delved into this topic, addressing their similarities to the Kupiec test, especially in the case of the Nass test. Since both approaches are based on asymptotic convergence to the chi-square distribution, they lead to less restrictive critical values, a tendency already identified in Section 5.1. This means that, by choosing the binomial test, the Basel Committee prioritized prudentiality over statistical refinement, therefore ensuring stricter capital requirements through conservative rejection thresholds. Even so, the results of this study reveal that the Nass test with $N = 8$ offers a balance between the regulatory and statistical perspectives, given its proximity to both VaR backtests. Furthermore, its theoretical framework aligns with the non-parametric nature of HS and the definition of ES. By contrast, the FRTB Traffic Light approach evaluates risk using the wrong target, that is, uses VaR to assess ES.

Based on these considerations, the Nass backtest with 8 levels emerges as the most suitable backtesting procedure among those tested and is going to serve as the primary focus on the subsequent analysis.

To further illustrate these backtesting results, I replicated the approach taken for VaR, for ES. Figure 5.3 provides a visualization of the daily classifications of HS estimates for ES, assigned by the Nass test with 8 levels, through the calculated p-value.

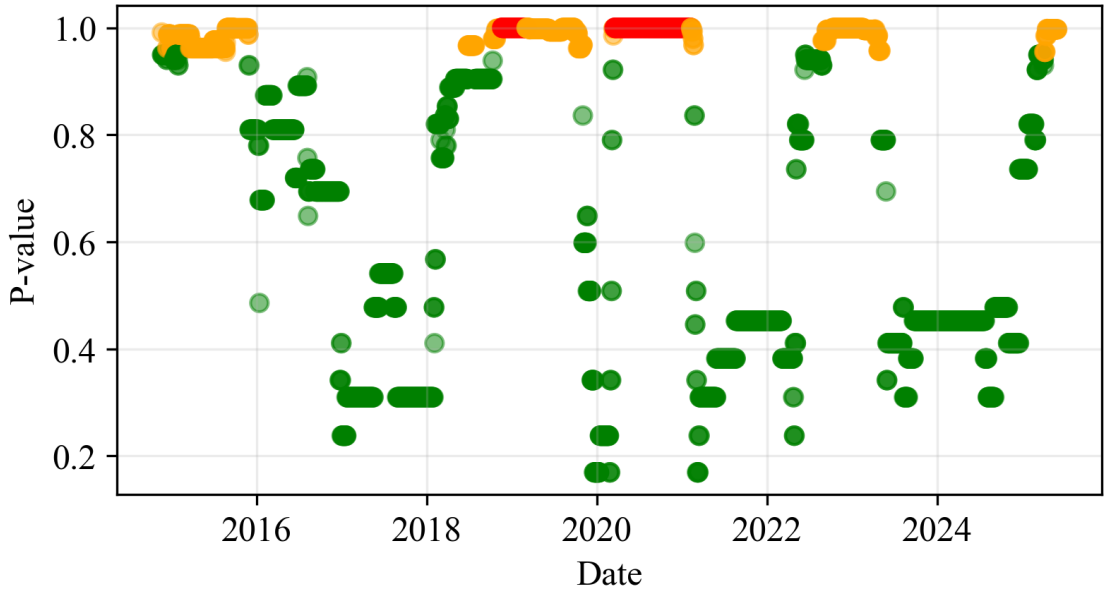

FIGURE 5.3. Daily green, amber and red classifications in the Nass test with $N = 8$, for $ES_{2.5\%}$ estimated through HS.

Consistent with expectations, the amber zones coincide with those identified in the Basel test in Figure 5.1. However, the incidence of red allocations in these scenarios is lower than what the Basel framework suggests.

An example of this is the absence of red allocations in 2025, a period of uncertainty in which the FRTB Traffic Light test returned such classifications. Figure 5.4 zooms in on the P&L distribution and on the VaR breaches that occurred in the last 250 days before May 30$^{\text{th}}$, 2025, the last day of the analysis.

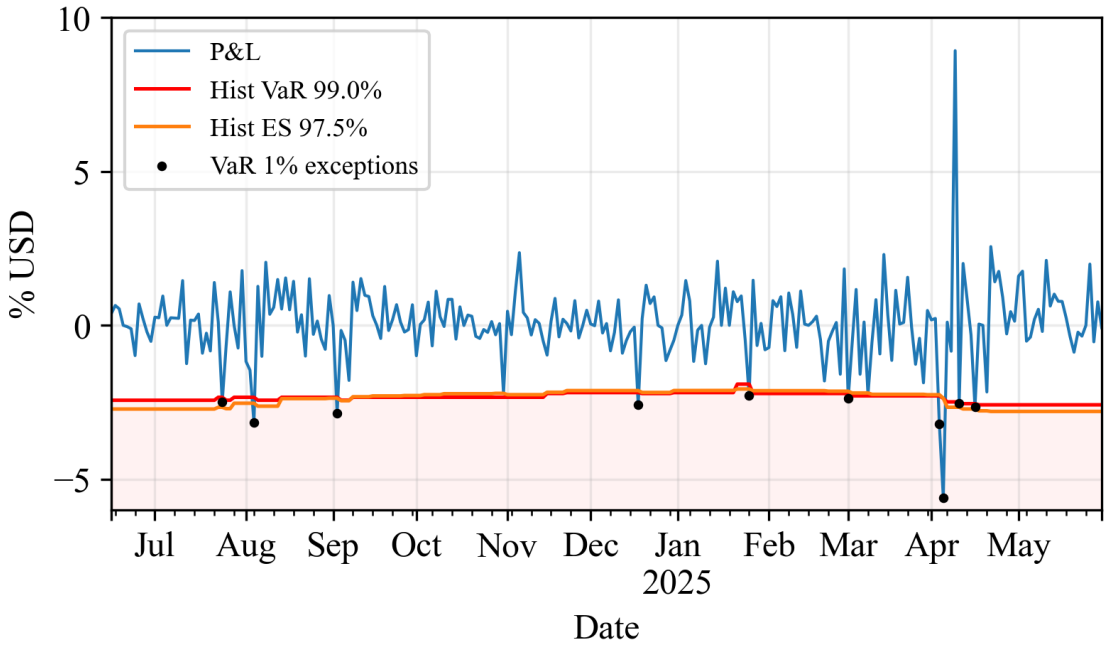

FIGURE 5.4. Portfolio P&L distribution, HS estimates for $VaR_{1\%}$ and $ES_{2.5\%}$ and $VaR_{1\%}$ breaches from June 14$^{\text{th}}$, 2024 to May 30$^{\text{th}}$, 2025.

Examining this figure, results demonstrated that VaR exceedances are sometimes covered by the ES estimates, highlighting the contrast between the Basel framework and the approach of the Nass test. This discrepancy could indicate that a model might fail under the Basel criteria for VaR but effectivity capture tail risk through ES. Yet, the contrary is also true.

Objectively, these findings are in line with the possible limitations associated with the Nass test with 8 levels, discussed previously, but also reflect the more nuanced assessment of model performance of ES-specific backtests. This means that the regulatory reliance on VaR breaches may exaggerate deficiencies in cases where ES remains properly calibrated, therefore reinforcing the need for consistency between risk measurement and model validation practices.

Building on the insight that the standard model underperformed for both measures and that the age-weighted scheme delivered significantly better results in the case of VaR, ES's results confirm the conclusion that, for WHS with a decay factor of 0.995, there are clear improvements. Figure 5.5 summarizes the results of the Nass test with $N = 8$, for this model.

