Indeks 371262
e-ISSN 2657-9545
ISSN 0033-2372

Przegląd Statystyczny
Statistical Review

Vol. 73 No. 2 2026

GŁÓWNY URZĄD STATYSTYCZNY
STATISTICS POLAND

INFORMATION FOR AUTHORS

Przegląd Statystyczny. Statistical Review publishes original research papers on theoretical and empirical topics in statistics, econometrics, mathematical economics, operational research, decision science and data analysis. The submitted manuscripts should significantly advance theoretical aspects of these fields or explore their practical applications. Manuscripts presenting important results of research projects are particularly welcome. Review papers, shorter papers reporting on major conferences in the field, and reviews of seminal monographs are eligible for publication based on the Editor-in-Chief's decision.

Since 1st May 2019, the journal has been publishing articles in English.

Any spelling style is acceptable as long as it is consistent within the manuscript.

All papers should be submitted to the journal through the Editorial System (https://www.editorialsystem.com/pst).

For details of the submission process and editorial requirements, please visit https://ps.stat.gov.pl/ForAuthors.

Przegląd Statystyczny
Statistical Review

Vol. 73 No. 2 2026

EDITORIAL OFFICE ADDRESS

Statistics Poland (GUS), Al. Niepodległości 208, 00-925 Warsaw, Poland

Language editing: Scientific Publications Division, Analyses and Dissemination Department, Statistics Poland
Technical editing and typesetting: Statistical Publishing Establishment, Statistical Computing Center – team supervised by Mariusz Męcina

Centrum Informatyki Statystycznej

Printed and bound by: Statistical Computing Center
Al. Niepodległości 208, 00-925 Warsaw, Poland, cis.stat.gov.pl

Website: ps.stat.gov.pl

ISSN 0033-2372
e-ISSN 2657-9545
Index 371262

Information on the sales of the journal: Statistical Computing Center
Phone no.: +48 22 608 32 10, +48 22 608 38 10

Order no. 169/2026

CONTENTS

Przegląd Statystyczny. Statistical Review, 2026, vol. 73, 2, 1–15 https://doi.org/10.59139/ps.2026.02.1
Received: 28.01.2026, revised: 11.05.2026, accepted: 15.05.2026

Higher-order Markov chains for capital market decision-making

Józef Stawicki,[a] Aleksandra Świetlicka[b]

Abstract. Analysing capital market returns is fundamental to decision-making by individual investors. Advanced methods require extensive knowledge and appropriate tools, whereas individual investors often make decisions intuitively or after a very simplified analysis. The aim of the study discussed in this paper is to present the idea of higher-order Markov chains and their models and to demonstrate that the combination of higher-order Markov chains with the technical analysis in its basic form provides support for investment decisions. This approach takes into account three aspects. The first one is the linguistic practice of observing rates of return through the construction of rate-of-return intervals (a large increase, a small decrease, no change, etc.), the second is related to investors' attitude towards risk through the aggregation of return intervals and the selection of investment strategies based on technical analysis, and the third concerns the investor's memory horizon through the construction of higher-order Markov chains.

Keywords: Markov chain, decision-making, capital market analysis
JEL: C58, F47, G17, G41

1. Introduction

The problem of decision-making under the conditions of uncertainty and risk has been analysed by many scholars. The scientific search for methods that could facilitate making investment decisions has been going on for over 100 years. It is linked to the scientific debate on preferences and probability (e.g. Keynes, 1921; Ramsey, 1928). Later, psychological aspects of decision-making were analysed (e.g. Allais, 1953; Kahneman & Tversky, 1979). Many methods have been developed, taking into account various measurement conditions and the determination of multiple criteria (Trzaskalik, 2014). In many cases, the selection of decision-supporting methods has been automated (Cinelli et al., 2020). The search for quick answers under the conditions of uncertainty and risk often leads to the simplest mechanism, i.e. 'flipping a coin' on whether the rate of return will increase or decrease.

The belief that the process of decision-making based on observation has a simple 'rise and fall' mechanism is so obvious that no effort is made to verify whether 'the coin is not fake'. Nor is the independence of the throws of this 'coin' verified. The coin can be

[a] Nicolaus Copernicus University in Toruń, Faculty of Economic Sciences and Management, Institute of Economics and Finance, Department of Applied Informatics and Mathematics in Economics, ul. Gagarina 13A, 87–100 Toruń, Poland, e-mail: stawicki@umk.pl, ORCID: https://orcid.org/0000-0002-5974-6216.
[b] Student at Faculty of Economic Sciences and Management, Institute of Economics and Finance, Nicolaus Copernicus University in Toruń, ul. Gagarina 13A, 87–100 Toruń, Poland, e-mail: 322406@stud.umk.pl, https://orcid.org/0009-0001-6550-9882.

replaced by a multidimensional dice, i.e., by analysing distributions in a finite n-dimensional state space, understanding states as small declines in rates of return, large declines in rates of return, and so on. The independence of the 'dice' throws should be replaced by the study of conditional distributions. Obviously, there are sophisticated econometric models for analysing the phenomenon of changes in the rate of return. Many of these methods incorporate behavioural determinants. There are several monographs, reprinted translations and current studies, concerning both models and decision analysis, within this popular field of research (Adamczyk-Kowalczuk, 2022; Borowski, 2014). Markov chains, including higher-order ones, have been widely used in the capital market analysis (Stawicki, 2004, 2016) and in many other analyses of economic processes, such as the analysis of business cycle test results (Podgórska & Decewicz, 2001). Modern analyses use the Markov mechanism to make decisions by means of state models in a binary-time representation (Stasiak, 2025; Stasiak et al., 2025). The proposal of an extremely simple mechanism, such as the Markov chain, allows such a model to be linked to the well-known and widely used analyses in the field of technical analysis (TA), particularly in the area of pattern analysis. These analyses are reduced to short observations, while retaining their analytical nature.

TA is the analysis of charts. Its purpose is to determine the best times to buy or sell a given security, or when to hold a decision. Alongside fundamental analysis, it is a basic tool for stock market investors. The new proposals are based on previously developed principles (Murphy, 2017). TA is essentially based on the following three basic assumptions:

- the market discounts everything, i.e. the price of a company reflects everything that is happening on the market and in the environment (microeconomic and macroeconomic situation, economic conditions, political conditions and all other information relating to a given security);
- prices are subject to trends, i.e. share prices follow specific trends, either downward (bear market) or upward (bull market). A change in a trend is clearly signalled;
- history repeats itself, i.e. using technical analysis, we examine the future based on the past, and individual formations that have occurred previously may provide information about the possible direction of change.

These charts usually refer to longer observations. The decisions are long-term. A similar analysis can be applied to subsequent daily observations. In such a case, decisions would concern a short period. It has to be remembered, though, that such a tool is only the aid to the decision-making process.

2. Higher-order Markov chains

Markov chains are a well-known tool used in economics (Ching & Ng, 2006; Decewicz, 2011; Kemeny & Snell, 1976; Podgórska et al., 2002; Stawicki, 2004; and others). A Markov process with a discrete time parameter and a discrete phase space is referred to as a Markov chain.

Let $\{Y_t\}$ be a stochastic process.

The Markov property, which is the basis for defining a Markov chain, has the following form:

$$Pr\{Y_{t+1}|Y_t, Y_{t-1}, \dots Y_1, Y_0\} = Pr\{Y_{t+1}|Y_t\}. \tag{1}$$

Let $S = \{s_1, s_2, \dots, s_r\}$ be a finite set of states in which process $\{Y_t\}$ is represented by observations $\{y_t\}$. We will define set S in an abbreviated form:

$$S = \{1, 2, \dots, r\}.$$

We will record our observations of process $\{Y_t\}$ as a sequence $\{y_t\}$. For example, the observation sequence presented in the next section will be the daily rate of return on securities listed on the stock exchange.

A Markov chain is defined by a sequence of stochastic matrices in the following form:

$$\boldsymbol{P}(t) = \left[p_{ij}(t)\right]_{r\times r}, \tag{2}$$

i.e. matrices with positive elements and satisfying additional conditions expressed by:

$$\forall_t \forall_i \ \sum_j p_{ij}(t) = 1, \tag{3}$$

where $p_{ij}(t) = Pr\{Y_t = j|Y_{t-1} = i\}$ is a conditional probability.
By denoting the vector of unconditional distribution of random variable Y_t with $\boldsymbol{D}_t$, i.e.

$$\boldsymbol{D}_t = [d_{1t}, \quad d_{2t}, \quad \cdots, \quad d_{rt}], \text{ where } d_{it} = Pr\{Y_t = i\}, \tag{4}$$

we determine the probability with which the process reaches the phase state in time *t*. The components of vector $\boldsymbol{D}_t$ satisfy the following conditions:

$$\forall_t \forall_i \quad d_{it} \geq 0, \tag{5}$$

and

$$\forall_t \quad \sum_i d_{it} = 1. \tag{6}$$

The dependence between unconditional distributions of random variables Y_t and Y_{t-1} is expressed by the formula resulting from the theorem on the total probability:

$$\boldsymbol{D}_t = \boldsymbol{D}_{t-1} \cdot \boldsymbol{P}(t). \tag{7}$$

Matrices $\boldsymbol{P}(t) = \left[p_{ij}(t)\right]_{r\times r}$ reflect the mechanism of changes in the distribution of the analysed random variable Y_t over time.

A Markov chain $\{Y_t, \quad t \in N\}$ with a phase space $S = \{1,\ 2,\ \ldots,\ r\}$ is called *a homogeneous Markov chain* if the conditional probabilities $p_{ij}(t)$ of transition from state i to state j within a time unit, i.e. in the time period from $(t-1)$ to t, do not depend on the choice of the moment t, that is:

$$\forall_t\ p_{ij}(t) = p_{ij}. \tag{8}$$

In the case of a homogeneous Markov chain, the dependence (7) takes the following form:

$$\boldsymbol{D}_t = \boldsymbol{D}_{t-1} \cdot \boldsymbol{P}. \tag{9}$$

If the Markov property, which is the basis for defining a higher-order Markov chain, has the following form:

$$Pr\{Y_{t+1}|Y_t, Y_{t-1}, \ldots, Y_1, Y_0\} = Pr\{Y_{t+1}|Y_t, Y_{t-1}, \ldots, Y_{t-k}\}, \tag{10}$$

a higher-order Markov chain order k can be represented by matrix $\boldsymbol{Q}$, where each row represents a k-element variation with repetitions from an r-element set of process states, and each column represents one of the states of this process. This matrix has dimensions $r^k \times r$.

By denoting the vector of unconditional distribution of random variable Y_t with $\boldsymbol{D}_t$, i.e.

$$\boldsymbol{D}_t = [d_{1t}, d_{2t}, \cdots, d_{rt}], \text{ where } d_{it} = Pr\{Y_t = i\},$$

and the vector of distribution of possible histories in the last k periods with $\boldsymbol{H}_t = \left[h_{1t}, h_{2t}, \ldots, h_{r^k t}\right]$, i.e.

$$h_{lt} = Pr\{[Y_{t-k}, Y_{t-k+1}, \ldots, Y_{t-1}]\} \tag{11a}$$

or

$$h_{lt} = Pr\{i(k)_{t-k}, i(k-1)_{t-k+1}, \dots, i(1)_{t-1}\}, \tag{11b}$$

we determine the probability with which the process reaches phase state i at time t, in the following form:

$$\boldsymbol{D}_t = \boldsymbol{H}_t \cdot \boldsymbol{Q}. \tag{12}$$

This history is observed in the form of a vector of realised state sequences, i.e.

$$\boldsymbol{H}_t = [0, 0, \dots, 0, 1, 0, \dots, 0].$$

Matrix $\boldsymbol{Q}$ contains conditional distributions, where the condition is the history in the last k periods.

The observation of the process is based on microdata. The parameters of matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ are obtained using the maximum likelihood estimator formula:

$$\hat{p}_{(i_k, i_{k-1}, \dots, i_1), j} = \frac{\sum_{t=k+1}^{T} v_{(i_k, i_{k-1}, \dots, i_1), j}(t)}{\sum_{t=k+1}^{T} n_{(i_k, i_{k-1}, \dots, i_1)}(t)}, \tag{13}$$

where

$$v_{(i_k, i_{k-1}, \dots, i_1), j}(t) = \begin{cases} 1, if\ at\ time\ t\ after\ history\ (i_k, i_{k-1}, \dots, i_1), state\ j\ occurred, \\ 0, otherwise \end{cases}$$

and

$$n_{(i_k, i_{k-1}, \dots, i_1)}(t) = \begin{cases} 1, if\ at\ time\ t, the\ observed\ history\ is\ (i_k, i_{k-1}, \dots, i_1), \\ 0, otherwise \end{cases}$$

Quantity $\sum_{t=k+1}^{T} n_{(i_k, i_{k-1}, \dots, i_1)}(t)$ from formula (13) determines the number of all the observed histories $(i_k, i_{k-1}, \dots, i_1)$, while quantity $\sum_{t=k+1}^{T} v_{(i_k, i_{k-1}, \dots, i_1), j}(t)$ determines the number of these histories followed by state j.

If the history is a single state i, the process becomes a classical Markov chain.

To test the hypothesis that a given row of matrix $\boldsymbol{Q}$ defining conditional transitions from history $(i_k, i_{k-1}, \dots, i_1)$ to individual states j is equal to the established

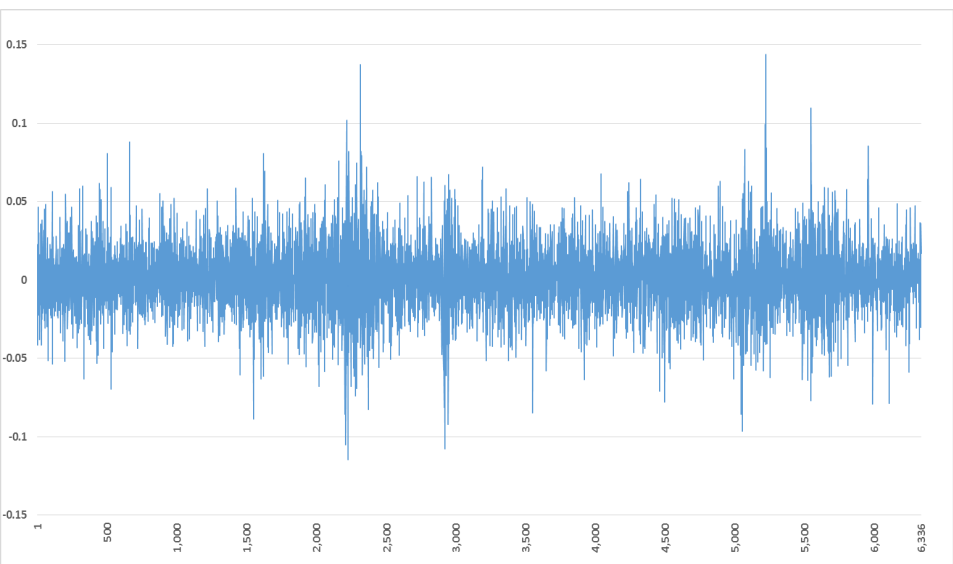
probabilities, we use a chi-squared test with (r–1) degrees of freedom, using the statistic:

$$\chi^2 = \sum_{t=k+1}^{T} n_{(i_k, i_{k-1}, \ldots, i_1)}(t) \cdot \sum_{j=1}^{r} \frac{\left(\hat{p}_{(i_k, i_{k-1}, \ldots, i_1), j} - p_{(i_k, i_{k-1}, \ldots, i_1), j}\right)^2}{p_{(i_k, i_{k-1}, \ldots, i_1), j}}. \quad (14)$$

3. Decision-making of individual investors: the example of Orlen company

We assume that investors on the stock market make decisions on the basis of the available observations of the rate of return on an asset. These observations cover a long period, but the decision is made on the basis of the last observation or several recent observations. The length of time covered by the memory is a specific attitude of the investor. Another specific feature of the investor is the way they view the observations. Investors may only be interested in the direction of changes in the rate of return or the ranges in which the rate of return is observed. The number of ranges and their sizes are determined by the decision-maker. In our study, the ranges will be determined according to specific rules.