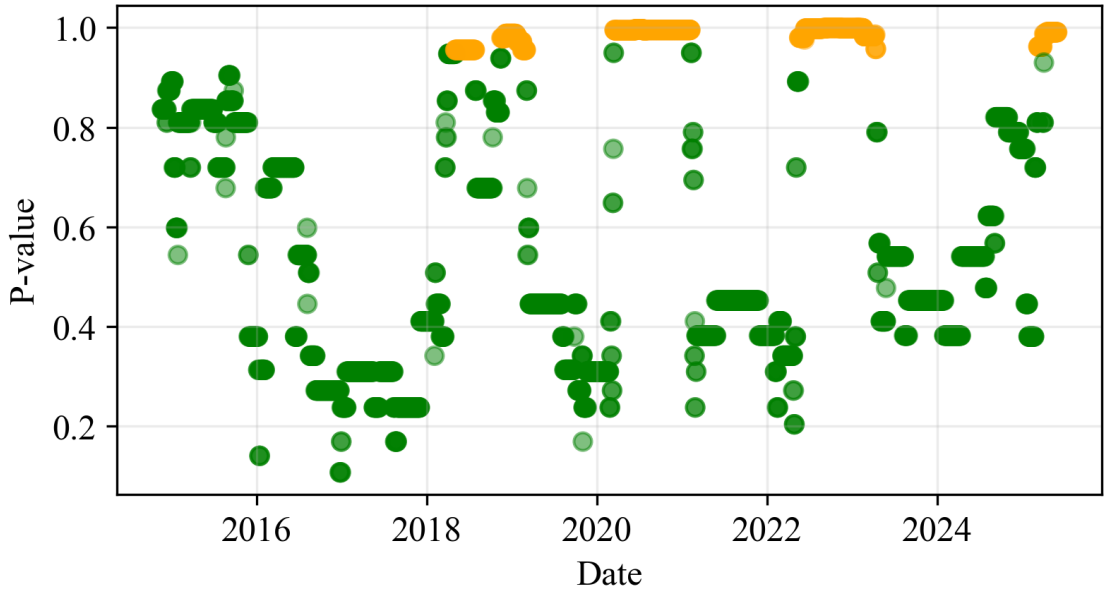

FIGURE 5.5. Daily green, amber and red classifications in the Nass test with $N = 8$, for $ES_{2.5\%}$ estimated through WHS,where $\lambda = 0.995$.

The prevalence of green and the total absence of red outcomes, as well as the alignment of results with those represented in Figure 5.2, not only confirms the conclusion that WHS outperforms standard HS, but also that the Nass test with $N = 8$ produces results very similar to those produced by the regulatory adopted test.

This evidence is of particular significance in the presence of market stress periods, when reliability is most critical. Although this model is not entirely immune to such fluctuations, the results remain accurate and notably consistent, reinforcing the advantages of choosing this model.

To visually demonstrate these findings and the notable decrease in VaR exceedances, focusing on the market instability during the first semester of 2025, Figure 5.6 zooms in on the P&L distribution and on the WHS VaR breaches that occurred in the last 250 days before May 30$^{\text{th}}$, 2025.

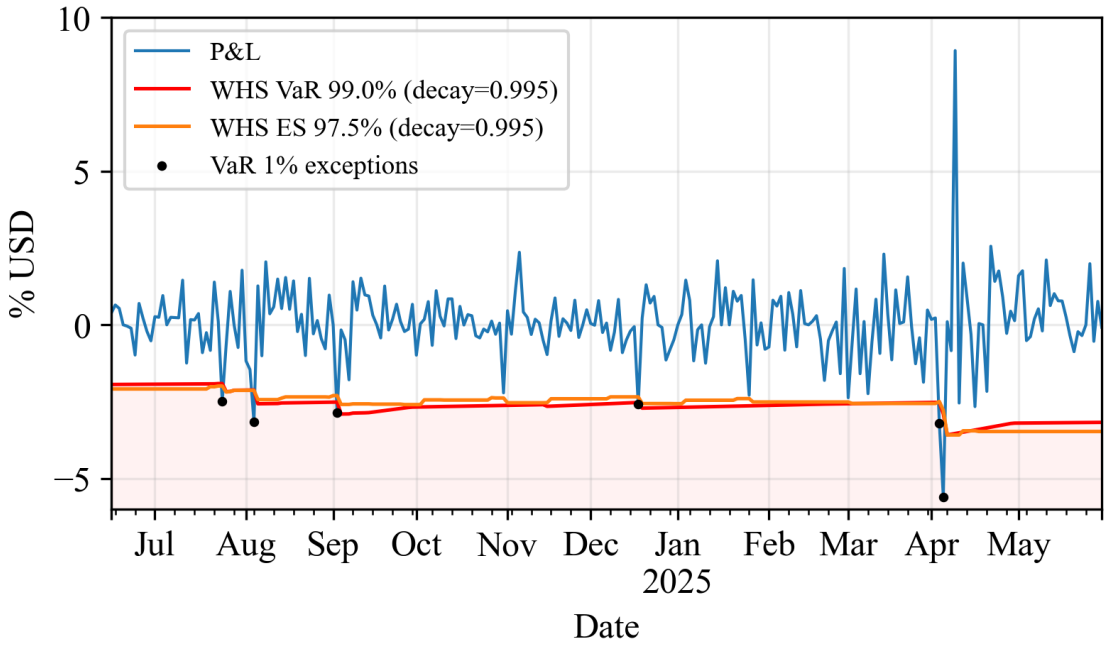

FIGURE 5.6. Portfolio P&L distribution, WHS, $\lambda = 0.995$, estimates for $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$ and $\text{VaR}_{1\%}$ breaches from June 14$^{\text{th}}$, 2024 to May 30$^{\text{th}}$, 2025.

As previously addressed, a slight adjustment that uses a weighting scheme with a slower decay already performed much better than standard HS. Nonetheless, the ES backtesting results summarized in Table 5.2 seem to indicate that a faster decay, $\lambda = 0.98$, produces statistically more accurate estimates.

Across multinomial tests, this model upholds the highest and lowest percentages of green and red outcomes, respectively, with only the Pearson test with 8 levels not fully eliminating red zone allocations. As first introduced in Subsection 5.1, these results are very close to those produced by Kupiec, but yield very differently from the Basel test. Not only that, they actually allocate more outcomes to the green zone, that were allocated to the amber zone by the FRTB test.

Using a lower parameter means that the weighting scheme decays faster, thus reacting more rapidly to recent market changes. On the one hand, this can improve statistical accuracy, as demonstrated by the multinomial and unconditional coverage tests. On the other hand, the increased sensitivity leads to more amber outcomes in the Basel backtest, even if they are more aligned to what is expected statistically.

In light of that, Figure 5.7 provides a visualization of the P&L distribution, as well as the WHS, with $\lambda = 0.98$, $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$ estimates and the VaR exceedences, across the entire period considered in this analysis.

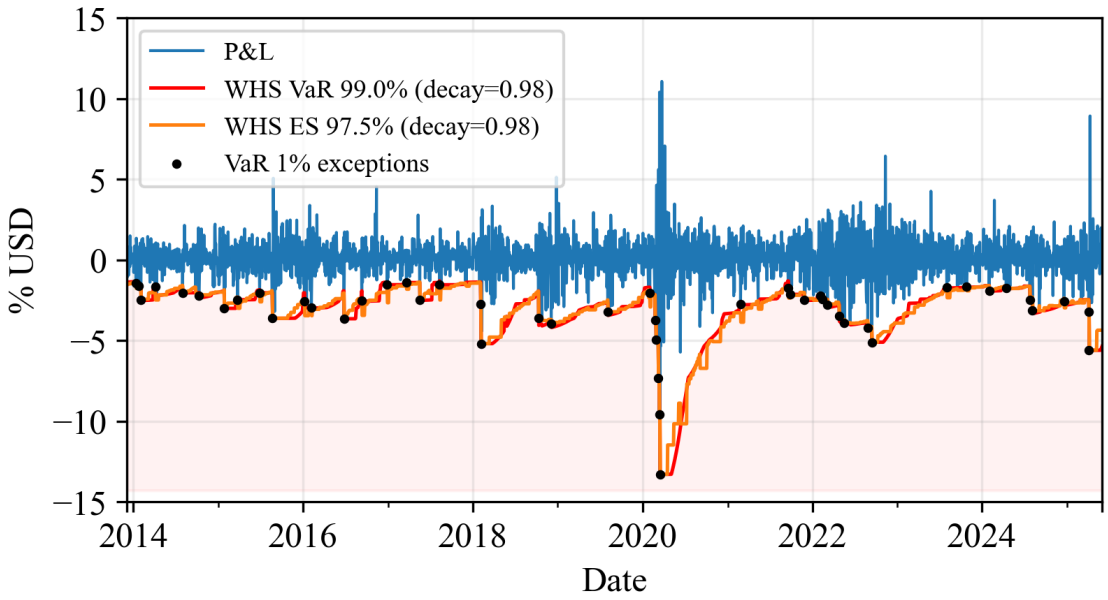

FIGURE 5.7. Portfolio P&L distribution, WHS estimates, with a decay parameter of $\lambda = 0.98$, for $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$, and $\text{VaR}_{1\%}$ breaches, from December 4^{th}, 2013 to May 30^{th}, 2025.

As this figure demonstrates, the $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$ estimates accurately follow the major losses in the P&L, justifying its statistical accuracy. Even so, several VaR exceedences are detected, which is the reason why, under the Basel test, this model underperforms.

Since this study is bound by the methodology used to backtest ES, even though WHS with $\lambda = 0.98$ performs better in the context of ES, for VaR regulatory compliance, the WHS model with $\lambda = 0.995$ is considered as the most accurate and reliable overall.

CHAPTER 6

Conclusions

Despite its long use in regulatory frameworks, VaR has well-documented shortcomings that led to its replacement with ES. Nevertheless, the new measure also faced persistent criticism. VaR is simpler and easier to backtest, consequently, due to the ongoing perception that ES lacks backtestability, regulation continues to rely on the use of VaR-based backtesting procedures, even when it is no longer the primary risk measure. Building on this, the contributions of this dissertation are its engagement in the long-standing debate over the backtestability of ES, while performing a cross-comparison between user-friendly estimation models and evaluating the performance of the chosen methods in the context of unstable market periods. For this analysis, it was taken into especial consideration the period of uncertainty lived in the first semester of 2025, in the US.

In light of that, I estimated VaR and ES through two non-parametric models based on HS, a widely used and favored method. Following that, I applied two multinomial backtesting procedures, the Pearson and Nass tests, to assess the backtestability of ES consistently with the HS framework.

Notably, the empirical analysis provides compelling evidence that demonstrates that ES can indeed be backtested.

On the one hand, these methods are relatively simple to compute and do not require model information, aligning with the model's nature. Furthermore, they are sustained by statistical results. On the other hand, they are less prudent than the currently recommended VaR-based backtest. Indeed, the Basel Committee opted for a more conservative backtest for VaR instead of the widely used unconditional coverage test. For that reason, and since the multinomial tests are closer to the Kupiec test, it is unlikely that these would be adopted for ES in their current form. Still, between the tested methods, the Nass test with 8 levels emerges as a simple and strong procedure that offers a good compromise between the two, providing greater coherence with the results expected by the Basel test. Not only that, this test also provides evidence that supports the importance of using ES-specific backtests, through a deeper analysis of VaR breaches between June 14^{th}, 2024 and May 30^{th}, 2025.

These findings corroborate the evidence provided by Catarino (2023) affirming that the Nass test with 8 levels is not a standalone replacement for the FRTB Traffic Light test but that it is a good complementary tool for internal validation. Furthermore, they reinforce the idea that backtesting ES is feasible and highlight the importance of addressing the incoherence of relying on VaR-based methods.

The comparative analysis of performance between HS and WHS clarified the implications associated with the chosen model. Standard HS is conceptually simple but exhibits some deficiencies, namely its inability to rapidly adapt to market regime shifts. As a direct consequence, this model produced a high incidence of red and amber allocations in periods of financial stress. By comparison, WHS significantly reduced these outcomes, actually eliminating red results in the most coherent ES backtest, the Nass test with $N = 8$, demonstrating its greater adaptation ability.

Although WHS improved performance, its adaptive property does not fully eliminate the amber classifications. In fact, reflecting the instability of the financial market, the amber outcomes of this model remain prevalent in the same episodes in which the HS model fails. This is the case of the classifications allocated to the ES estimates of the first semester of 2025, reflecting the uncertainty lived in the US, due to the tariffs applied by President Donald Trump. From a regulatory point of view, the persistence of these classifications would require closer monitoring and, most likely, trigger model modifications.

Fundamentally, WHS accomplished the expected results, outperforming standard HS, by introducing an age-weighted scheme that allows it to respond faster to changing market conditions. In addition, this model has the advantage of retaining the same intuitive and user-friendly properties of the standard approach, which favors its use by practitioners. This is in line with the conclusions of Mozumder et al. (2024), which advocate for the reliability of user-friendly approaches.

Conversely, there is an additional limitation to this model, compared to HS, as its effectiveness is bound by the choice of decay parameter. With a slower decay, $\lambda = 0.995$, this model achieved strong regulatory compliance as well as statistical validity. Moreover, the results of this estimating process are smoother and more stable across the different tests, particularly under Basel's framework. Alternatively, $\lambda = 0.98$ offers greater statistical accuracy, although at the cost of greater variability. The use of a faster decay allows the model to adapt more rapidly to market changes, providing estimates that better align to what is expected by the multinomial and Kupiec tests. Nevertheless, results diverge under the Basel framework, where WHS with $\lambda = 0.995$ is more accurate. Therefore, for a more balanced approach, the model with the higher decay factor must be considered. In the the regulatory context, that is the Traffic Light test used for VaR, this is in line with the evidence provided by Boudoukh et al. (1998), favoring the higher parameter. On the contrary, for ES, the model with the lower parameter performs better because it adapts more rapidly to the financial market changes.

This research offers valuable insights. In summary, it confirms the backtestability of ES, emphasizes the relevance of backtesting it using ES-based approaches and advocates for the use of effective user-friendly approches, such as WHS. Moreover, despite the finding that no model or test fully eliminated the challenges of tail risk measurement, this study provides evidence supporting the conclusion that combining robust methodologies

with adaptive models enhances both the reliability of forecasts and their regulatory acceptability.

This study offers a comprehensive analysis but it is not without limitations. Firstly, it focuses only on non-parametric, simple approaches to model the risk measures, thereby disregarding more sophisticated models such as the Hull and White volatility adjusted HS, EWMA and GARCH approaches. Even though the restriction of using HS and WHS aligns with the objective of using user-friendly approaches, it might also narrow the scope of comparison, overlooking models that might enhance performance during stress events. Additionally, this thesis only uses easy to compute backtests that might not capture the full spectrum of emerging ES backtesting techniques. Finally, this reasearch is also limited to fixed window lenghts, decay factors and asset classes, which might impact the results obtained. The exclusive focus on equities is justified by the chosen period of analysis, that is, the 2025 announced tariff measures, in which the equity market was very affected and exposed to significant losses. However, by excluding other asset classes, this analysis overlooks other channels of risk transmission and other dimensions of diversification, thereby limiting the replicability of its findings.

Stemming from this, future research could extend this analysis to other backtesting methodologies, for a more in-depth assessment of ES. Furthermore, using portfolios composed of multiple asset classes, other financial markets and varying time windows could also provide a deeper understanding of the studied approaches. Finally, additional follow-up studies could explore other weighting schemes or adapt the current WHS method, by analyzing other decay factors or enhancing the current weighting procedure, to develop a more robust approach to estimate ES, while maintaining its user-friendly characteristics.

In conclusion, no single risk measure, model, or backtesting procedure can fully mitigate the gaps in risk management. Nonetheless, the methods examined in this thesis provide practitioners with a practical framework for evaluating tail risk, offering valuable insights into the vulnerabilities associated with stressed market conditions. Not only that, they are also in line with key regulatory objectives by enhancing transparency and promoting consistency in the process of evaluating model performance.

References

Acerbi, C., and Szekely, B. (2014). Backtesting expected shortfall. *Risk*, 27(11), 76-81.

Acerbi, C., and Szekely, B. (2017). General Properties of Backtestable Statistics. *Available at SSRN Electronic Journal; ISSN 1556-5068.*

Acerbi, C., and Tasche, D. (2002). On the coherence of Expected Shortfall. *Journal of Banking and Finance* 26(7), 1487-1503.

Almudhaf, F. (2018). Backtesting expected shortfall: evidence from European securitized real estate. *Applied Economics Letters*, 25(3), 176–182.

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

Barone-Adesi, G., and Giannopoulos, K. (2001). Non parametric VaR Techniques. Myths and Realities. *Economic Notes*, 30(2).

Bayer, S. and Dimitriadis, T. (2022). Regression-Based Expected Shortfall Backtesting. *Journal of Financial Econometrics*, 20(3) 437–471.

BCBS – Basel Committee on Banking Supervision (2013). Fundamental review of the trading book: a revised market risk framework. *Bank for International Settlements.* `https://www.bis.org/publ/bcbs265.pdf`

BCBS – Basel Committee on Banking Supervision (2019). Minimum capital requirements for market risk. *Bank for International Settlements.* `https://www.bis.org/bcbs/publ/d457.pdf`

Boudoukh, J., Richardson, M., and Whitelaw, R. (1998). The best of both worlds: a hybrid approach to calculating value at risk. *Risk*, 11(5), 64-67.