Figure 1. Daily rate of return – closing price for Orlen (January 2000 – April 2025)



Source: authors' work.

Figure 1 shows the rate of return quotations that will be the basis for the analysis.

If the decision-maker only analyses stock market ups and downs, we consider a Markov chain with two states $s_1 = up, s_2 = down$. Therefore $\boldsymbol{D}_t = (u_t, d_t)$. For the entire studied period,

$$D = (0.5066, 0.4934).$$

It can be said that this vector represents the coin-tossing accurately. The Chi-square statistics is:

$$\chi^2 = 0.5347 \text{ where chi-square critical value is } \chi^2_{0.05} = 3.841.$$

The corresponding p-value is 0.3011.

The first-order Markov chain model shows that conditional distributions are also distributions $\overline{p} = (0.5\ , 0.5)$.

The matrix for this model has the following form:

$$\boldsymbol{P} = \begin{bmatrix} 0.4944 & 0.5056 \\ 0.5176 & 0.4824 \end{bmatrix},$$

and the corresponding chi-squared statistics are: for the first row (u is the condition of the growth of the rate of return at the previous time) $\chi^2 = 0.384$, and for the second row (d is the condition of the decrease in the rate of return at the previous time) $\chi^2 = 3.617$, where the chi-square critical value is $\chi^2_{0.05} = 3.841$. The corresponding p-values are 0.5354 and 0.0572.

The second-order Markov chain model is represented by matrix $\boldsymbol{Q}$:

$$\boldsymbol{Q} = \begin{matrix} uu \\ ud \\ du \\ dd \end{matrix} \begin{bmatrix} 0.4580 & 0.5420 \\ 0.5007 & 0.4993 \\ 0.5266 & 0.4734 \\ 0.5348 & 0.4652 \end{bmatrix}.$$

Additionally, the form of the condition was marked with matrix $\boldsymbol{Q}$. Chi-squared statistics for each row of matrix $\boldsymbol{Q}$ and the p-value are presented in Table 1 below.

Table 1. Chi-square statistics and p-value for the second-order Markov chain

Condition	Chi-square statistic	p-value
uu	10.252	0.0014
ud	0.003	0.9584
du	4.150	0.0416
dd	6.688	0.0097

Source: authors' work.

The chi-square critical value is $\chi^2_{0.05} = 3.841$. Therefore, the conditions (uu, du, dd) significantly change the probabilities of transition to states u and d.

This is even more evident when looking at the history going back three times. The matrix representing the third-order Markov chain model has the following form:

$$\boldsymbol{Q} = \begin{matrix} uuu \\ uud \\ udu \\ udd \\ duu \\ dud \\ ddu \\ ddd \end{matrix} \begin{bmatrix} 0.4537 & 0.5463 \\ 0.4823 & 0.5177 \\ 0.5251 & 0.4749 \\ 0.5239 & 0.4761 \\ 0.4511 & 0.5489 \\ 0.5253 & 0.4747 \\ 0.5289 & 0.4711 \\ 0.5510 & 0.4490 \end{bmatrix}.$$

For each row of matrix $\boldsymbol{Q}$, the p-values for the chi-square statistics are as presented in Table 2.

Table 2. Chi-square statistics and p-value for third-order Markov chain

Condition	Chi-square statistic	p-value
uuu	5.556	0.0184
uud	0.958	0.3277
udu	1.810	0.1785
udd	1.624	0.2026
duu	7.243	0.0071
dud	1.720	0.1897
ddu	2.371	0.1236
ddd	6.522	0.0107

Source: authors' work.

The chi-square critical value is $\chi^2_{0.05} = 3.841$. Therefore, the conditions (uuu, duu, ddd) significantly change the probabilities of the transition to state u and state d.

Seeing the last three quotations as upwards, the decision-maker should take into account different probabilities of the occurrence of states *u* and *d*.

The second example, based on the same observations of the Orlen company's return quotation process, is constructed using Markov chain states as intervals in which the observed rate of return may be included. These states are defined as: s1– an increase, s2 – no change, and s3 – a decrease. The intervals for these states are determined by the decision-maker. Decision-makers differ from one another and establishing the same ranges for everyone is extremely difficult. The intervals below have been set so that each of them contains 1/3 of the observations from the entire period of the studied rate of return. This example refers to 'tossing a coin' or rather to 'throwing a three-sided dice'.

Table 3 below describes the states.

Table 3. Intervals and distribution for the entire observation period

	Return intervals for quotations	Number of observations	Unconditional probability
s1	to –0.0076	2,099	0.3314
s2	(–0.0076; 0.00818)	2,109	0.3328
s3	from 0.00818	2,128	0.3358

Source: authors' work.

A second-order Markov chain model is defined by matrix $\boldsymbol{Q}$:

$$\boldsymbol{Q} = \begin{matrix} \text{s1,s1} \\ \text{s1,s2} \\ \text{s1,s3} \\ \text{s2,s1} \\ \text{s2,s2} \\ \text{s2,s3} \\ \text{s3,s1} \\ \text{s3,s2} \\ \text{s3,s3} \end{matrix} \begin{bmatrix} 0.3246 & 0.2768 & 0.3986 \\ 0.3233 & 0.3262 & 0.3505 \\ 0.3253 & 0.2999 & 0.3748 \\ 0.3170 & 0.3511 & 0.3319 \\ 0.2925 & 0.3592 & 0.3483 \\ 0.3462 & 0.3805 & 0.2733 \\ 0.3433 & 0.3188 & 0.3379 \\ 0.3455 & 0.3581 & 0.2964 \\ 0.3647 & 0.3265 & 0.3088 \end{bmatrix}.$$

Therefore, the estimated matrix does not yield equal distributions identical to the distribution:

$$\bar{p} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right).$$

For each conditional distribution, the p-values for the chi-square test are presented in Table 4 below.

Table 4. A chi-square statistic and a p-value for the second-order Markov chain for the three states

Condition	Chi-square statistic	p-value
s1,s1	15.5739	0.0004
s1,s2	0.8822	0.6433
s1,s3	6.5141	0.0385
s2,s1	1.1822	0.5537
s2,s2	5.6408	0.0596
s2,s3	12.5923	0.0018
s3,s1	0.7302	0.6941
s3,s2	4.5534	0.1026
s3,s3	3.3294	0.1892

Source: authors' work.

Observing the presented process two periods back, the distributions were significantly different from the uniform distribution only for two conditions.

The third example refers directly to the technical analysis used by decision-makers investing in the stock market. Technical analysis is a method of predicting the movement of stock prices by analysing charts of historical prices. Higher-order Markov chains fit into this understanding of price process analysis (return on investment).

In this example, six states of the Markov chain were assumed and described in linguistic terms. A state is understood as a range in which the rate of return will fall or rise. Table 5 presents the states in the studied example.

Table 5. Description of the Markov chain states

State	Linguistic description	Interval
s1	A large drop in the rate of return	–0.02 or less
s2	A drop in the rate of return	(–0.02 , –0.01)
s3	A small drop in the rate of return	(–0.01 , 0.00)
s4	A small increase in the rate of return	(0.00 , 0.01)
s5	An increase in the rate of return	(0.01 , 0.02)
s6	A large increase in the rate of return	0.02 or more

Source: authors' work.

Each decision-maker establishes an individual interval. This depends on the decision-maker's individual characteristics, their attitude towards risk (inclination or aversion), their knowledge and experience, and external circumstances influencing the course of economic processes. The proposed ranges are illustrative only and do not represent a specific decision-maker. The adoption of specific ranges generates the distribution of the rate of return over the entire period and the behaviour of the process within the examined time period. One such proposal is presented in Table 6.

Table 6. Probability distribution by the state of the rate of return on Orlen quotations

State	Probability
s1	0.141730
s2	0.143308
s3	0.224432
s4	0.195391
s5	0.143782
s6	0.151357

Source: authors' work.

Understanding the states is an individual characteristic of the investor. It is the investor who determines the ranges described by, e.g. 'a small increase in the rate of return' or 'a small decrease in the rate of return'.

We present the transition matrix for a second-order Markov chain and the chi-square test for each conditional distribution in Table 7.

Table 7. Conditional distribution and the p-values for a second-order Markov chain

	s1	s2	s3	s4	s5	s6	p-value for a chi-square test
s1s1	0.2273	0.0584	0.1234	0.1883	0.1623	0.2403	0.0000
s1s2	0.1626	0.1301	0.1220	0.1789	0.1951	0.2114	0.0401
s1s3	0.1761	0.1056	0.1901	0.1901	0.1831	0.1549	0.4295
s1s4	0.1429	0.1429	0.1565	0.1701	0.1769	0.2109	0.1524
s1s5	0.1597	0.1345	0.2017	0.1849	0.1597	0.1597	0.9673
s1s6	0.1250	0.1369	0.1726	0.1726	0.1310	0.2619	0.0053
s2s1	0.1773	0.1560	0.2057	0.1489	0.1489	0.1631	0.6459
s2s2	0.1635	0.1346	0.1923	0.1635	0.2019	0.1442	0.5636
s2s3	0.1438	0.1918	0.1781	0.2192	0.1233	0.1438	0.4536
s2s4	0.1374	0.1209	0.2143	0.1923	0.1813	0.1538	0.7694
s2s5	0.1324	0.1691	0.1838	0.2279	0.1397	0.1471	0.7787
s2s6	0.1579	0.1053	0.1729	0.2331	0.1353	0.1955	0.3156
s3s1	0.1438	0.1438	0.1688	0.2250	0.1250	0.1938	0.3890
s3s2	0.1807	0.1145	0.1687	0.2229	0.1687	0.1446	0.2646
s3s3	0.1240	0.1240	0.1983	0.2273	0.1818	0.1446	0.3396
s3s4	0.0781	0.1641	0.2500	0.2188	0.1563	0.1328	0.0648
s3s5	0.0978	0.1250	0.2283	0.2663	0.1250	0.1576	0.1466
s3s6	0.1631	0.1560	0.2340	0.1915	0.1206	0.1348	0.9166
s4s1	0.1812	0.2101	0.1594	0.1667	0.1087	0.1739	0.0589
s4s2	0.1397	0.1285	0.1564	0.2626	0.1229	0.1899	0.0645
s4s3	0.1401	0.1440	0.2335	0.2179	0.1634	0.1012	0.3331
s4s4	0.1076	0.1659	0.2601	0.1973	0.1256	0.1435	0.4800
s4s5	0.0990	0.2178	0.2178	0.2178	0.0941	0.1535	0.0129
s4s6	0.1585	0.1402	0.2622	0.1768	0.1037	0.1585	0.6142
s5s1	0.1770	0.1858	0.1593	0.1150	0.1681	0.1947	0.0709
s5s2	0.1702	0.1206	0.1986	0.2270	0.1560	0.1277	0.6808
s5s3	0.1173	0.1676	0.2179	0.2346	0.1508	0.1117	0.4340
s5s4	0.1277	0.1702	0.2340	0.1809	0.1809	0.1064	0.3260
s5s5	0.1284	0.1743	0.2477	0.1651	0.1468	0.1376	0.8782
s5s6	0.1450	0.1679	0.2137	0.1985	0.1221	0.1527	0.9571
s6s1	0.1544	0.1342	0.1879	0.1879	0.1342	0.2013	0.5881
s6s2	0.1860	0.1318	0.1938	0.2171	0.1473	0.1240	0.6380
s6s3	0.1497	0.1551	0.2032	0.2674	0.1230	0.1016	0.1006
s6s4	0.1455	0.2061	0.1818	0.1576	0.2000	0.1091	0.0267
s6s5	0.1982	0.1081	0.1802	0.1892	0.1261	0.1982	0.2784
s6s6	0.2138	0.1887	0.2075	0.1321	0.1195	0.1384	0.0296

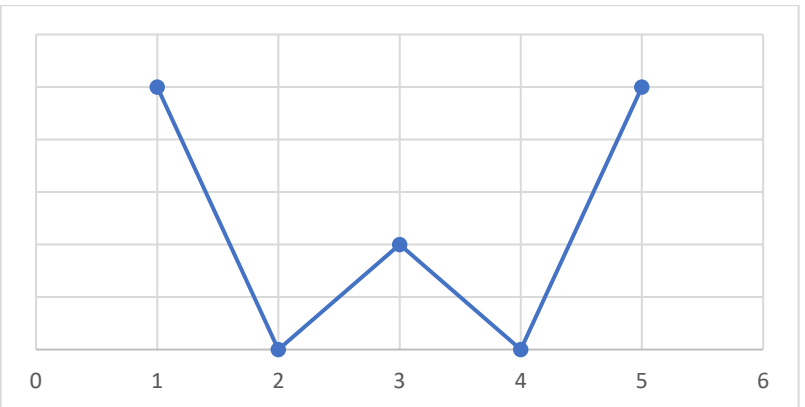
Source: authors' work.

By looking only at the last two observations, we can check the chances of an increase or a decrease in the price of the observed asset before making an investment decision. In the case under study, if the last two observations were significantly upward,

a decline should be expected with a probability of over 0.6, or more precisely, the sum of the transition probabilities from state (s6,s6) to states s1, s2, s3 is 0.2138+0.1887+ +0.2075 = 0.61.