Catarino, J. (2023), Backtesting expected shortfall: Comparative study and impact analysis on capital requirements, MSc Thesis in Finance, *ISCTE.*

Costanzino, N., and Curran, M. (2018). A Simple Traffic Light Approach to Backtesting Expected Shortfall. *Risks*, 6(1), 2.

Del Brio, E. B., Mora-Valencia, A., and Perote, J. (2020). Risk quantification for commodity ETFs: Backtesting value-at-risk and expected shortfall. *International Review of Financial Analysis*, 70, 101163.

Deng, K. and Qiu, J. (2021). Backtesting expected shortfall and beyond. *Quantitative Finance*, 21(7): 1109–1125.

Du, Z. and Escanciano, J. C. (2017). Backtesting Expected Shortfall: Accounting for Tail Risk. *Management Science*, 63(4), 940–958.

Du, Z., Pei, P., Wang, X., and Yang, T. (2024). Powerful Backtests for Historical Simulation Expected Shortfall Models. *Journal of Business and Economic Statistics*, 42(3), 864–874.

Emmer, S., Kratz, M., and Tasche, D. (2015). What is the best risk measure in practice? A comparison of standard measures. *Journal of Risk* 18(2), 31-60.

Fissler, T., and Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4), 1680-1707.

Fissler, T., Ziegel, J. F., Gneiting, T. (2016). Expected Shortfall is jointly elicitable with Value at Risk- Implications for backtesting. *RePEc: Research Papers in Economics.*

García-Risueño, P. (2025). Historical Simulation Systematically Underestimates the Expected Shortfall. *Journal of Risk and Financial Management*, 18(1), 34.

Gneiting, T. (2011). Making and Evaluating Point Forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.

Hull, J., and White, A. (1998). Incorporating volatility updating into the historical simulation method for value-at-risk. *Journal of risk*, 1(1), 5-19.

Hull. J. C. (2023). Risk Management and Financial Institutions (6th ed.). *Wiley.*

Jorion, P. (2006). Value at risk: The new benchmark for managing financial risk (3rd ed.). *McGraw-Hill.*

JPMorgan/Reuters (1996). RiskMetrics - Technical Document. (4th ed.) *Morgan Guaranty Trust Company.*

Kratz, M., Lok, Y. H., and McNeil, A. J. (2018). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall. *Journal of Banking and Finance*, 88, 393–407.

Kupiec, P. H. (1995). Techniques for Verifying the Accuracy of Risk Measurement Models. *The Journal of Derivatives*, 3(2), 73–84.

Lambert, N. S., Pennock, D. M., and Shoham, Y. (2008). Eliciting properties of probability distributions. *SIGecom Exchanges*, 7(3), 1–5.

Moldenhauer, F. and Pitera, M. (2019). Backtesting expected shortfall: A simple recipe? *Journal of Risk*, 22(1): 17-42.

Mozumder, S., Abedin, M. Z., Lalon, R., and Hossain, A. (2024). Which User-Friendly Model is the Best for BASEL-III? An Emerging Market Study. *Computational Economics*, 64(5), 3049–3086.

Nass, C. A. G. (1959). The χ^2-test for small expectations in contingency tables, with special reference to accidents and absenteeism. *Biometrika*, 46(3/4), 365-385.

Nolde, N., Ziegel, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics*, 11(4).

Novales, A., and Garcia-Jorcano, L. (2019). Backtesting extreme value theory models of expected shortfall. *Quantitative Finance*, 19(5), 799–825.

Patton, A. J., Ziegel, J. F., and Chen, R. (2019). Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk). *Journal of Econometrics*, 211(2), 388–413.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157–175.

Appendix A

TABLE A.1. Descriptive Statistics of the Diversified US Equity Portfolio.

	Mean	Median	Min	Max	Standard Deviation	Skewness	Excess Kurtosis
MSFT	0.0009	0.0002	-0.1474	0.1422	0.0162	0.1141	8.5567
NVDA	0.0021	0.0009	-0.1876	0.2981	0.0281	0.6358	9.8435
BRK.B	0.0006	0.0002	-0.0959	0.1161	0.0114	-0.0232	11.4496
JPM	0.0007	0.0000	-0.1496	0.1801	0.0164	0.2655	13.3103
HD	0.0007	0.0004	-0.1979	0.1375	0.0143	-0.7532	17.6029
MCD	0.0004	0.0002	-0.1588	0.1813	0.0118	0.4161	32.4077
GOOGL	0.0008	0.0003	-0.1163	0.1626	0.0170	0.2927	8.0574
JNJ	0.0003	0.0000	-0.1004	0.0800	0.0106	-0.2573	9.9382
GE	0.0005	0.0000	-0.1516	0.1473	0.0198	0.1389	7.9596
WWT	0.0005	0.0002	-0.1138	0.1171	0.0125	0.1927	16.3316
XOM	0.0002	0.0000	-0.1222	0.1269	0.0158	0.0399	7.5891
NEE	0.0005	0.0006	-0.1342	0.1369	0.0141	-0.2418	11.1600
AMT	0.0005	0.0001	-0.1516	0.1222	0.0150	0.0859	8.7458
LIN	0.0005	0.0000	-0.1028	0.1169	0.0132	0.1396	7.4941
Portfolio	0.0009	0.0006	-0.1329	0.1107	0.0116	-0.2186	14.6786

Appendix B

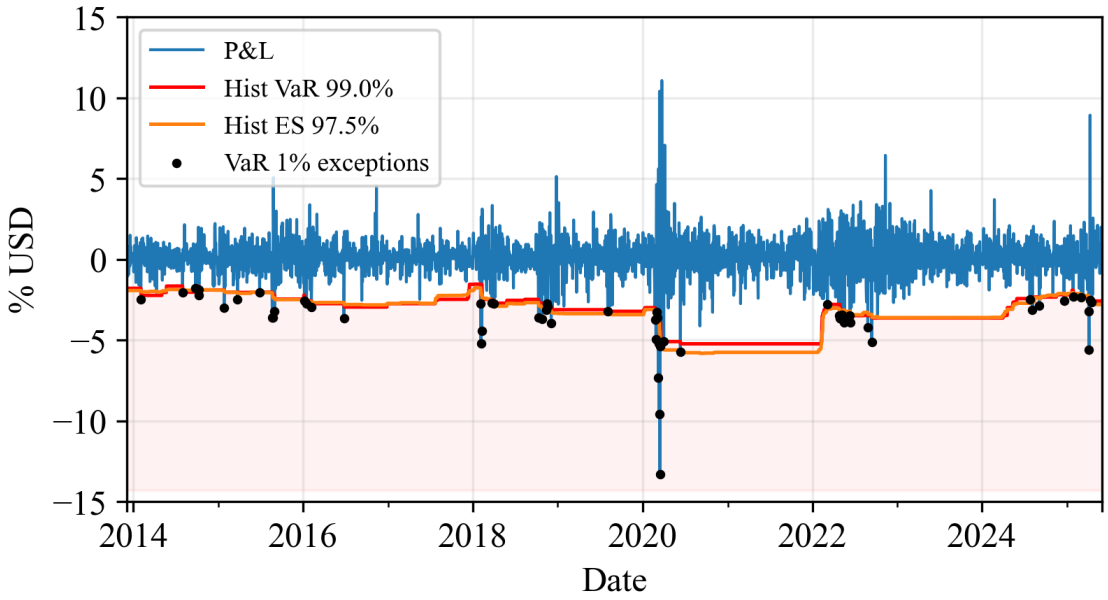

FIGURE B.1. Portfolio P&L distribution, HS estimates for $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$, and $\text{VaR}_{1\%}$ breaches, from December 4$^{\text{th}}$, 2013 to May 30$^{\text{th}}$, 2025.

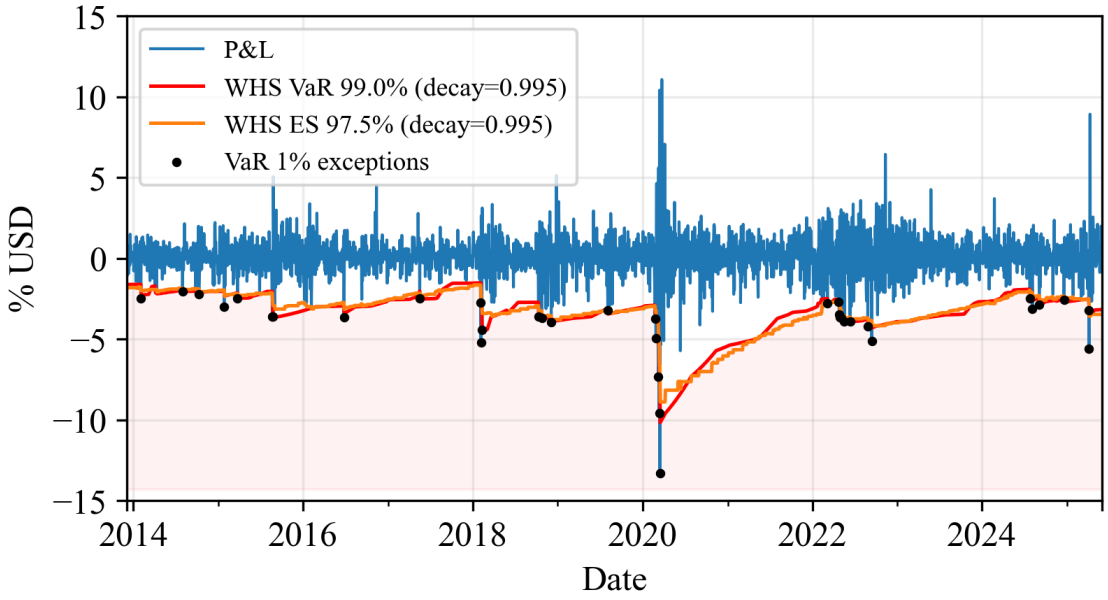

FIGURE B.2. Portfolio P&L distribution, WHS estimates, with a decay parameter of $\lambda = 0.995$, for $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$, and $\text{VaR}_{1\%}$ breaches, from December 4$^{\text{th}}$, 2013 to May 30$^{\text{th}}$, 2025.

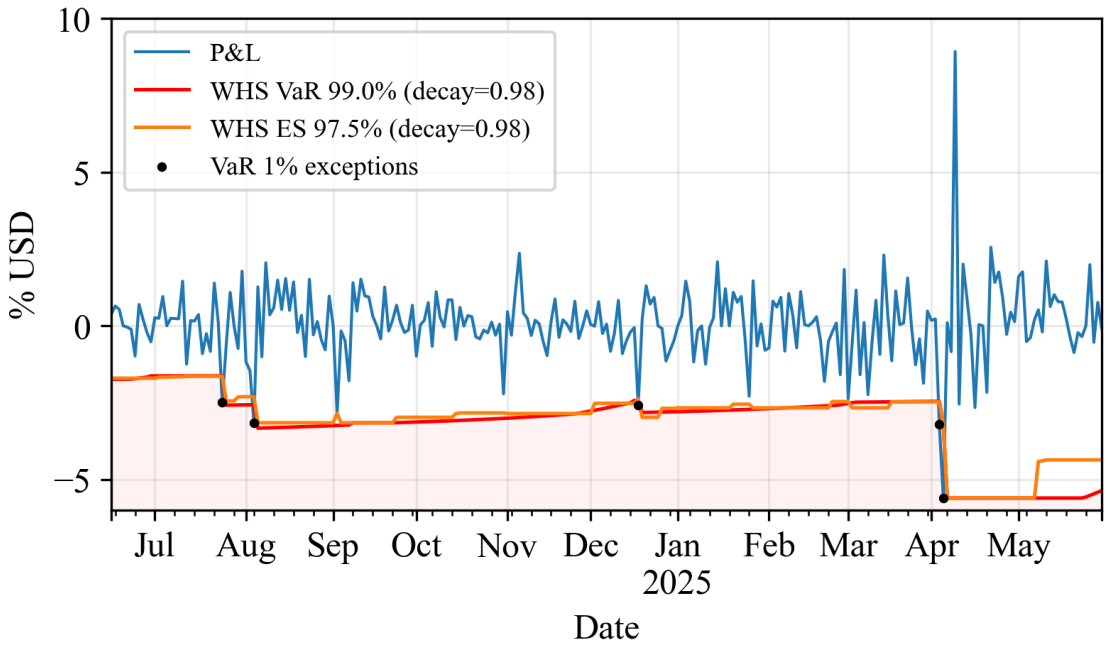

FIGURE B.3. Portfolio P&L distribution, WHS, $\lambda = 0.98$, estimates for $\text{VaR}_{1\%}$ and $\text{ES}_{2.5\%}$ and $\text{VaR}_{1\%}$ breaches from June 14^{th}, 2024 to May 30^{th}, 2025.

Appendix C

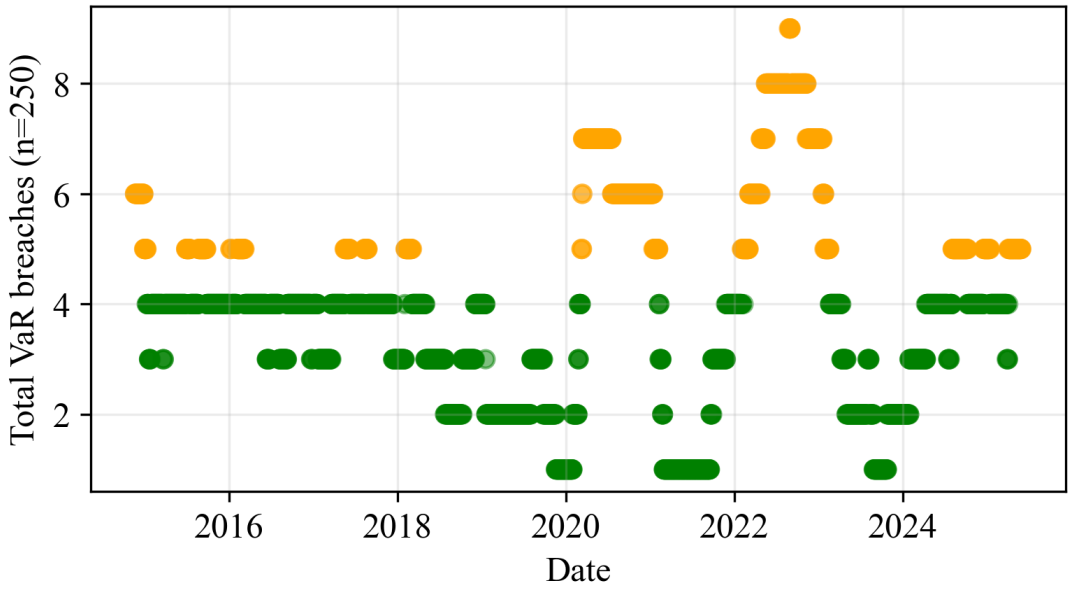

FIGURE C.4. Daily green, amber and red classifications in the FRTB Traffic Light test for $\text{VaR}_{1\%}$ estimated through WHS, where $\lambda = 0.98$.

FIGURE C.5. Daily green, amber and red classifications in the Kupiec test for $\text{VaR}_{1\%}$ estimated through HS.

FIGURE C.6. Daily green, amber and red classifications in the Kupiec test for $\text{VaR}_{1\%}$ estimated through WHS, where $\lambda = 0.995$.