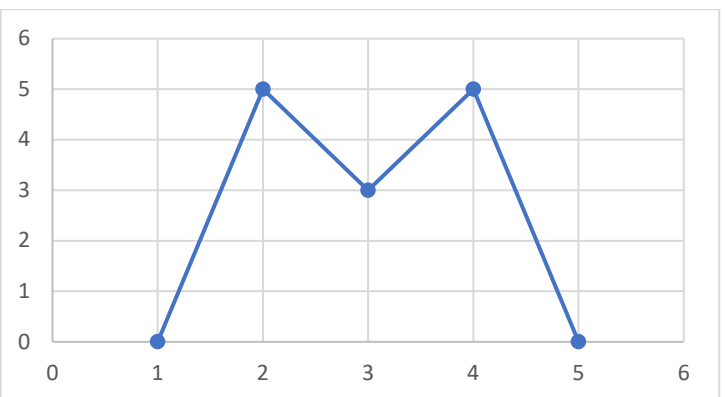
The analysis of somewhat complex patterns in technical analysis mentioned in the introduction requires a large number of observations. To examine the consequences of a double bottom or double top pattern in the short term, it is sufficient to know the last four observations, which become a condition in the fourth-order Markov chain (Figures 2 and 3).

Figure 2. Double bottom pattern



Source: authors' work.

Figure 3. Double top pattern



Source: authors' work

For the rate of return change process, this will be a sequence of successive states observed in history:

- a double bottom pattern, i.e. a large decline, a small increase, a small decline, a large increase;
- a double top pattern, i.e. a large increase, a small decline, a small increase, a large decline.

The fuzzy nature of such patterns can be assumed, including a similar behaviour of the rate of return in terms of declines and increases in the assumed sequence. The sequence of the last observed states, namely s1, s4, s3 and s6, is representative of such a sequence. For the double bottom pattern, the set of state sequences was assumed {s1,s4,s3,s6; s1,s5,s2,s6; s1,s4,s2,s6; s1,s5,s3,s6; s2,s4,s3,s5}. Similarly, for the double top pattern, the set of state sequences was assumed {s6,s2,s4,s1; s6,s3,s5,s2; s6,s3,s4,s1; s6,s2,s5,s1; s5,s3,s4 s2}.

Analysing the fourth-order Markov chain with states as in the last example and estimating conditional probabilities, we obtain the results presented in Table 8.

Table 8. Selected conditional probabilities for a fourth-order Markov chain

	s1	s2	s3	s4	s5	s6
double bottom pattern	0.2105	0.0000	0.2632	0.2105	0.1579	0.1579
double top pattern	0.0000	0.2500	0.1000	0.2500	0.1500	0.2500

Source: authors' work.

By comparing the obtained conditional distribution with the distribution presented in Table 6 and using the chi-square test, we obtain the following p-values: 0.4594 for the double bottom pattern, and 0.3141 for the double top pattern.

To support decision-making, any sequence of the observed states can be used to estimate the conditional distribution. For an observation period that is too short, estimating the conditional distribution may prove undesirable. However, if the investor's memory horizon is not too distant, the conditional distribution becomes a supportive forecast.

For example, if the observations at times t-4, t-3, t-2 and t-1 are as follows: s1, s4, s3, s6, then a vector of conditional probabilities takes the form of a double bottom pattern. The conditional distribution in this situation is presented in Table 9.

Table 9. Special case of conditional probability for a fourth-order Markov chain

	s1	s2	s3	s4	s5	s6
s1, s4, s3, s6	0.1667	0.0000	0.1667	0.5000	0.0000	0.1666

Source: authors' work.

This means that investors can expect growth rather than decline (the sum of the probabilities of transitions to states s4, s5, s6), and if there is a decline, it is most likely to be minor (the probability of transition to state s1 and s3). The small number of such observed sequences of states does not allow far-reaching conclusions. Comparing this distribution with the pattern in Table 6 gives a p-value of 0.5619.

4. Conclusions

The proposed approach to investment decisions is based on a very simple Markov chain mechanism. To utilise this 'old-school' tool, we adopted a personalised approach to capital market decision-making. The numerical intervals of the rate of return were determined by incorporating a linguistic description of the changes in the rate. This allows an individualised approach for each investor. The decision-maker's memory horizon for the observed rates of return has also been taken into account, and the assumption of constant transition probabilities has been adopted in the analysis. Modifying the transition probability matrix by incorporating the characteristics for specific periods (e.g. macroeconomic observations, company developments or specific global situations influencing stock prices) will support investment decisions more effectively. The proposed simple approach to analysing rates of return can be used by individual investors in a situation where they do not see the need for a sophisticated analysis and want to make decisions quickly. This method may prove useful in an algorithmic trading system based on models with binary-time representation.

References

Adamczyk-Kowalczuk, M. (2022). *Behawioralne determinanty decyzji inwestycyjnych na rynku kapitałowym*. Polskie Wydawnictwo Ekonomiczne.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica*, *21*(4), 503–546. https://doi.org/10.2307/1907921.

Borowski, K. (2014). *Finanse behawioralne. Modele*. Difin.

Ching, W. K., & Ng, M. K. (2006). *Markov Chains. Models, Algorithms and Applications*. Springer Science + Business Media. https://doi.org/10.1007/0-387-29337-X.

Cinelli, M., Kadziński, M., Gonzalez, M., & Słowiński, R. (2020). How to support the application of multiple criteria decision analysis? Let us start with a comprehensive taxonomy. *Omega*, *96*, 1–22. https://doi.org/10.1016/j.omega.2020.102261.

Decewicz, A. (2011). *Probabilistyczne modele badań operacyjnych*. Oficyna Wydawnicza SGH.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185.

Kemeny, J. G., Snell, J. L. (1976). *Finite Markov Chains*. Springer-Verlag.

Keynes, J. M. (1921). *A Treatise on Probability*. MacMillan and Co.

Murphy, J. J. (2017). *Analiza techniczna rynków finansowych*. Księgarnia Maklerska.

Podgórska, M., & Decewicz, A. (2001). Modele Markowa w analizie wyników testu koniunktury. In E. Adamowicz, M. Męczarski, M. Podgórska (Eds.), *Analiza tendencji rozwojowych w polskiej gospodarce na podstawie testu koniunktury. Metody i wyniki* (pp. 119–155). Szkoła Główna Handlowa.

Podgórska, M., Śliwka, P., Topolewski, M., & Wrzosek, M. (2002). *Łańcuchy Markowa w teorii i w zastosowaniach*. Szkoła Główna Handlowa.

Ramsey, F. P. (1928). Truth and Simplicity. *The British Journal for the Philosophy of Science*, *58*(3), 379–386.

Stasiak, M. D. (2025). Algorithic Trading System with Adaptive State Model of a Binary-Temporal Representation. *Risks*, *13*(8), 1–12. https://doi.org/10.3390/risks13080148.

Stasiak, M. D., Staszak, Ż., Siwek, J., & Wojcieszak, D. (2025). Application of State Models in Binary –Temporal Representation for the Prediction and Modelling of Crude Oil Prices. *Energies*, *18*(3), 1–14. https://www.mdpi.com/1996-1073/18/3/691.

Stawicki, J. (2004). *Wykorzystanie łańcuchów Markowa w analizie rynku kapitałowego*. Wydawnictwo UMK.

Stawicki, J. (2016). Using the First Passage Times in Markov Chain Model to Support Financial Decisions on Stock Exchange. *Dynamic Econometric Models*, *16*(1), 37–47. https://doi.org/10 .12775/DEM.2016.003.

Trzaskalik, T. (2014). Wielokryterialne wspomaganie decyzji. Przegląd metod i zastosowań. *Zeszyty Naukowe Politechniki Śląskiej*, *74*, 239–263.

Przegląd Statystyczny. Statistical Review, 2026, vol. 73, 2, 16–34 https://doi.org/10.59139/ps.2026.02.2
Received: 13.03.2026, revised: 2.06.2026, accepted: 2.06.2026

Yield curve forecasting using the Nelson-Siegel model. A comparison of ARIMA, VAR and Random Forest approaches – evidence from the USA

Maciej M. Olszewski[a]

Abstract. The aim of this article is to compare different approaches to forecasting the US yield curve factors derived using the Nelson-Siegel (NS) model. Using daily US swap yield data from 1990 to 2026, we assess the Autoregressive Integrated Moving Average (ARIMA), Vector Autoregression (VAR) and Random Forest (RF) models in a 1-day-ahead and 1- to 20-day-ahead forecasting competition. The principal finding of this study is that ARIMA significantly outperforms the RF in forecasting the NS model factors, as does VAR, although only in terms of the Level and Curvature factors. The results of this study thus suggest that the use of machine learning methods, in the case of the US yield curve, is not always superior.

Keywords: yield curve, forecasting, Nelson-Siegel model, machine learning
JEL: C53, C58, E43

1. Introduction

The yield curve is a key element of financial markets which carries information about the state of the economy. For that reason, a good understanding of its dynamics and determinants facilitates decision-making processes in areas such as monetary policy and risk management or when developing trading strategies.

The aim of this article is to compare different approaches to forecasting the US yield curve factors based on the Nelson-Siegel (1987, NS) model, which decomposes the yield curve into three factors: Level (L), Slope (S) and Curvature (C). The Level factor describes the level of long-term interest rates, the Slope factor represents the difference between the level of short- and long-term interest rates, while the Curvature factor is responsible for yield curve convexity.

1.1. Classical approaches

The NS model is widely applied due to its simplicity and economic interpretability. It can be used to forecast the entire yield curve. As proposed by Diebold and Li (2006), this can be done in two steps: forecasting the NS factors and reconstructing the future shape of the yield curve. There are numerous ways of forecasting the NS factors, including the use of Autoregressive Integrated Moving Average (ARIMA) or

[a] Student at SGH Warsaw School of Economics, al. Niepodległości 162, 02–554 Warszawa, Poland, e-mail: maciek.m.olszewski@gmail.com, ORCID: https://orcid.org/0009-0006-2351-8867.

Vector Autoregression (VAR) models (Diebold & Li, 2006) or the state-space model and the Kalman filter (Diebold et al., 2006). Additionally, there are noteworthy arbitrage-free approaches such as the Arbitrage-Free Dynamic Nelson-Siegel model proposed by Christensen, Diebold and Rudebusch (2011).

1.2. Machine learning approaches

Recently, with the growing popularity of machine learning (ML) methods, the yield curve factors are also forecasted using models such as Support Vector Machines (SVM), Group Method of Data Handling (GMDH; Kim et al., 2020) and different variations of neural networks, e.g. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM; Kim et al., 2020; Richman & Scognamiglio, 2024). Another approach is to use decision tree-based methods, such as Random Forests (RF; Kostyra & Rubaszek, 2020; Puglia & Tucker, 2020; Rayeni & Naderi, 2025), LightGBM (Puglia & Tucker, 2020) or XGBoost (Puglia & Tucker, 2020; Rayeni & Naderi, 2025; Santos Soares, 2025; Zhang, 2024). Decision trees can also be used as an enhancement of the dynamic NS model to identify different interest rate regimes and predict their changes (Bie et al., 2024). Additionally, tree-based methods were successfully applied to yield-curve-derived recession prediction (see Cadahia Delgado et al., 2022).

1.3. Methodological concerns in machine learning forecasts

The advantages of ML models over traditional time-series models are not obvious. While most of the published articles indicate that ML methods deliver more accurate forecasts than traditional benchmarks, it is worth noting that this might result from publication bias, when only the results indicating that ML methods are successful in forecasting are accepted for publication or data leakage in the design of the forecasting competition (see Hewamalage et al., 2022 for a general discussion on data leakage in the context of time-series forecasting). Puglia and Tucker (2020), who worked on US Treasury data, showed that the performance of ML methods in yield curve forecasting depends heavily on the choice of the training and cross-validation samples. As stated by the authors, 'strategies which eliminate data "peeking" produce lower, and perhaps more realistic, estimates of forecast accuracy' (Puglia & Tucker, 2020, p. 2). Rubaszek and Sznajderska (2026), who discuss the topic of data leakage in exchange rate forecasting, show that XGBoost models produce significantly more accurate forecasts than the random walk benchmark only when data leakage is allowed.

Considering the above, the aim of this article is to explore the suitability of the Random Forest (RF) framework in forecasting NS factors extracted from the US swap yield curve in the years 1990–2026. For this purpose, we estimate the ARIMA, VAR and RF models

for each of the NS model factors and evaluate the next-day forecasts. We also check the predictive content of selected financial variables.

2. Methodology

2.1. Nelson-Siegel model

Nelson and Siegel (1987) proposed a parsimonious model that allows the reproduction of the commonly observed shapes of yield curves. They proposed the following functional form:

$$R_m = L + S\left(\frac{1 - e^{-m\lambda}}{m\lambda}\right) + C\left(\frac{1 - e^{-m\lambda}}{m\lambda} - e^{-m\lambda}\right), \tag{1}$$

where:
R_m is the yield for maturity m,
L is the parameter for the long-term level of the interest rates,
S is the parameter responsible for the slope of the yield curve,
C is the parameter that affects the curvature of the yield curve,
λ is the parameter responsible for the shape of latent factors (see Figure 1).

As explained by Rubaszek (2012), the formula above can be derived by substituting the equation for the instantaneous forward rate (F_m):

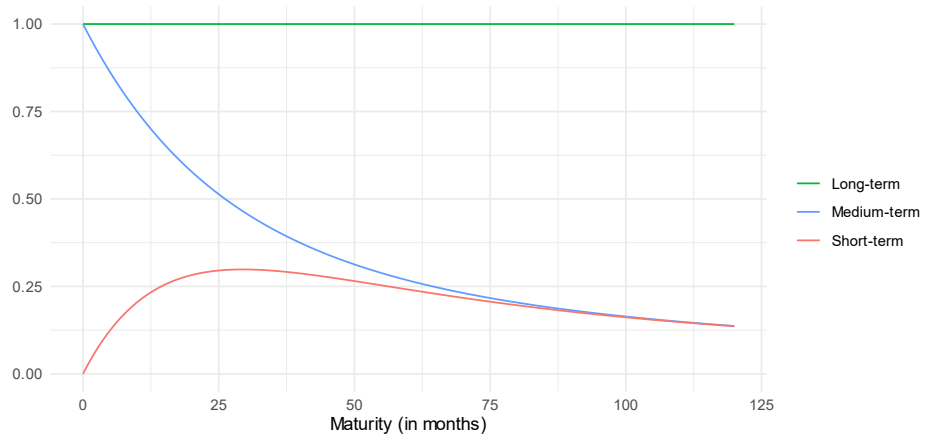
$$F_m = L + Se^{-m\lambda} + C\left(m\lambda \times e^{-m\lambda}\right) \tag{2}$$

to the definition of the spot rate:

$$R_m = \frac{1}{m}\int_0^m F_s ds. \tag{3}$$

To estimate the values of the NS factors, one can fix the λ parameter and use ordinary least squares (OLS) to derive the L, S and C factors. In this article, we follow Diebold and Li (2006), where $\lambda = 0.0609$ so that the peak of the Curvature factor is at a 30-month horizon (see Figure 1). This method also allows the comparability of NS factor estimates across different dates, as their interpretation depends on the value of λ.