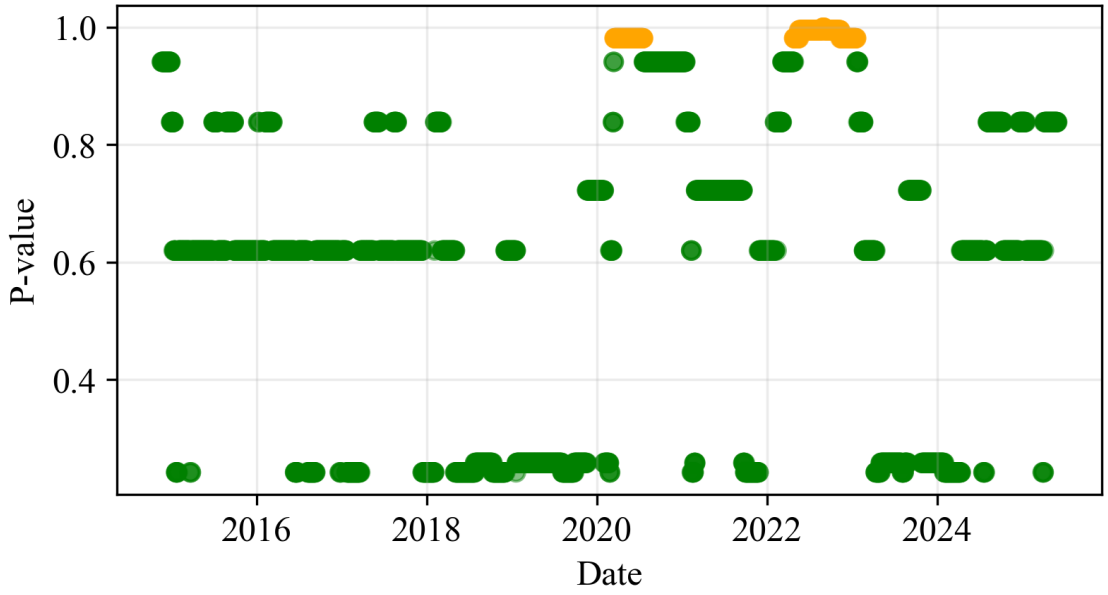

FIGURE C.7. Daily green, amber and red classifications in the Kupiec test for $\text{VaR}_{1\%}$ estimated through WHS, where $\lambda = 0.98$.

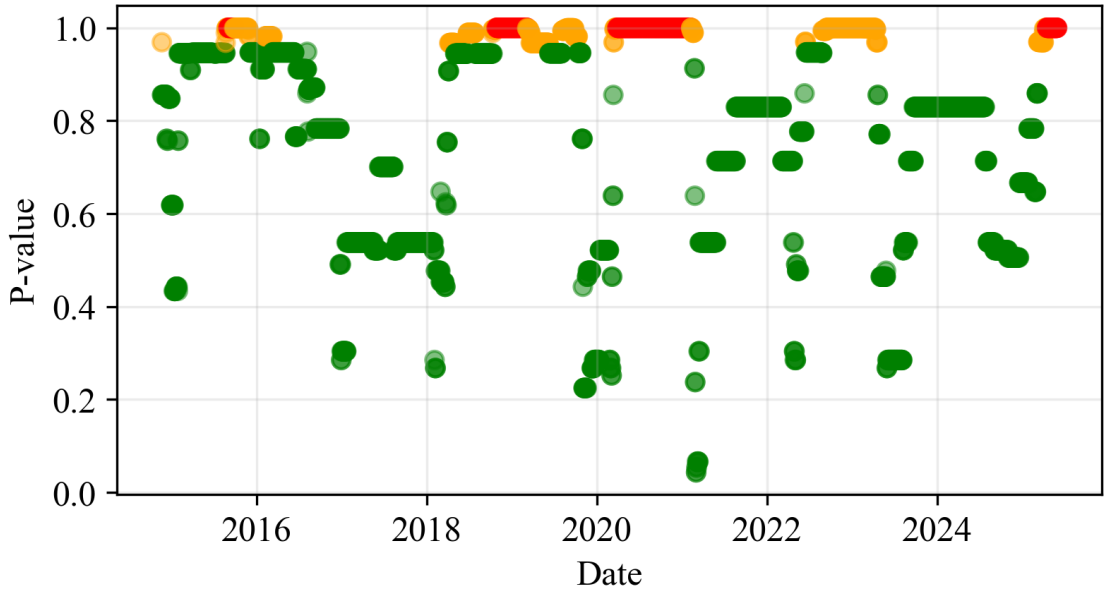

FIGURE C.8. Daily green, amber and red classifications in the Pearson $N = 4$ test for $ES_{2.5\%}$ estimated through HS.

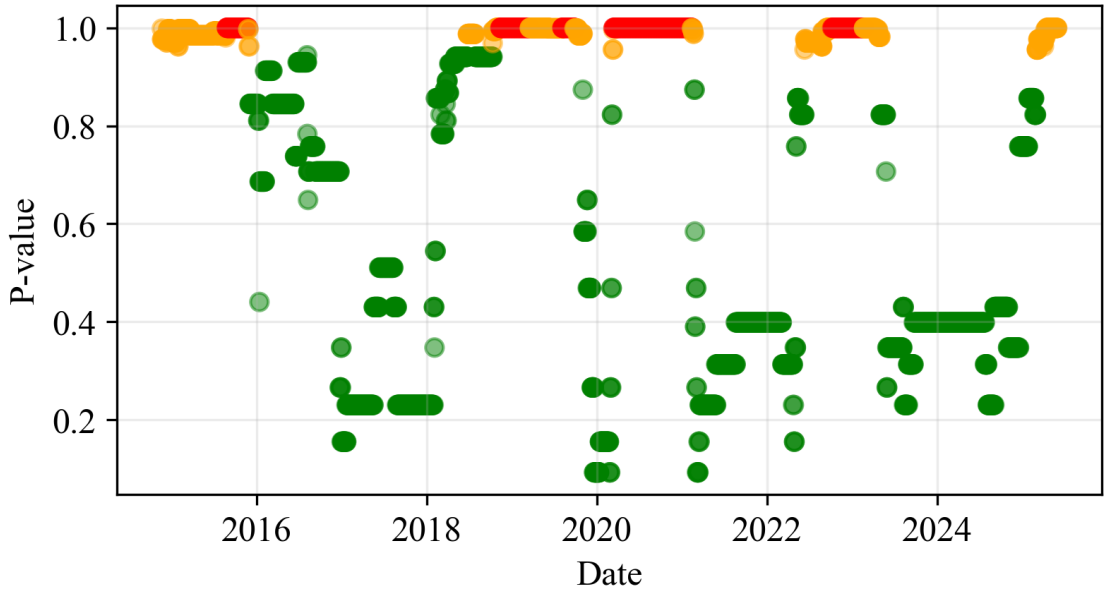

FIGURE C.9. Daily green, amber and red classifications in the Pearson $N = 8$ test for $ES_{2.5\%}$ estimated through HS.

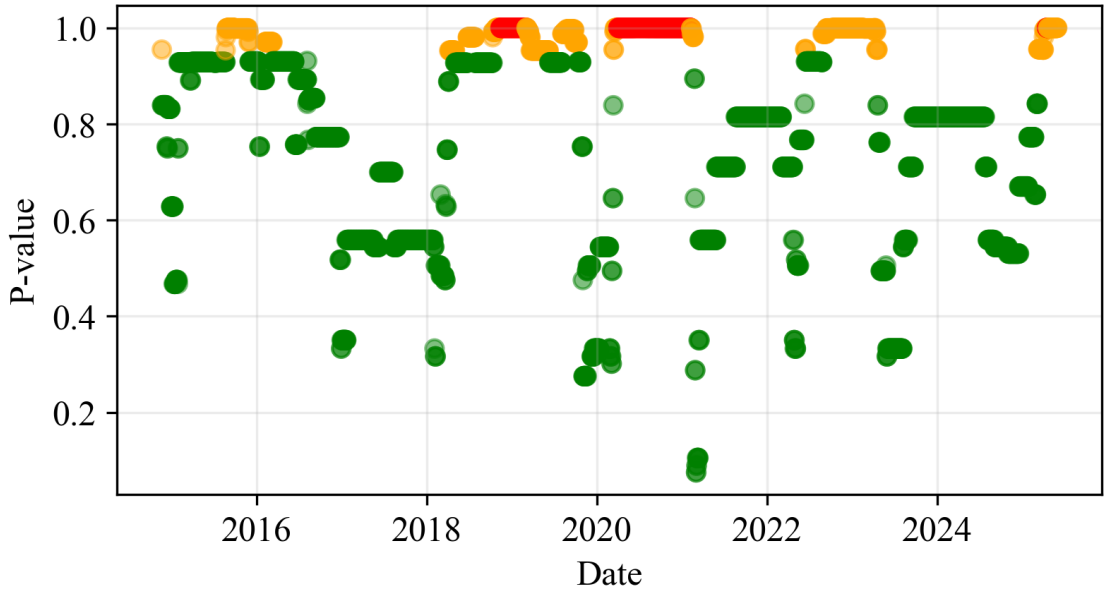

FIGURE C.10. Daily green, amber and red classifications in the Nass $N = 4$ test for $\mathrm{ES}_{2.5\%}$ estimated through HS.