Figure 1. Components of the spot rate



Source: author's calculations.

2.2. ARIMA

The first model used in forecasting NS latent factors is a well-known ARIMA (p, d, q) model of the following form:

$$\left(1 - \varphi_1 L - \varphi_2 L^2 - \ \ldots \ - \varphi_p L^p\right)\Delta^d y_t = \alpha_0 + \left(1 + \theta_1 L + \theta_2 L^2 + \ \ldots \ + \theta_q L^q\right)\varepsilon_t, \quad (4)$$

where:

y_t is the dependent variable,
L is the lag operator,
Δ is the differencing operator,
p is the order of the autoregressive process,
d is the order of the differencing of the dependent variable,
q is the order of the moving average process,
$\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ is the error term.

This model is described in detail in Hamilton (1994, pp. 43–71).

2.3. VAR

The VAR model, introduced by Christopher A. Sims (1980), represents a data-driven approach. Its main theoretical advantage over ARIMA is that it allows interactions between the variables in the equation system. The basic specification is given by:

$$\boldsymbol{y_t} = \boldsymbol{c} + \boldsymbol{\Phi_1 y_{t-1}} + \boldsymbol{\Phi_2 y_{t-2}} + \cdots + \boldsymbol{\Phi_p y_{t-p}} + \boldsymbol{\varepsilon_t}, \tag{5}$$

where:

$\boldsymbol{y_t}$ is an $(n \times 1)$ vector containing n variables in period t,

$\boldsymbol{\Phi_p}$ is the matrix of coefficients corresponding to the vector of p lagged values of the dependent variables,

$\boldsymbol{\varepsilon_t} \sim \boldsymbol{\mathcal{N}(0, \Omega)}$ is the error term.

This model is discussed in Hamilton (1994, pp. 291–350).

2.4. Random Forrest

The RF is the third model used in the forecasting competition of NS latent factors. It is an ensemble method introduced by Breiman (2001). Its main objective is to combine predictions of many de-correlated regression trees via bagging (bootstrap aggregating).

As described by James et al. (2021), the construction of a single regression tree for a response variable, in our case latent factor y_t, using a set of predictors $x_{1t}, x_{2t}, \ldots, x_{pt}$ relies on partitioning the p-dimensional predictor space into J distinct and non-overlapping regions $R_1, R_2, \ldots, R_J$. For each observation from the training sample, we assign it to given region R_i and calculate the predicted value of the response variable as the mean of all observations from R_i. The algorithm for creating J regions is called recursive binary splitting and consists of a loop in which one predictor variable is chosen and its value is used to divide the predictor space into two. The loop ends when a stopping criterion is reached. This criterion may rely on reaching a given (low) number of observations in each node or reaching the maximum tree depth. The algorithm for selecting the best split is based on finding the variable and its cut point value that minimises the following expression:

$$\sum_{t: x_{jt} \in R_1(j,s)} \left(y_t - \hat{y}_{R_1}\right)^2 + \sum_{t: x_{jt} \in R_2(j,s)} \left(y_t - \hat{y}_{R_2}\right)^2, \tag{6}$$

where $\hat{y}_{R_1}, \hat{y}_{R_2}$ are the means of the response variable in two regions created after the split across variable x_j and cut point s.

One of the drawbacks of a single regression tree is its relatively high variance. The results we obtain from a regression tree are very sample-sensitive. James et al. (2021) provide an explanation that e.g. splitting the dataset into two parts and fitting separate regression trees to both halves will produce two quite different trees, unlike in a low variance procedure such as linear regression, where the results would be somewhat similar (in cases where the number of observations is much higher than the number

of regressors). The solution to this problem relies on growing many independent trees and then averaging their predictions. To ensure that the trees remain independent, two kinds of solutions are applied. The first one is bagging, which involves growing trees on bootstrapped (different) samples. The second one deals with strong predictors that the algorithm selects repeatedly in every tree, in which case the resulting ensemble of trees is usually highly correlated. Here, bagging decreases the variance to a limited extent. Therefore, the number of candidate predictors for each split is restricted so that the algorithm will not be allowed to continuously use the strongest predictors, thus returning less correlated trees.

The RF framework above was developed for cross-section regression rather than time-series predictions. For that reason, one of the assumptions of RFs is that the realisations of the response variable are independent and identically distributed (i.i.d.). This is clearly at odds with many time-series observations, which are characterised by serial correlation. Thus, in this article, we use an extension of the RF approach for time-series data proposed by Goehry et al. (2023). The authors suggested a modification to the bootstrapping algorithm that draws blocks of consecutive observations rather than observations from the whole dataset. Their numerical experiments proved that the moving block bootstrap is the best choice for selecting these blocks for the bootstrap, regardless of the values of other hyperparameters. Its mechanism is straightforward and involves drawing blocks of consecutive observations of a predefined length.

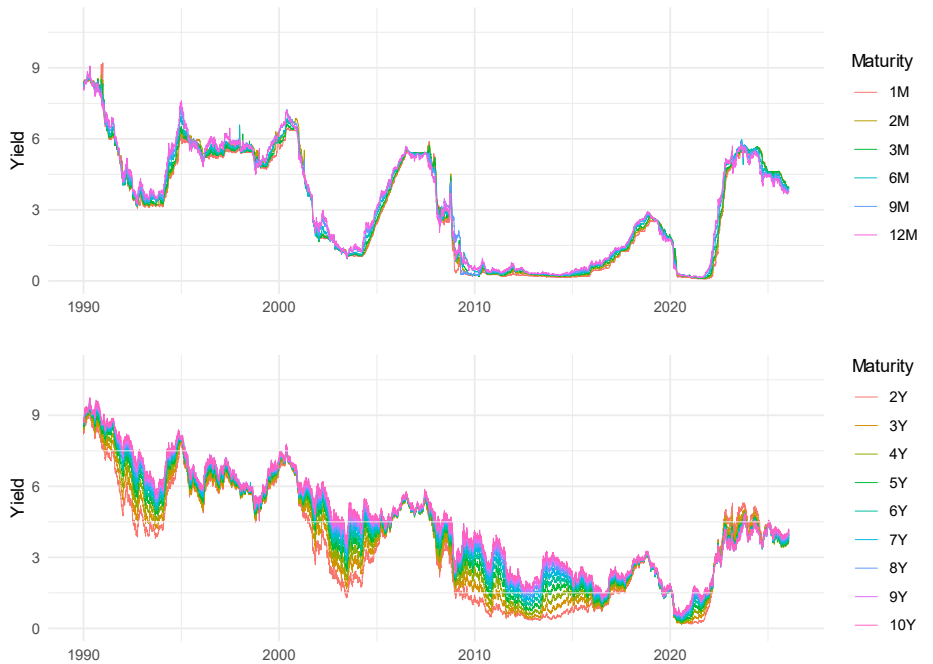
3. Empirical Analysis

3.1. Data

The analysis covers daily United States swap yield curve data from January 1990 to January 2026, obtained from Refinitiv Workspace. The dataset contains yields for the following maturities: 1M, 2M, 3M, 6M, 9M, 1Y, 2Y, 3Y, 4Y, 5Y, 6Y, 7Y, 8Y, 9Y, and 10Y. Swap rates are used rather than government bond yields, as they are derived from actual daily market transactions and do not require interpolation for missing maturities. Moreover, the swap market is believed to be more liquid.

As the NS model is derived for continuously compounded rates, raw data are transformed using the formula: $R_{m,t} = \ln\left(1 + \frac{r_{m,t}}{100}\right) \cdot 100$. Figure 2 shows how yields of short- and long-term maturities evolved over time.

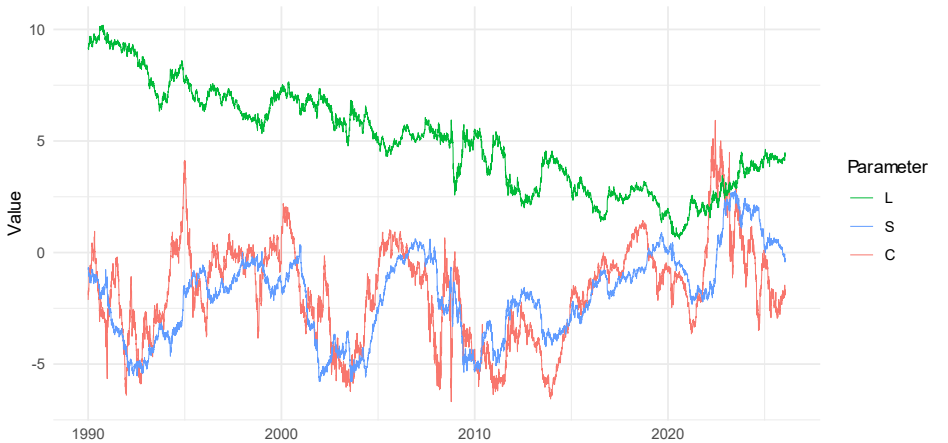
Figure 2. US swap yields



Source: author's calculations based on Refinitiv Workspace data.

In the next step, for each day t, we fitted the NS model to the yield curve. As already mentioned, following Diebold and Li (2006), the value of the λ parameter was fixed at 0.0609, whereas the values of L_t, S_t and C_t were estimated using the OLS estimator. The resulting estimates of the factor loadings are presented in Figure 3. It shows that the long-term level of interest rates decreased gradually up until 2020, followed by a sharp increase. Throughout the research period, the slope of the yield curve fluctuated from high term-premium to nearly none (flat yield curve), only rarely inverting. In most cases, this situation occurred during financial and/or global crises, mainly in 2001 (dotcom crisis), 2007–2008 (global economic crisis), 2020 (COVID-19 pandemic) and following 2022 (Russian invasion on Ukraine). C_t exhibited primarily negative values throughout the analysed period. The yield curve was positively humped during three major events, namely the bond market crisis in 1994, and the 2001 and 2022 crises. This means that the market expected interest rate hikes in the medium term.

Figure 3. Loadings of the NS model factors



Source: author's calculations.

Next, the stationarity of the resulting time series is tested using the Augmented Dickey-Fuller (ADF) test at a significance level of 0.05. The key conclusion from the results presented in Table 1 is that L_t and S_t are I(1), while C_t is stationary.

Table 1. Descriptive statistics and ADF test for NS model factors

Variable	Mean	SD	Minimum	Maximum	Skewness	Kurtosis	JB statistic	ADF test levels	ADF test differences
L	5.00	2.24	0.60	10.19	0.21	2.19	307.73	−2.02	−68.51
S	−1.91	1.90	−5.78	2.78	0.04	2.42	125.70	−1.85	−68.73
C	−1.75	2.19	−6.69	5.93	0.04	2.77	22.22	−3.66	–

Note. The critical value for the ADF test is -3.43 at the significance level of 0.01 and -2.86 at the significance level of 0.05.

Source: author's calculations.

The NS factor predictors were downloaded from the Federal Reserve Economic Database[1] (FRED). The following variables were selected:

- Chicago Board Options Exchange (CBOE) Volatility Index (VIX), which is a measure of the uncertainty of the US stock market; Bekaert et al. (2013) demonstrated that uncertainty and risk appetite are negatively related to interest rates;
- Kansas City Fed's Rate Uncertainty which reflects market expectations regarding short-term rates, calculated using publicly traded options contracts;

[1] FRED tickers for these variables are respectively: VIXCLS, KCPRU, DEXUSK, DEXJPUS, DEXSZUS, DCOILWTICO, DAAA.

- Exchange rates of the US dollar against the British pound (USD/GBP), Japanese yen (USD/JPY) and Swiss franc (USD/CHF). For Japan, Akram and Li (2024) demonstrated that exchange rates influence interest rates, as whenever a depreciation (or appreciation) of the domestic currency occurs, the central bank, following a Taylor-type rule, reacts by adjusting the interest rates to counteract the inflationary (or deflationary) pressures;
- West Texas Intermediate (WTI) crude oil prices; according to Akram and Li (2024), commodity prices affect inflation and thus interest rates;
- Moody's Seasoned Aaa Corporate Bond Yield (Aaa yield); Gilchrist and Zakrajšek (2012) showed that positive shocks to the excess bond premium have a negative impact on economic activity, thus necessitating monetary policy easing.

The descriptive statistics of these variables are provided in Table 2. In addition to the basic statistics, we computed skewness, kurtosis and the Jarque-Bera (JB; 1980) test statistic in order to assess whether the variables are normally distributed. The results in the table indicate that none of the variables is normally distributed. Interestingly, the VIX shows the highest skewness and is leptokurtic, while the other variables demonstrate a moderate degree of asymmetry and have platykurtic distributions.

Table 2. Descriptive statistics of the financial variables

Variable	Mean	SD	Minimum	Maximum	Skewness	Kurtosis	JB statistic
VIX	19.44	7.76	9.14	82.69	2.21	11.72	35,742.63
KCPRU	0.95	0.39	0.17	2.18	0.04	2.12	292.04
USD/GBP	1.55	0.21	1.07	2.11	0.28	2.52	205.68
USD/JPY	114.31	17.96	75.72	161.73	0.31	3.02	139.58
USD/CHF	1.17	0.25	0.73	1.82	0.48	2.07	666.64
WTI	51.60	29.05	–36.98	145.31	0.44	2.14	570.80
Aaa yield	5.60	1.75	2.01	9.68	0.23	2.23	299.64

Note. The JB test statistic has a $\chi^2(2)$ $distribution$ that has the critical values of 5.99 at the 0.05 significance level and 9.21 at the significance level of 0.01.

Source: author's calculations.

3.2. Forecasting competition design

We divide the whole dataset of 8,930 observations (2nd January 1990–27th January 2026) into three subsets: training (6,100 observations from the 2nd January 1990–29th August 2014 period), validation (2,575 observations from the 30th August 2014–16th January 2025 period) and testing (255 observations from the 17th January 2025–27th January 2026 period). The first two datasets are used to determine the optimal specification of the ARIMA, VAR and RF models.

In the forecasting competition, we compare four models (ARIMA, VAR, AR RF, AR RF + exogenous variables) over a 1-day horizon and three models (ARIMA, VAR, AR RF)

across horizons up to 20 days. For each factor, we fit an ARIMA, VAR and RF model with autoregressive predictors and RF with autoregressive and exogenous predictors. The exact tuning procedures for each model are described below.

For the ARIMA, we use the training set and `auto.arima()` function in the `R forecast` package, which returns the specification that optimises the Bayesian Information Criterion (BIC).