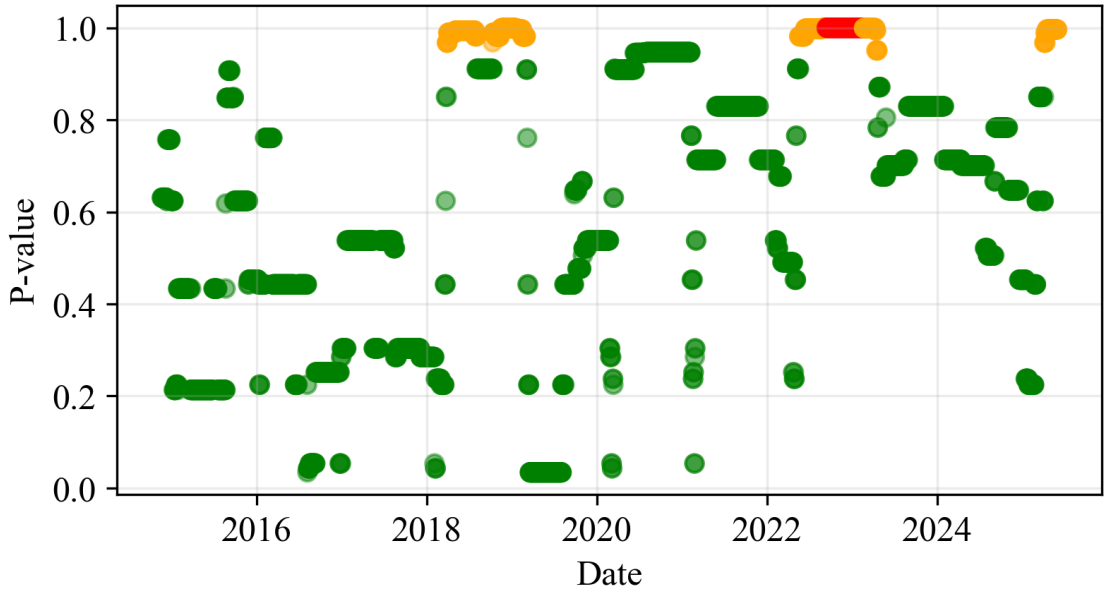

FIGURE C.11. Daily green, amber and red classifications in the Pearson $N = 4$ test for $\mathrm{ES}_{2.5\%}$ estimated through WHS, where $\lambda = 0.995$.

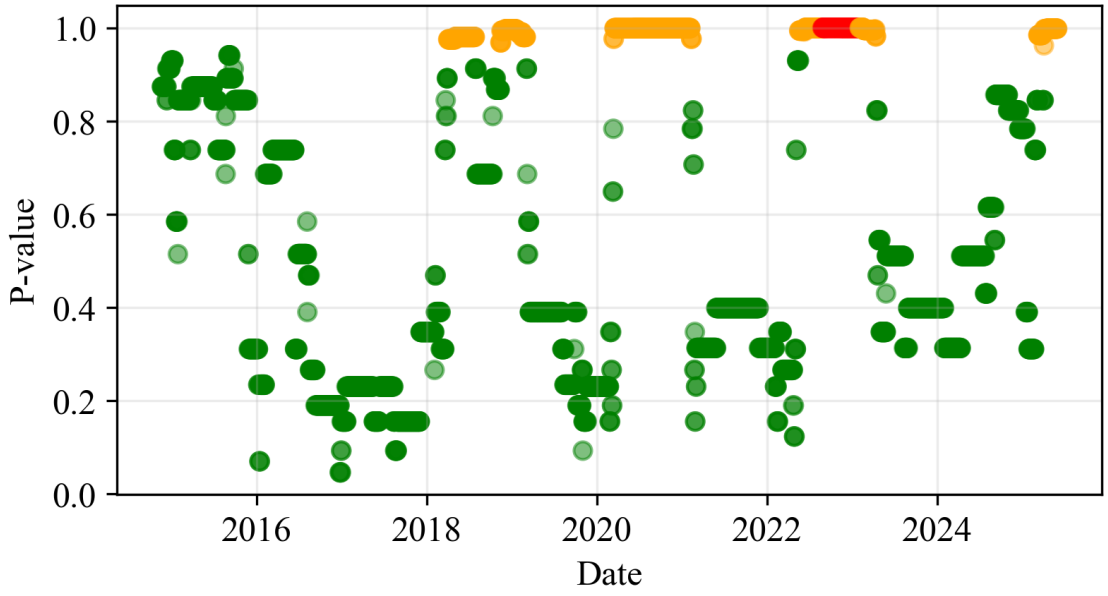

FIGURE C.12. Daily green, amber and red classifications in the Pearson $N = 8$ test for $ES_{2.5\%}$ estimated through WHS, where $\lambda = 0.995$.

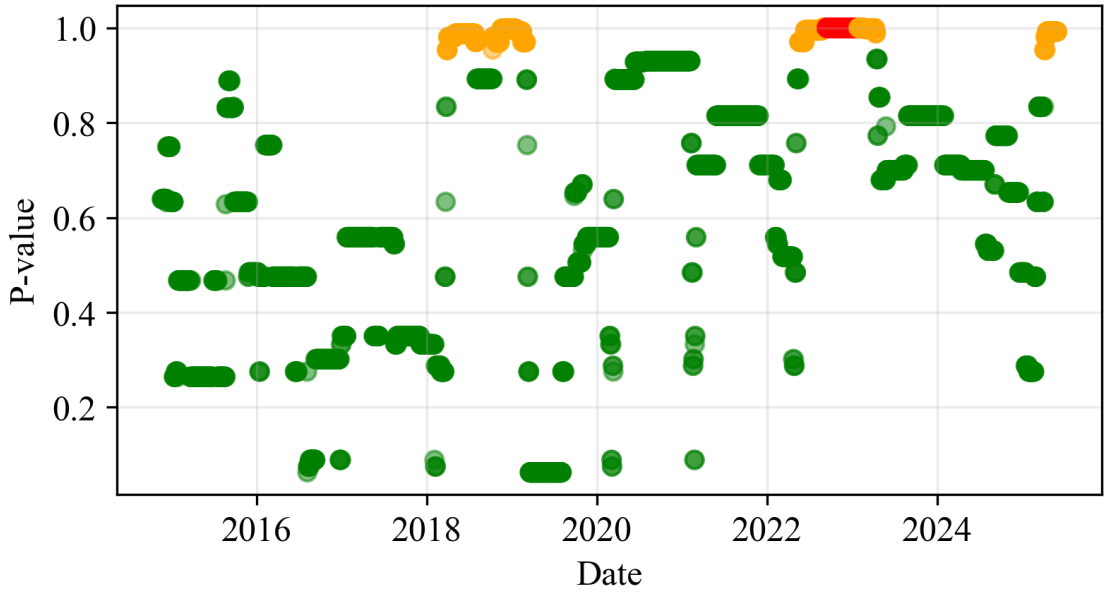

FIGURE C.13. Daily green, amber and red classifications in the Nass $N = 4$ test for $ES_{2.5\%}$ estimated through WHS, where $\lambda = 0.995$.