We utilise the `VARselect()function` from the `vars` package for the VAR model, through which we obtained the optimal lag order that minimises the BIC.

For the RF models, we used five lags of the dependent variable and one lag of the exogenous predictors. This approach is called predictive regression (see Stambaugh, 1999), which generates 1-day-ahead forecasts without the need to forecast exogenous predictors. As regards the hyperparameter tuning for the RF model, their values are derived from the following algorithm. For a given set of hyperparameters, we fitted the model using the `rangerts` package (based on the `ranger` package by Wright & Ziegler, 2017) to the training set. We then used its predictions in the validation set to compute the Root Mean Square Error (RMSE). We selected the hyperparameters that optimised the RMSE statistic. The exact list of the tuned hyperparameters and their feasible values are discussed in Section 3.3 of this article.

The testing set was used to compare the forecasting accuracy of the competing models. For this purpose, we employed a rolling origin setup (see Hewamalage et al., 2022). Consequently, starting from the end of the validation set (observation 8,675), we fitted the model to generate a 1-day – ahead forecast and then added the actual observation to the set. The process was repeated until reaching the end of the test set. The hyperparameters were fixed at their validation-set optima, meaning that only the model coefficients were re-estimated at each forecast origin.

To obtain longer horizon forecasts, we applied the recursive forecasting approach for the three autoregressive models. Specifically, to obtain $\hat{y}_{t+h+1|t}$ (forecast for the next out-of-sample value), we used all the necessary lags as either forecasted values if they were from periods $t+h, \ldots, t+1$ or actual observations for periods $t, t-1$ and so on.

3.3. Tuning results

The specification of the ARIMA models is presented in Table 3. For each of the three NS model factors, the optimal model specification utilised the first differences ($d = 1$) approach. The application of a second-order autoregressive component best captured the L_t and S_t variables, which exhibited high inertia. Conversely, the optimal model for C_t consisted of a first-order moving average only.

Table 3. Optimal specification of the ARIMA models

Parameter	*L*	*S*	*C*
p	2	2	0
d	1	1	1
q	0	0	1

Source: author's calculations.

Next, in the RF tuning procedure, the algorithm selected the optimal values from the feasible ones (indicated in square brackets) for the following hyperparameters: the number of variables to possibly split at each node (m_{try})[AR RF – 1:5; AR RF + X – 1:11], the maximum number of splits between the beginning and end of the tree ($max\ depth$) [3:30 for both models], the length of the bootstrap block ($block\ length$)[1:100 for both models] and the regularisation parameter that controls overfitting ($minimum\ node\ size$) [5:100 for both models].

As previously mentioned, the estimation was based on the moving-block bootstrap, which adapts the RF method to time series analyses. To accelerate the search for the optimal hyperparameter setting, we applied the random search approach described by Bergstra and Bengio (2012). Instead of evaluating all possible hyperparameter combinations, this approach samples with replacement from the feasible set. A total of 500 hyperparameter combinations were thus sampled.

The values of the hyperparameters that performed best in the validation set are presented in Tables 4 and 5. From among all the variables and models, the optimal $block\ length$ comprised approximately 15 observations, which is the equivalent to three trading weeks. The models for C_t had grown deeper than for the other variables; this may mean that more intricacy is necessary for the proper forecasting of this factor than in the case of the other two. Regarding the m_{try} parameter in the autoregressive models, the value of four across all factors meant that the algorithm used four out of the five available lags to construct one tree. In the models incorporating exogenous predictors, the algorithm naturally selected a larger number of regressors per tree; for L_t, on the other hand, the requirement was lower by two. The $minimum\ node\ size$ hyperparameter is closely tied to the $max\ depth$. The deeper the tree, the smaller the final node size and vice versa. Both regularisation hyperparameters were optimised to address two different sources of overfitting.

Table 4. Optimal hyperparameters for RF with autoregressive predictors

Hyperparameter	*L*	*S*	*C*
$block\ length$	15	15	12
$max\ depth$	9	9	17
m_{try}	4	4	4
$minimum\ node\ size$	18	18	19

Source: author's calculations.

Table 5. Optimal hyperparameters for RF with autoregressive and exogenous predictors

Hyperparameter	*L*	*S*	*C*
$block\ length$	14	17	17
$max\ depth$	7	25	25
m_{try}	8	10	10
$minimum\ node\ size$	35	6	6

Source: author's calculations.

3.4. Accuracy of the forecast

The aim of the presented research is to verify whether autoregressive RFs are able to deliver additional forecasting power in comparison with ARIMA and VAR models and whether adding financial predictors improves the forecast accuracy of the latent factors. For this purpose, we analysed the 1-day-ahead forecasts generated by each of the four competing methods. Then, we discuss the forecasts of the three strictly autoregressive models over longer forecast horizons.

3.4.1. One-day-ahead forecasts

Table 6. Root mean square forecast error of NS model factors; 1-day-ahead forecasts

Factor	ARIMA	Autoregressive RF	VAR	Autoregressive RF with exogenous variables
L	0.0548	0.0593	0.0542	0.0572
S	0.0563	0.0603	0.0575	0.0604
C	0.1581	0.1672	0.1609	0.1688

Source: author's calculations.

We calculated the root mean square forecast error (RMSFE) for each model and factor. The results in Table 6 show that for each factor, ARIMA outperformed both RF types. It was evident that C_t was more difficult to forecast than L_t and S_t. Interestingly, adding financial covariates to the RF increased the forecast accuracy only for the Level factor. Additionally, despite its regularisation hyperparameters, RF proved susceptible to overfitting on the slightly noisy daily data.

In the final phase of our research, we tested whether the differences in forecast accuracy between the models is statistically significant. For this purpose, we performed the Diebold-Mariano (1995, DM) test, which uses ARIMA as a benchmark. Table 7, which contains the *p*-values of the DM tests, indicates that the use of RFs leads to statistically significant deterioration in forecasting accuracy rather than an improvement, compared to the traditional ARIMA and VAR benchmarks. These findings are at odds with the other studies discussed in the Introduction. Since our analysis is limited to the US market, the discussion that follows concerns the validation design of these studies rather than a direct comparison of forecast accuracy. The observed discrepancy may stem from data leakage, which is not always controlled by other researchers. For example, Kim et al. (2020) do not provide a clear validation strategy as they discuss the partition of their dataset into training and test subsamples only; in addition, they do not disclose which country the data come from. Furthermore, they did not ensure a level playing field for classical and ML models, e.g. they restricted the classical approach to an NS model with AR(1) factors only. This is consistent with the broader concerns raised by Hewamalage et al. (2022) and Puglia and Tucker (2020) regarding leakage (even unintentional) and validation design. Additionally, due to publication bias, many articles that do not find a significant advantage of using ML methods are simply not published.

Table 7. DM test *p*-values for 1-day-ahead forecasts

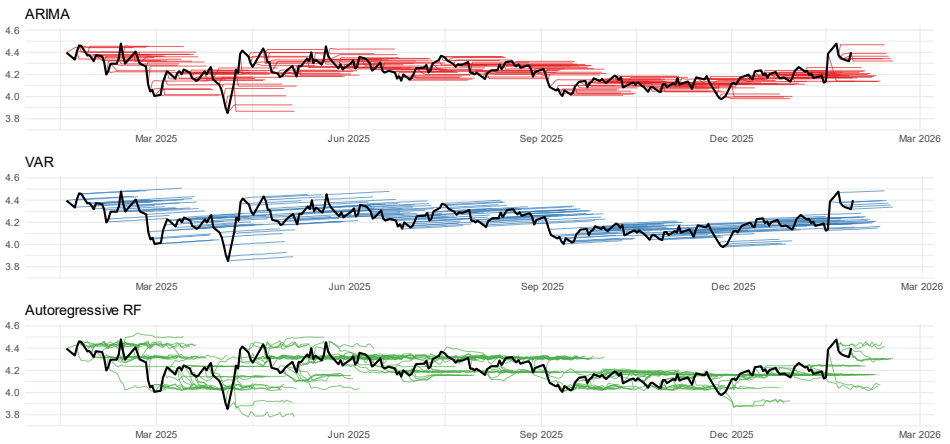
Factor	ARIMA vs VAR	ARIMA vs Autoregressive RF	VAR vs Autoregressive RF	ARIMA vs Autoregressive RF with exogenous variables	VAR vs Autoregressive RF with exogenous variables	Autoregressive RF with exogenous variables vs Autoregressive RF
L	0.7883	0.0044	0.0006	0.0317	0.0074	0.0282
S	0.0352	0.0044	0.0362	0.0141	0.0638	0.5354
C	0.0488	0.0233	0.0603	0.0248	0.0669	0.6543

Note. H1: Model 1 is better than model 2.
Source: author's calculations.

3.4.2. Forecasts for 1- to 20-day horizons

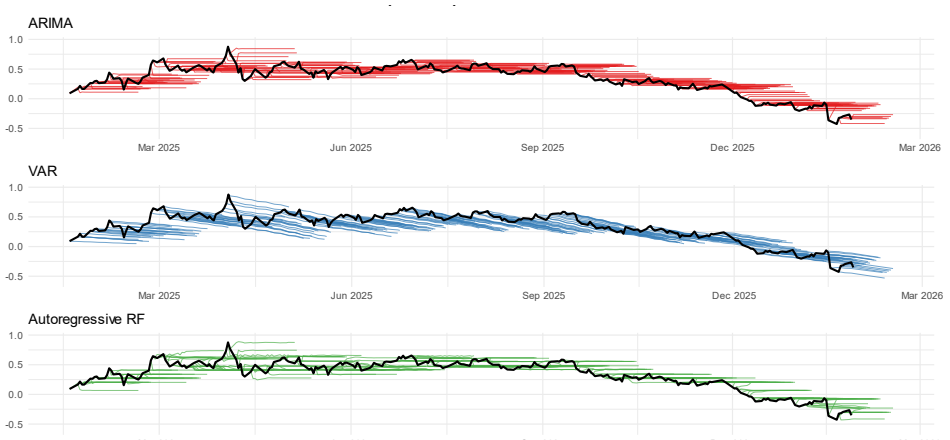
In this section, we compare forecasts from the three autoregressive models across horizons from 1 to 20 days. Figures 4–6 demonstrate that all the models tend to provide mean-reverting forecasts. Furthermore, they fail to predict sudden, sharp changes of the latent factors. It is worth noting that the autoregressive RFs return volatile forecasts, whereas ARIMA and VAR models provide smooth trajectories of future values.

Figure 4. Sequential forecasts for the Level factor



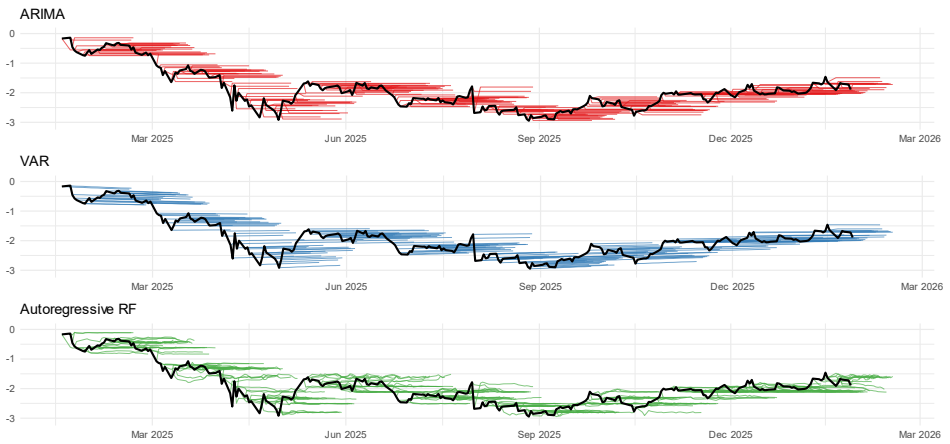
Source: author's calculations.

Figure 5. Sequential forecasts for the Slope factor



Source: author's calculations.

Figure 6. Sequential forecasts for the Curvature factor



Source: author's calculations.

Table 8 shows the RMSFEs of the three autoregressive models across longer horizons, while Table 9 presents the p-values of the DM test that allows the verification of the significance of pairwise forecast accuracy differences. To account for serial correlation in the loss differential at longer horizons, the long-run variance is estimated using autocovariances up to lag h-1, where h denotes the forecast horizon (see more in Diebold & Mariano, 1995).

Table 9 suggests that at a 0.05 significance level, ARIMA outperforms considerably VAR just for 7 out of 20 horizons for L_t, and for S_t and C_t in the case of 1-day-ahead forecasts only. ARIMA yields better results than the autoregressive RF across all 20 horizons for L_t and S_t; similarly, it dominates in forecasting C_t up to h = 14 (with the exception of h = 2 and h = 10). VAR generates more accurate forecasts than the autoregressive RF for L_t up to h = 13; however, it fails to outperform the RF significantly at any of the horizons for S_t (except h = 1). For C_t, VAR performs more efficiently for horizons from h = 3 to h = 13.

Table 8. RMSFEs of NS model factors for h = 1, …, 20 forecast horizons

Variable	h = 1	h = 2	h = 3	h = 4	h = 5	h = 6	h = 7	h = 8	h = 9	h = 10	h = 11	h = 12	h = 13	h = 14	h = 15	h = 16	h = 17	h = 18	h = 19	h = 20
ARIMA																				
L	0.0548	0.0784	0.0915	0.1034	0.1120	0.1171	0.1217	0.1254	0.1277	0.1293	0.1291	0.1302	0.1298	0.1284	0.1282	0.1271	0.1263	0.1264	0.1264	0.1268
S	0.0563	0.0799	0.0942	0.1066	0.1169	0.1226	0.1284	0.1335	0.1372	0.1409	0.1434	0.1465	0.1470	0.1474	0.1492	0.1501	0.1518	0.1548	0.1579	0.1610
C	0.1581	0.2006	0.2313	0.2637	0.2868	0.3067	0.3232	0.3447	0.3668	0.3911	0.4083	0.4261	0.4421	0.4538	0.4657	0.4776	0.4913	0.5014	0.5153	0.5230
VAR																				
L	0.0542	0.0786	0.0923	0.1046	0.1136	0.1186	0.1230	0.1270	0.1300	0.1323	0.1325	0.1341	0.1338	0.1327	0.1332	0.1323	0.1321	0.1326	0.1334	0.1344
S	0.0575	0.0812	0.0958	0.1090	0.1204	0.1269	0.1339	0.1400	0.1451	0.1501	0.1544	0.1594	0.1621	0.1645	0.1687	0.1721	0.1761	0.1818	0.1872	0.1930
C	0.1609	0.2014	0.2314	0.2636	0.2869	0.3058	0.3210	0.3414	0.3622	0.3862	0.4023	0.4196	0.4350	0.4460	0.4566	0.4672	0.4800	0.4882	0.5018	0.5080
Autoregressive RF																				
L	0.0593	0.0862	0.1027	0.1156	0.1282	0.1349	0.1385	0.1428	0.1477	0.1499	0.1452	0.1441	0.1430	0.1409	0.1400	0.1385	0.1403	0.1433	0.1435	0.1466
S	0.0603	0.0828	0.0980	0.1103	0.1218	0.1286	0.1328	0.1371	0.1412	0.1455	0.1485	0.1525	0.1541	0.1553	0.1570	0.1590	0.1607	0.1623	0.1641	0.1676
C	0.1672	0.2059	0.2433	0.2820	0.3031	0.3254	0.3371	0.3598	0.3800	0.4010	0.4166	0.4326	0.4513	0.4623	0.4720	0.4829	0.4971	0.5032	0.5157	0.5271

Source: author's calculations.

Table 9. DM test *p*-values for h = 1, …, 20 forecast horizons

Variable	h = 1	h = 2	h = 3	h = 4	h = 5	h = 6	h = 7	h = 8	h = 9	h = 10	h = 11	h = 12	h = 13	h = 14	h = 15	h = 16	h = 17	h = 18	h = 19	h = 20
ARIMA vs. VAR																				
L	0.7883	0.4396	0.1611	0.0120	0.0101	0.0967	0.1615	0.1514	0.1025	0.0585	0.0595	0.0479	0.0472	0.0391	0.0303	0.0463	0.0563	0.0756	0.0866	0.0969
S	0.0352	0.1874	0.2653	0.2598	0.2313	0.2456	0.2385	0.2413	0.2273	0.2141	0.1949	0.1727	0.1585	0.1541	0.1493	0.1486	0.1557	0.1576	0.1631	0.1654
C	0.0488	0.3146	0.4970	0.5357	0.4823	0.6122	0.7188	0.7952	0.8597	0.8494	0.8751	0.8785	0.8760	0.8810	0.9072	0.9227	0.9335	0.9472	0.9397	0.9383
ARIMA vs. Autoregressive RF																				
L	0.0044	0.0051	0.0021	0.0025	0.0004	0.0000	0.0001	0.0000	0.0000	0.0000	0.0002	0.0013	0.0037	0.0033	0.0028	0.0037	0.0003	0.0004	0.0028	0.0069
S	0.0044	0.0246	0.0098	0.0416	0.0140	0.0117	0.0207	0.0536	0.0446	0.0403	0.0493	0.0383	0.0343	0.0375	0.0288	0.0185	0.0069	0.0057	0.0001	0.0010
C	0.0233	0.1768	0.0022	0.0001	0.0001	0.0031	0.0301	0.0145	0.0284	0.0553	0.0403	0.0146	0.0311	0.1062	0.1136	0.1273	0.1707	0.3279	0.4703	0.2344
VAR vs. Autoregressive RF																				
L	0.0006	0.0032	0.0042	0.0071	0.0014	0.0001	0.0002	0.0001	0.0001	0.0000	0.0003	0.0058	0.0177	0.0325	0.0814	0.1465	0.0871	0.0661	0.1052	0.1175
S	0.0362	0.2234	0.2329	0.3716	0.3803	0.3985	0.5580	0.6241	0.6478	0.6620	0.6909	0.7069	0.7103	0.7149	0.7391	0.7375	0.7452	0.7699	0.7821	0.7823
C	0.0603	0.1969	0.0016	0.0009	0.0025	0.0040	0.0137	0.0064	0.0141	0.0351	0.0272	0.0310	0.0238	0.0682	0.0715	0.0559	0.0533	0.0552	0.1013	0.0575

Note. H1: Model 1 is better than model 2.

Source: author's calculations.

4. Conclusions

The principal finding of this study is that the persistent use of ML methods for modelling the US yield curve is not always superior to classical approaches. We have found that using ARIMA and VAR for modelling the NS model parameters outperforms the RF (with time-series specific bootstrap) approach in 1-day-ahead forecasts. Adding financial variables to the autoregressive RFs significantly improved the forecasts for the Level factor only. These conclusions suggest that the RFs tended to overfit to the noise in the daily data and that the daily NS factors are just heavily autocorrelated. Therefore, more parsimonious models such as ARIMA or VAR remain effective in capturing the whole signal and providing accurate forecasts.

Likewise, for longer forecast horizons (up to 20 trading days), ARIMA and VAR models outperformed the RF. Compared to the RF, ARIMA generated more accurate forecasts for all the factors, whereas VAR provided better forecasts for the Level and Curvature factors. Interestingly, ARIMA generated more accurate forecasts for the Level and Slope factors (although not statistically significant as far as the second variable is concerned) than VAR. Curvature forecasts, however, were more accurate when made from the VAR model, probably because lagged values of the two other factors allow a better understanding of the yield curve's curvature.

The yield forecasts from all three models were sometimes systematically higher or lower than the actual values. This indicates the presence of a cross-section error stemming from a poor fit of the NS model to the actual yield curve. It is worth noting that this is an inherent fitting error as the NS decomposition is only an approximation of the yield curve. This may also result from using a fixed λ that can vary from the optimal λ for each day. The main advantage of using a fixed λ is that it allows the comparability with other studies.

This study additionally emphasises the need for rigorous data-leakage protection and addressing the problem of publication bias in the ML field.

References

Akram, T., & Li, H. (2024). Empirical Models of JGB Yields Using Daily Data. *Journal of Economic Issues*, *58*(3), 1011–1034. https://doi.org/10.1080/00213624.2024.2382051.

Bekaert, G., Hoerova, M., & Lo Duca, M. (2013). Risk, uncertainty and monetary policy. *Journal of Monetary Economics*, *60*(7), 771–788. https://doi.org/10.1016/j.jmoneco.2013.06.003.

Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, *13*(10), 281–305. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf.

Bie, S., Diebold, F. X., He, J., & Li, J. (2024). *Machine Learning and the Yield Curve: Tree-Based Macroeconomic Regime Switching*. https://doi.org/10.2139/ssrn.4934442.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A :1010933404324.

Cadahia Delgado, P., Congregado, E., Golpe, A. A., & Vides, J. C. (2022). The Yield Curve as a Recession Leading Indicator. An Application for Gradient Boosting and Random Forest. *International Journal of Interactive Multimedia and Artificial Intelligence*, *7*(3), 7–19. https://doi.org/10.9781/ijimai.2022.02.006.

Christensen, J. H. E., Diebold, F. X., & Rudebusch, G. D. (2011). The affine arbitrage-free class of Nelson–Siegel term structure models. *Journal of Econometrics*, *164*(1), 4–20. https://doi.org/10.1016/j.jeconom.2011.02.011.

Diebold, F. X., & Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, *130*(2), 337–364. https://doi.org/10.1016/j.jeconom.2005.03.005.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599.

Diebold, F. X., Rudebusch, G. D., & Aruoba, S. B. (2006). The macroeconomy and the yield curve: a dynamic latent factor approach. *Journal of Econometrics*, *131*(1–2), 309–338. https://doi.org/10 .1016/j.jeconom.2005.01.011.

Gilchrist, S., & Zakrajšek, E. (2012). Credit Spreads and Business Cycle Fluctuations. *The American Economic Review*, *102*(4), 1692–1720. https://doi.org/10.1257/aer.102.4.1692.

Goehry, B., Yan, H., Goude, Y., Massart, P., & Poggi, J.-M. (2023). Random Forests for Time Series. *REVSTAT – Statistical Journal*, *21*(2), 283–302. https://doi.org/10.57805/revstat.v21i2.400.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hewamalage, H., Ackermann, K., & Bergmeir, C. (2022). Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, *37*, 788–832. https://doi.org/10.1007/s10618-022-00894-5.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*(3), 255–259. https://doi.org/10.1016 /0165-1765(80)90024-5.

Kim, W. J., Jung, G., & Choi, S.-Y. (2020). Forecasting CDS Term Structure Based on Nelson–Siegel Model and Machine Learning. *Complexity*, *1*, 1–23. https://doi.org/10.1155/2020/2518283.

Kostyra, T. P., & Rubaszek, M. (2020). Forecasting the Yield Curve for Poland. *Econometric Research in Finance*, *5*(2), 103–117. https://doi.org/10.2478/erfin-2020-0006.

Nelson, C. R., & Siegel, A. F. (1987). Parsimonious Modeling of Yield Curves. *The Journal of Business*, *60*(4), 473–489.

Puglia, M., & Tucker, A. (2020). *Machine Learning, the Treasury Yield Curve and Recession Forecasting* (Finance and Economics Discussion Series 2020-038). https://doi.org/10.17016/feds.2020.038.

Rayeni, A., & Naderi, H. (2025). Predicting the Canadian Yield Curve Using Machine Learning Techniques. *International Journal of Financial Studies*, *13*(3), 1–30. https://doi.org/10.3390 /ijfs13030170.

Richman, R., & Scognamiglio, S. (2024). Multiple yield curve modeling and forecasting using deep learning. *ASTIN Bulletin*, *54*(3), 463–494. https://doi.org/10.1017/asb.2024.26.

Rubaszek, M. (2012). *Modelowanie polskiej gospodarki z pakietem R*. Oficyna Wydawnicza SGH.

Rubaszek, M., & Sznajderska, A. (2026). *Data leakage in time-series forecasting: Lessons from exchange rate prediction*.

Santos Soares, S. A. (2025). *Eu-Bonds Yield Curve Forecast: Comparing ARIMA and XGBoost Models* [master's thesis, Lisbon School of Economics & Management].

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, *48*(1), 1–48. https://doi.org/10.2307/1912017.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, *54*(3), 375–421. https://doi.org/10.1016/s0304-405x(99)00041-0.

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01.

Zhang, J. (2024). *Forecasting Chinese Government Bond Yield Curves: An Empirical Comparison of DNS (Dynamic-Nelson-Siegel) Model and Machine Learning Approaches* [master's thesis, University of Chicago].

Przegląd Statystyczny. Statistical Review, 2026, vol. 73, 1, 35–55 https://doi.org/10.59139/ps.2026.02.3
Received: 10.04.2026, revised: 28.05.2026, accepted: 29.05.2026

Risk mitigation in a volatile US equity market: A comparative analysis of hedging with index futures and investing in gold as a safe haven

Krzysztof Echaust,[a] Agnieszka Lach[b]

Abstract. The aim of this paper is to compare two alternative strategies for protecting a US stock market portfolio against market risk during four major stock market crashes: the Global Financial Crisis, the European Sovereign Debt Crisis, the COVID-19 pandemic and the Russia-Ukraine war. Hedging with the S&P 500 index futures is compared with investing in gold as a safe haven based on the risk minimisation criterion. The effectiveness of the protection strategies is verified for portfolios that differ in terms of the number of constituents which range from single-asset to well-diversified portfolios. The results vary depending on the specific crisis, portfolio size and time horizon. However, hedging with index futures tended to provide more effective long-term protection, particularly for larger portfolios.

Keywords: hedging, safe haven, futures, gold, portfolio
JEL: G11, G13, C58

1. Introduction

In the last 20 years, four major events have shaken global financial markets. The most significant was the Global Financial Crisis (GFC) of 2007–2009, triggered by subprime lending, inflated housing prices and poorly regulated mortgage-backed securities. The bankruptcy of the Lehman Brothers on 15th September 2008 was the most dramatic moment, pulling financial markets down. The European Sovereign Debt Crisis (ESDC) was a multi-year financial crisis that began after the GFC, affecting several Eurozone member states. It was triggered by a combination of high government debt, vulnerable banking systems and weak economic growth, particularly in countries such as Greece, Ireland, Italy, Portugal and Spain. In early August 2011, global financial markets sharply declined due to the growing fears of a Greek default and contagion throughout the Eurozone. A more recent crisis was caused by the COVID-19 pandemic, which was officially declared as such by the World Health Organization on 11th March 2020. It led to an unprecedented global economic shock. Widespread lockdowns, supply chain disruptions and a collapse in consumer demand triggered

[a] Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Operations Research and Mathematical Economics, Al. Niepodległości 10, 61–875 Poznań, Poland, e-mail: krzysztof.echaust@ue.poznan.pl, ORCID: https://orcid.org/0000-0002-3855-256X.
[b] Poznań University of Economics and Business, Institute of Informatics and Quantitative Economics, Department of Operations Research and Mathematical Economics, Al. Niepodległości 10, 61–875 Poznań, Poland, e-mail: agnieszka.lach@ue.poznan.pl, ORCID: https://orcid.org/0000-0002-2831-6336.

a global recession. The fourth crisis began with Russia's invasion of Ukraine on 24th February 2022, causing further economic disruption across Europe. The conflict resulted in a major energy crisis, involving spikes in gas, electricity and crop prices, which accelerated inflation.

Portfolio diversification aims to reduce risk by investing in uncorrelated assets. However, it cannot eliminate systematic risk, which affects the entire market. During market crashes, both volatility and correlations between assets tend to increase (Sandoval & Franca, 2012). Thus, diversification weakens when most needed, prompting investors to shift from risky assets like stocks and corporate bonds to safe havens. According to Baur and Lucey (2010), safe-haven assets are uncorrelated or negatively correlated with the reference asset in times of a market crash.

Based on the Modern Portfolio Theory, such investments allow investors to reduce the overall risk of their portfolios. A wide range of instruments serving as safe havens have been analysed in the recent literature, including precious metals (Baur & Lucey, 2010; Baur & McDermott, 2010; Beckmann et al., 2015; Boubaker et al., 2020; Echaust & Just, 2022), bonds (Drobetz et al., 2020; Pisedtasalasai, 2021), oil (Batten et al., 2021; Disli et al., 2021), agricultural commodities (Ali et al., 2020), foreign exchange rates (Cho & Han, 2021; Dong et al., 2021; Siemaszkiewicz, 2023), cryptocurrencies (Będowska-Sójka & Kliber, 2021; Conlon & McGee, 2020; Goodell & Goutte, 2021; Just & Echaust, 2024; Kliber et al., 2019; Mizerka et al., 2020), and other non-traditional assets (Siemaszkiewicz & Let, 2020). Among them, gold is perceived as the traditional and most efficient safe haven.