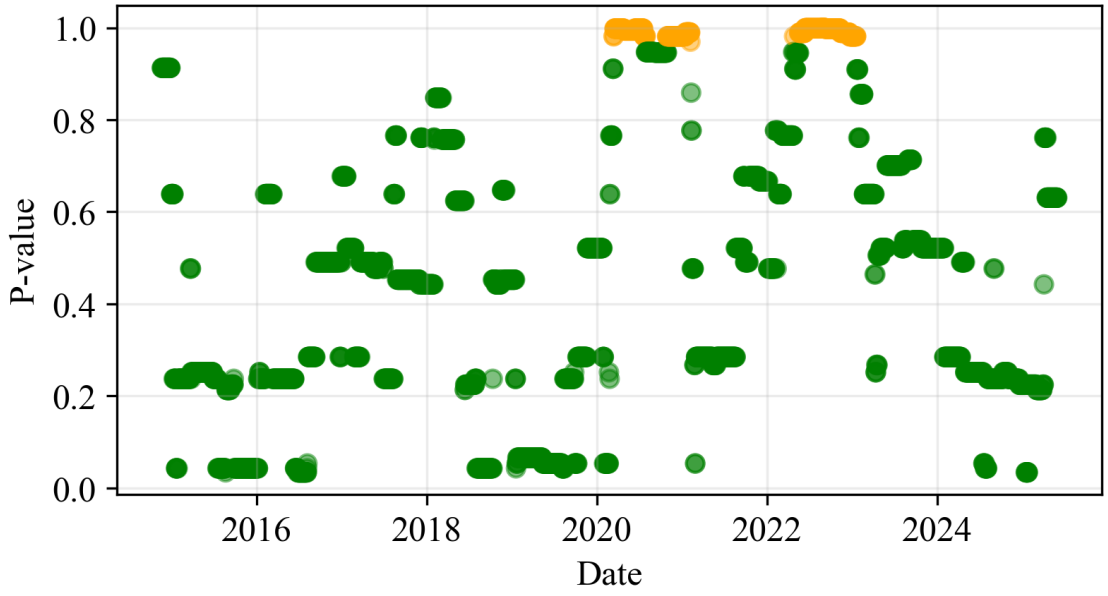

FIGURE C.14. Daily green, amber and red classifications in the Pearson $N = 4$ test for $ES_{2.5\%}$ estimated through WHS, where $\lambda = 0.98$.

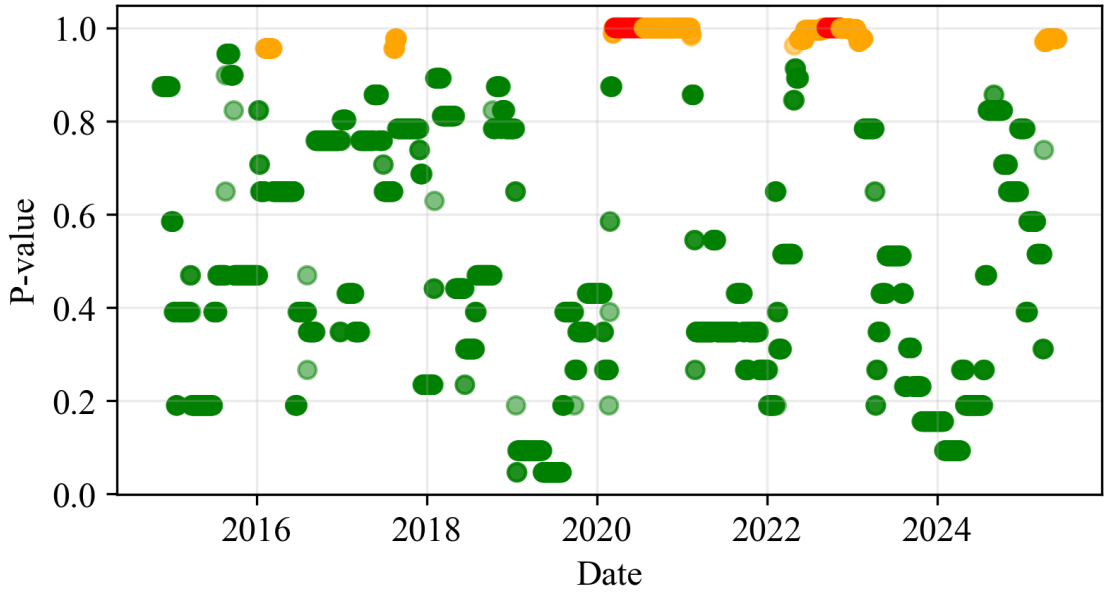

FIGURE C.15. Daily green, amber and red classifications in the Pearson $N = 8$ test for $ES_{2.5\%}$ estimated through WHS, where $\lambda = 0.98$.

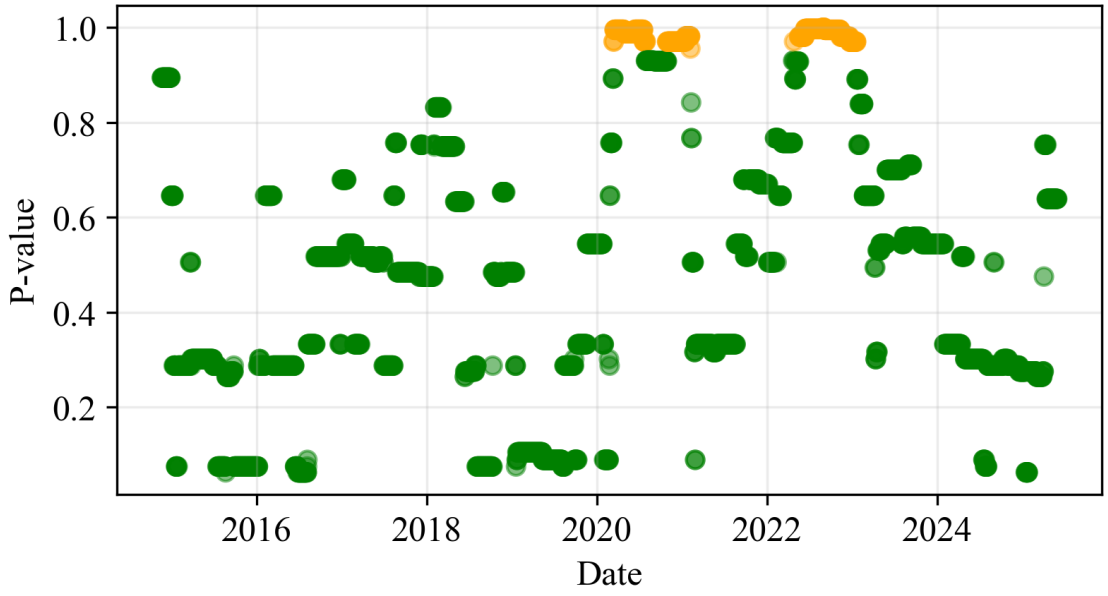

FIGURE C.16. Daily green, amber and red classifications in the Nass $N = 4$ test for $\text{ES}_{2.5\%}$ estimated through WHS, where $\lambda = 0.98$.

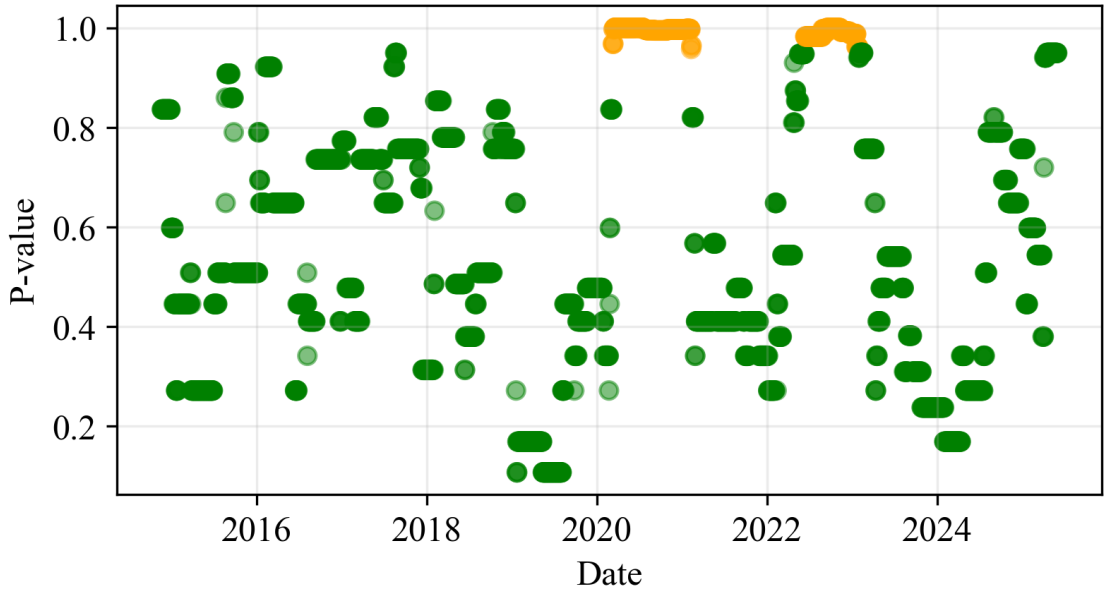

FIGURE C.17. Daily green, amber and red classifications in the Nass $N = 8$ test for $\text{ES}_{2.5\%}$ estimated through WHS, where $\lambda = 0.98$.