A large body of literature emphasises the role of gold as a safe-haven asset. Baur and McDermott (2010) examine the safe-haven properties of gold for equities across 53 emerging and developed countries over a 30-year period (1979–2009). They show evidence that gold is a strong-form safe haven for most major developed stock markets. Boubaker et al. (2020) provide long-run evidence, showing that gold functions as a hedge against risk over a period of more than seven centuries, depending on the model specification. Similarly, Klein (2017) finds that gold and silver act as safe havens in developed markets, although their effectiveness weakens after 2013. Beckmann et al. (2015) show that gold's safe-haven property is market-specific across 18 stock markets and five regional indices, likely reflecting differences in market structure and capital flows. Focusing on crisis episodes, Dong et al. (2021) and Ji et al. (2020) demonstrate that gold can protect portfolios against extreme equity losses, particularly during the GFC and the COVID-19 pandemic. Banerjee and Pradhan (2024) confirm the similar safe-haven behaviour of gold for U.S. equities using high-frequency data during COVID-19. In emerging markets, Wen and Cheng (2018) also support the safe-haven role of this commodity using copula methods, while Ryan et al. (2024) show that gold should be used as a safe haven against S&P 500 risk during periods of macroeconomic uncertainty.

An alternative investment strategy used to protect portfolios against losses during market crashes is hedging with derivatives. It allows investors to temporarily offset portfolio losses with profits generated from these derivatives. The use of linear derivatives such as forwards or futures requires, unlike safe-haven instruments, a high correlation between these contracts and the underlying risk exposure. Correlation is therefore a key factor in determining whether a particular asset is suitable for a given hedging strategy, as periods of financial distress not only disrupt the functioning of financial markets through the transmission of shocks, but also significantly increase market interdependence (Forbes & Rigobon, 2002). This heightened interdependence may amplify crises, disrupting both financial markets and real economic activity. This proves that accounting for market interconnections is crucial to understanding financial market behaviour (Fałdziński & Pietrzak, 2015).

Numerous studies have examined the same protective assets – predominantly gold, oil and cryptocurrencies – against stock market risk, considering their roles in long positions as safe havens as well as in short positions as hedging instruments. However, these two strategies require assets with fundamentally different characteristics. While instruments like gold, oil or cryptocurrencies may serve as safe havens, they are inappropriate for a hedging strategy against stock market risk. Echaust et al. (2024) show that these instruments cannot compete with index futures in a hedging role. We aim to compare the effectiveness of both strategies within the minimum variance framework in the context of four major market crashes: the Global Financial Crisis, the European Sovereign Debt Crisis, the COVID-19 pandemic, and the Russia-Ukraine war. Unlike previous studies, we compare gold as a safe-haven asset with S&P 500 futures as a hedging instrument against the risk exposure of equity portfolios of varying sizes. Although the safe-haven strategy may be perceived as offering longer-term protection than hedging, Baur and Lucey (2010) showed that in practice gold is a safe haven only in the short run. Safe havens have been one of the most extensively explored subjects in the financial literature in recent years (Anas et al., 2024); therefore, it appears both reasonable and indeed necessary to compare these two key risk-mitigation strategies.

We contribute to the existing literature in two ways. First, we compare a safe-haven strategy in the spirit of Baur and Lucey (2010) with a hedging strategy. We evaluate their effectiveness in the U.S. stock market during four global stock market crashes. Our study considers gold as a safe haven and S&P 500 futures as a hedging instrument against stock portfolio risk. These assets are likely to be the first choice for many investors, which makes our study highly relevant from a practical perspective. Second, unlike most existing studies, we examine the protection of stock portfolios that vary in the number of constituents, rather than using broad market indices. Limiting the analysis to market indices would place the hedging strategy in a favourable position, as it would result in near-perfect hedging. However, small equity portfolios typically show lower correlation with index futures, which limits the effectiveness of hedging.

2. Methods

2.1. The choice of a stock portfolio

Let us assume that an investor on the brink of a financial crash (the last day of the portfolio construction period shown in Figure 1) holds a stock portfolio ranging from a single stock to a fully diversified one, including all S&P 500 constituents. The stocks in the portfolio are selected based on a variance minimisation criterion calculated from a one-year sample, with short selling restricted. This assumption ensures consistency in our subsequent analysis of portfolio protection, as both the hedging ratio and the optimal weights of the safe-haven asset are determined using the same criterion. Furthermore, we consider two approaches to portfolio construction: the first is the minimum variance portfolio according to Markowitz (1952), and the second is based on equal-weighting.

2.2. Investing in a safe haven

Let us assume that the investor holds a long stock portfolio position chosen according to Subsection 2.1. and a long position in a safe haven. The overall portfolio (after adding a safe-haven asset to the stock portfolio) return $r_{P,t}$, at time t is

$$r_{P,t} = w_{S,t} \cdot r_{S,t} + w_{SH,t} \cdot r_{SH,t}, \tag{1}$$

where $r_{S,t}$ and $r_{SH,t}$ denote return on the stock portfolio and the safe haven, respectively, and $w_{S,t}$ and $w_{SH,t}$ denote their weights in a portfolio, such that $w_{S,t} + w_{SH,t} = 1$.

The formula for an optimal (minimum variance) weight of a safe-haven asset is given by

$$w_{SH,t} = \begin{cases} 0 & \text{for} \quad w^*_{SH,t} \leq 0 \\ w^*_{SH,t} & \text{for} \quad 0 < w^*_{SH,t} \leq 1, \\ 1 & \text{for} \quad w^*_{SH,t} > 1 \end{cases} \tag{2}$$

where

$$w^*_{SH,t} = \frac{\sigma^2_{S,t} - Cov(r_{S,t}, r_{SH,t})}{\sigma^2_{SH,t} + \sigma^2_{S,t} - 2 \cdot Cov(r_{S,t}, r_{SH,t})}, \tag{3}$$

where $\sigma^2_{S,t}$, $\sigma^2_{SH,t}$ denote variances of the stock portfolio and the safe haven, respectively, and $Cov(r_{S,t}, r_{SH,t})$ denotes the covariance between the stock portfolio and safe haven returns.

2.3. Short hedging with futures

Let us assume that the investor holds a long stock portfolio position chosen according to Subsection 2.1. and adopts a short futures position to hedge the portfolio. The overall portfolio (after hedging the stock portfolio with a futures contract) return $r_{P,t}$, at time t is

$$r_{P,t} = r_{S,t} - h_t \cdot r_{F,t}, \tag{4}$$

where $r_{S,t}$ and $r_{F,t}$ denote returns on the stock portfolio and the futures, respectively, and h_t denotes the hedge ratio. The minimum variance optimal hedge ratio at time t is as follows:

$$h_t = \frac{Cov(r_{S,t}, r_{F,t})}{\sigma^2_{F,t}}. \tag{5}$$

2.4. Effectiveness of the protection strategy

In order to compare the performance of the chosen protection strategy, we employ a well-known performance measure for the variance minimisation problem, designed to evaluate the effectiveness of a hedging strategy (Ederington, 1979), namely:

$$Effectiveness = 1 - \frac{\text{Variance of a protected portfolio}}{\text{Variance of a stock portfolio}}. \tag{6}$$

This ratio reflects how well a protective strategy minimises the variance (in percentage terms) of a stock portfolio. The ratio of 0 means the protection strategy provides no risk reduction, while the ratio of 1 indicates perfect effectiveness, fully reducing the variance of the underlying exposure.

3. Data

The data were obtained from the Refinitiv Eikon database and include stock prices, E-mini S&P 500 Index Futures (ESc1) and gold spot prices quoted in U.S. dollars per ounce (XAU=). The sample includes S&P 500 constituents with complete price histories, defined separately at the onset of each crisis. For every crisis, the portfolio construction period is defined as the one-year interval immediately preceding the crisis onset, while the portfolio evaluation begins on the crisis onset date (15th September 2008, 1st August 2011, 11th March 2020 and 24th February 2022). The evaluation will be conducted over four different horizons: weekly, monthly, semi-annual and annual. The details are provided in Table 1 and Figure 1.

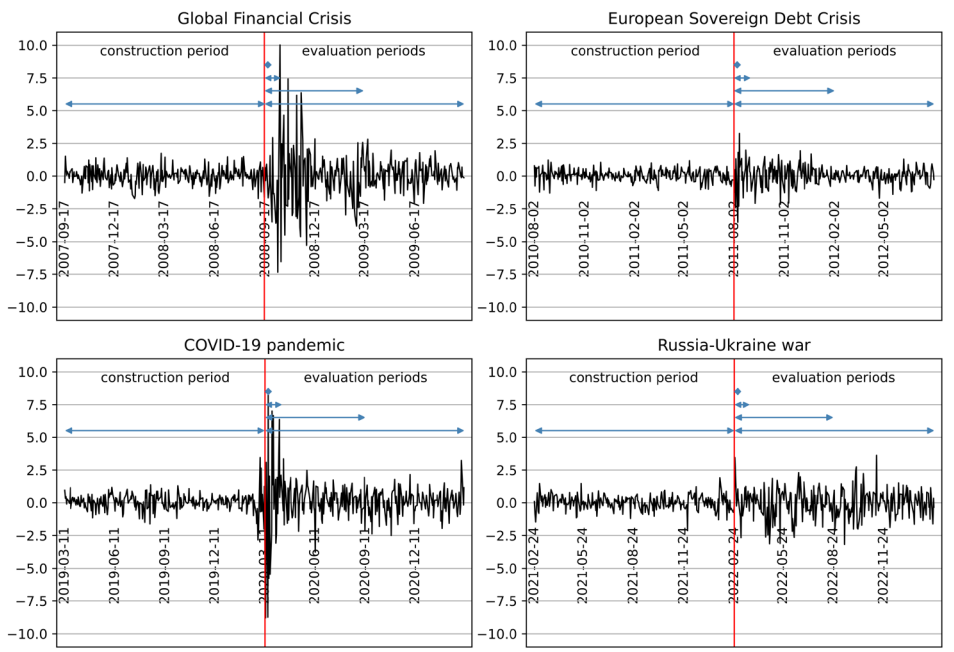
Table 1. Portfolio construction and evaluation periods

Crisis	Construction period	Full evaluation period
Global Financial Crisis	15-09-2007 – 14-09-2008	15-09-2008 – 14-09-2009
European Sovereign Debt Crisis	01-08-2010 – 31-07-2011	01-08-2011 – 31-07-2012
COVID-19 pandemic	11-03-2019 – 10-03-2020	11-03-2020 – 10-03-2021
Russia-Ukraine war	24-02-2021 – 23-02-2022	24-02-2022 – 23-02-2023

Source: authors' work.

Figure 1 clearly shows that the assumed crash start dates separate the period of low volatility (the portfolio construction period) prior to the outbreak of the crash from the subsequent increase in volatility during the evaluation periods. The most spectacular increase in volatility is observed during the periods of the GFC and COVID-19 pandemic.

Figure 1. Construction and evaluation periods



Note. The figure illustrates returns of the Markowitz portfolio comprising 10 constituents.
Source: authors' work.

4. Empirical study

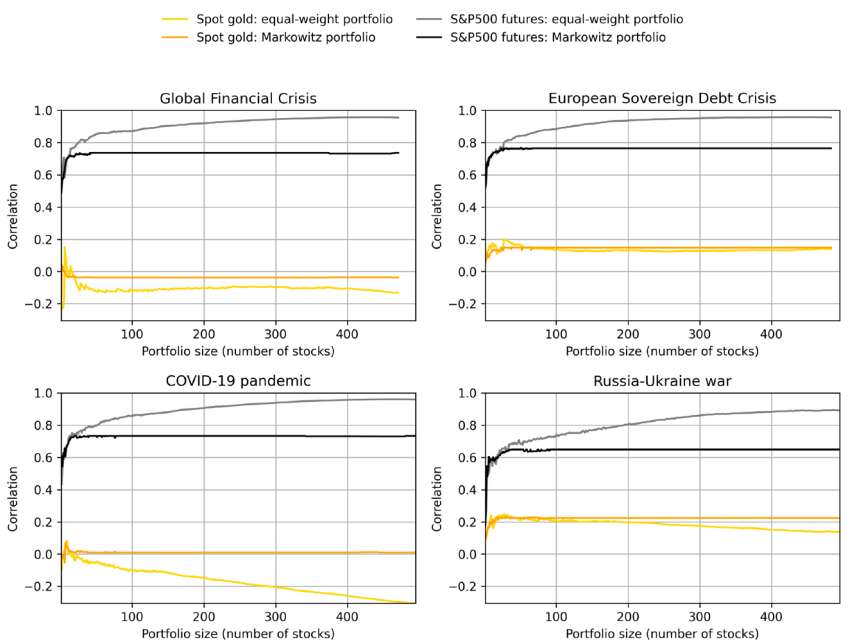
This section focuses on the construction phase, analysing such key factors as correlation and variance used in the static approach. Weights and hedge ratios are calculated on the last day of the construction period and remain fixed throughout the evaluation phase.

4.1. Correlations, variances, weights and hedge ratios during portfolio construction phases

Figure 2 shows the correlations between stock portfolios of different sizes and index futures or gold during construction phases and Figure 3 shows variances of stock portfolio, futures and gold. The correlation between stock portfolios and gold is expected to be close to zero or negative, whereas the correlation between stock portfolios and index futures should be close to one for effective hedging. Correlations with Markowitz portfolios tend to stabilise when the number of assets exceeds a few dozen constituents, as beyond a certain portfolio size, the optimisation algorithm adds new stocks but assigns them zero weight. For equally-weighted portfolios, correlations with gold decrease, while correlations with index futures increase along with the number of constituents.

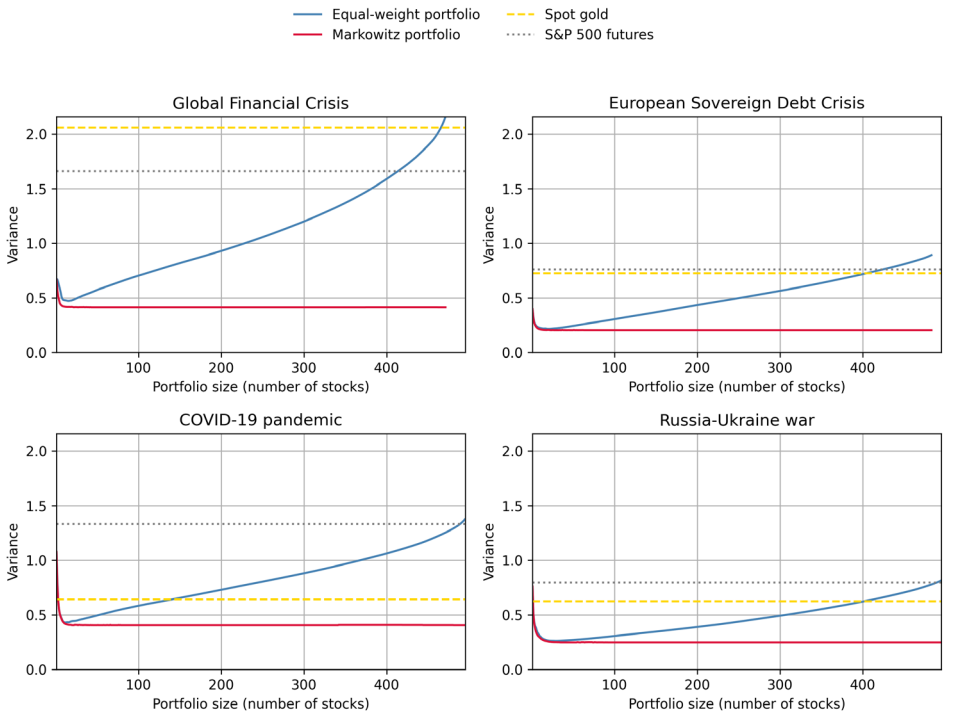
Markowitz portfolios consistently exhibit lower variance than equally-weighted portfolios (see Figure 3). The variance of a Markowitz portfolio consisting of several stocks becomes nearly constant. This supports the findings of Elton and Gruber (1977) and Eom et al. (2021) that a portfolio should consist of maximum 20–50 stocks to significantly reduce the unsystematic risk through diversification. To further reduce variance, the portfolios will be combined with a safe-haven asset in a long position or an index futures contract in a short position.

Figure 2. Correlations between stock portfolios and gold, and between stock portfolios and index futures during the construction period



Source: authors' work.

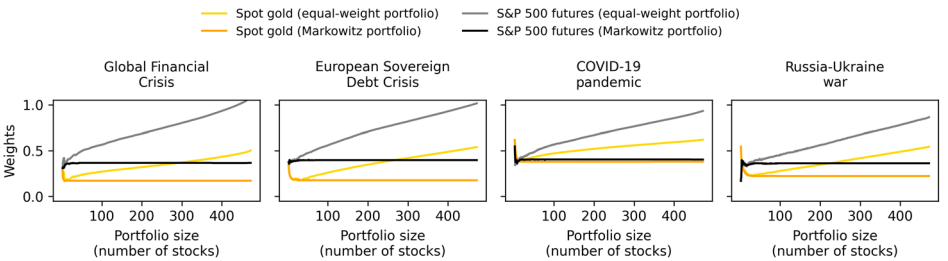
Figure 3. Variances of stock portfolios, gold and S&P 500 futures across the construction phases



Source: authors' work.

Figure 4 presents the asset weights in portfolios composed of a stock portfolio and a protection asset. For the Markowitz portfolio, the compositions are fairly stable, except for small portfolios. Equally-weighted portfolios, on the other hand, change with each added stock. The weight of gold in an equally-weighted portfolio reaches minimum for a portfolio consisting of several stocks and then increases as the portfolio size expands. The hedging ratio for an equally-weighted portfolio systematically converges to one as the portfolio size increases, indicating full hedging. The highest gold weights are observed in the year preceding the onset of the COVID-19 pandemic. This results, on the one hand, from the lowest correlation between the constructed portfolios and gold (Figure 2) and, on the other hand, from the exceptionally low volatility of gold (Figure 3). Assuming these relationships persist throughout the evaluation period, gold may be expected to be the most effective safe haven during this crash period.

Figure 4. Portfolio weights (hedging ratios in the case of the hedging strategy) for the considered strategies



Source: authors' work.

4.2. Effectiveness of protection strategies

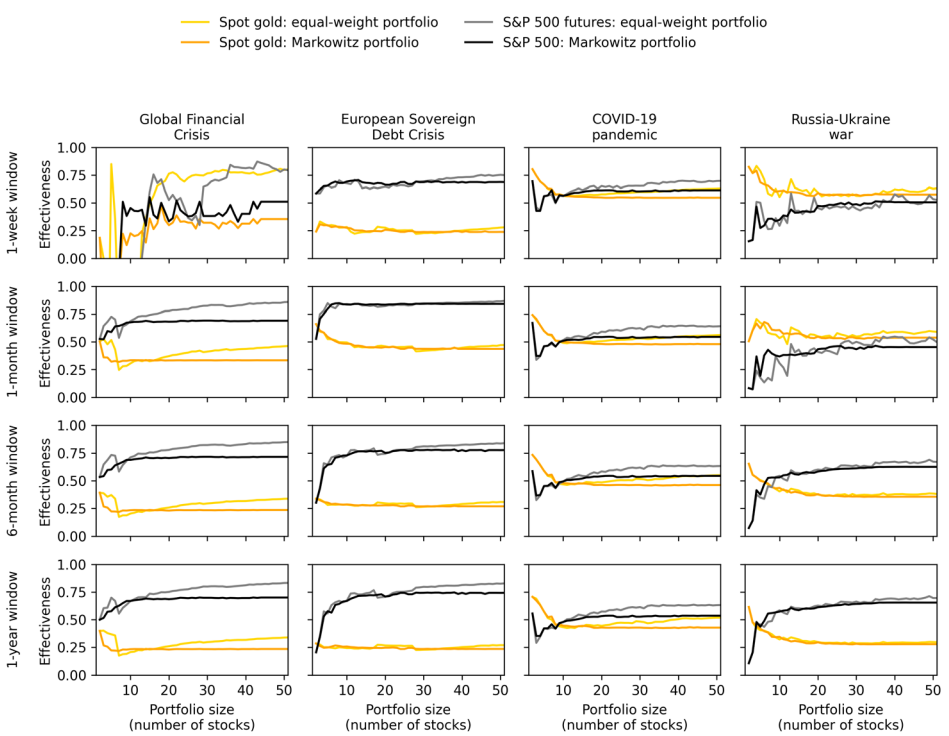
Given the predetermined portfolio allocations, we examine the out-of-sample performance of the protected portfolio across four time horizons: one week, one month, six months and one year, beginning at the onset of the market crash. The evaluation focuses on portfolios containing no more than 50 stocks, as clear trends form in this range and small portfolios exhibit the highest volatility.

The effectiveness shown in Figure 5 depends on the chosen portfolio construction method, the portfolio size, the tested crash and the time horizon of the analysis. For small portfolios, the effectiveness measure fluctuated considerably, especially at the beginning of the GFC. During its first week, gold proved to be an effective safe haven, in particular for an equal-weighted stock portfolio with at least a dozen constituents. Its effectiveness was on average 72% (for portfolios of 15–30 stocks), compared to around 55% for futures contract. In the other horizons considered, the S&P 500 futures contract reduced portfolio volatility much more effectively, regardless of the portfolio's size. During the ESDC, hedging with futures definitely outperformed the safe-haven strategy with the effectiveness higher by 35%–57% (for portfolios of 11–50 stocks), regardless of the hedging horizon. During the COVID-19 pandemic, gold offered superior protection for small portfolios (up to nine constituents). For larger portfolios, hedging turned out as the more effective risk-mitigation strategy, although the differences between both approaches were considerably less pronounced than during the two preceding crises. The analysis of the last crash, caused by the Russia-Ukraine war, yields more ambiguous results. Gold provided better protection during the initial phase of the war (week and month). The effectiveness advantage of gold over futures was significantly higher (by 19%–67%) for portfolios of 1–10 stocks, but this advantage diminished as the size of the portfolio increased. In the long run, gold also outperformed futures for portfolios with a few constituents. However, for larger

portfolios it was the opposite: the larger the portfolio, the greater the efficiency advantage in favour of the hedging strategy.

In all the cases, variance reduction was greater for equally-weighted portfolios than for Markowitz portfolios. This could be expected, as equally-weighted portfolios are typically characterised by higher variance. Furthermore, when hedging with index futures, the reduction in the variance increases along with the growing portfolio size. However, for safe-haven assets, the variance reduction mostly decreases as the portfolio grows. The most effective strategy is hedging an equally-weighted stock portfolio with index futures. This result is intuitive: as the number of stocks increases, the equally-weighted portfolio converges towards the S&P 500 index, which can then be effectively hedged with index futures. In contrast, the safe-haven strategy is found to be the least effective for the Markowitz portfolio among all strategies considered. A safe-haven strategy is essentially diversification through a specific asset. For a well-diversified portfolio, gold is not able to further significantly enhance the diversification effect.

Figure 5. Effectiveness (static approach – 1-year construction window)



Note. The figure illustrates the effectiveness of the protection strategies calculated according to Equation (6).
Source: authors' work.

The results presented in Table 2 show the average effectiveness of the considered protection strategies for underlying portfolios consisting of 1–50 constituents. Comparisons across crises should be interpreted with some caution, as each of these periods was characterised by a different level of underlying risk. It is therefore more informative to examine effectiveness across evaluation horizons. Except for the weekly horizon, which is very short and produces highly varied results, a general pattern can be observed: as the protection horizon increases, in most cases the effectiveness of the protection strategies tends to decrease. This likely reflects market conditions changing over time, while the optimal strategy does not fully adjust to these dynamics, leading to a gradual loss of effectiveness.

Table 2. Average effectiveness of protection strategies

		Static approach				Dynamic approach			
Evaluation period	Crisis	Gold EW	Sp500 EW	Gold M	SP500 M	Gold EW	SP500 EW	Gold M	SP500 M
1W	GFC	0.59	0.51	0.28	0.38	0.55	0.67	0.21	0.76
	ESDC	0.26	0.69	0.25	0.68	0.40	0.70	0.40	0.70
	COVID-19	0.61	0.64	0.57	0.60	0.93	0.71	0.92	0.63
	RUS_INV	0.61	0.45	0.60	0.45	0.71	0.42	0.61	0.39
1M	GFC	0.40	0.78	0.34	0.67	0.60	0.85	0.59	0.81
	ESDC	0.47	0.83	0.46	0.83	0.66	0.88	0.65	0.87
	COVID-19	0.54	0.59	0.50	0.53	0.77	0.56	0.76	0.48
	RUS_INV	0.58	0.42	0.56	0.41	0.57	0.34	0.54	0.33
6M	GFC	0.29	0.78	0.24	0.70	0.43	0.83	0.41	0.80
	ESDC	0.29	0.77	0.28	0.74	0.35	0.78	0.33	0.75
	COVID-19	0.52	0.58	0.48	0.52	0.73	0.56	0.72	0.48
	RUS_INV	0.40	0.58	0.39	0.57	0.53	0.61	0.53	0.62
1Y	GFC	0.29	0.75	0.24	0.68	0.42	0.79	0.39	0.77
	ESDC	0.25	0.74	0.24	0.70	0.31	0.75	0.29	0.70
	COVID-19	0.49	0.57	0.45	0.51	0.66	0.55	0.64	0.47
	RUS_INV	0.32	0.61	0.31	0.60	0.42	0.69	0.42	0.69

Note. This table presents the average effectiveness of the analysed strategies. The left panel (static approach) shows the results for the strategies described in this section, while the right panel the results for the strategies described in Section 5.2. EW denotes equal-weighted portfolios, while M refers to Markowitz portfolios.

Source: authors' work.

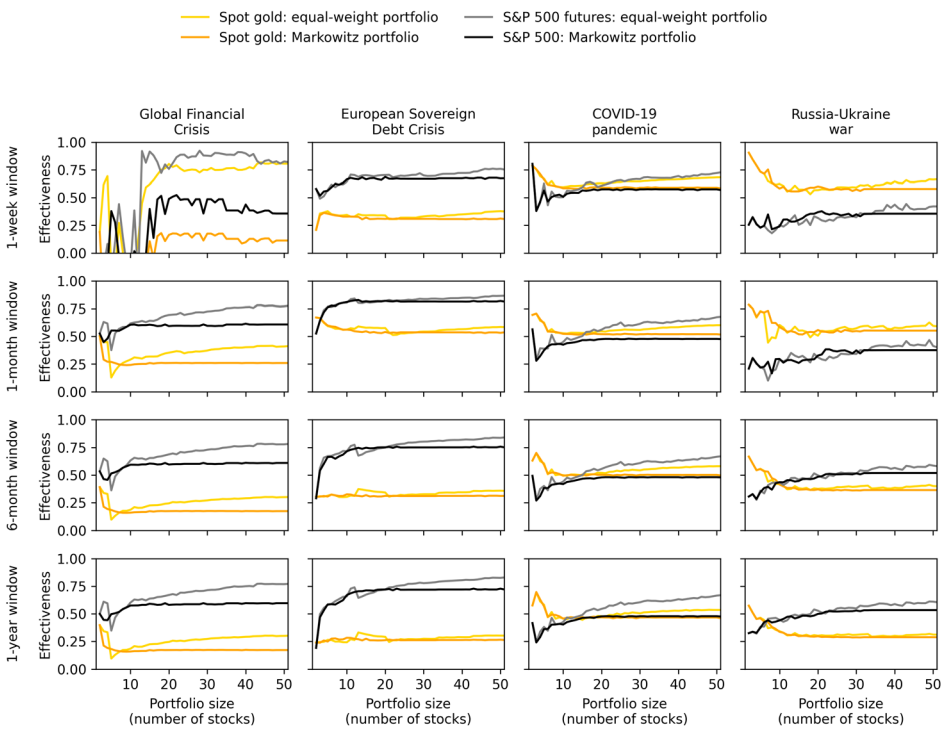
5. Robustness check

5.1. Effectiveness of protection strategies: 6-month construction period

In this subsection, we present the results as in the baseline approach, replacing the annual construction period with a half-year period. The results of the protection effectiveness are presented in Figure 6. Overall, the findings remain largely unchanged, indicating that the results are robust to the length of the portfolio formation window. The relative effectiveness of hedging and safe-haven strategies

across different crashes, portfolio sizes and investment horizons remains virtually unchanged compared to the baseline specification, with only negligible differences observed.

Figure 6. Effectiveness (static approach – 6-month construction window)



Note. The figure illustrates the effectiveness of the protection strategies calculated according to Equation (6).
Source: authors' work.

5.2. Effectiveness of protection strategies – dynamic approach

In this subsection, we extend the baseline analysis by allowing the protection strategy to be rebalanced on a daily basis. On the one hand, this approach does not restrict the evaluation of the strategy's effectiveness to only two dates, thereby enabling investors to adjust their decisions in a more flexible and potentially more efficient manner. On the other hand, more frequent rebalancing generates higher expenses and may reduce investment profitability (Latoszek & Ślepaczuk, 2020). While in the baseline approach transaction costs have only a limited impact on the effectiveness of the strategy, as they occur only once at the initial stage, daily rebalancing would substantially increase their importance. Nevertheless, in this part of the study we do not consider transaction costs either. Transaction costs in the futures market are typically charged on a per-contract

basis rather than as a percentage of transaction value, as is commonly the case in equity markets. Accounting for such costs would therefore require additional assumptions regarding the portfolio size.

We compare the effectiveness of both strategies using a dynamic approach based on the Engle (2002) DCC model. Model-driven futures-hedging strategies are not explicitly addressed in this study. However, related evidence described in Michańków et al. (2023) shows that the effectiveness of hedging depends to a large extent on the choice of signal-generation methods within algorithmic investment strategies. Their results indicate that model-based forecasts can materially affect diversification and hedging performance, particularly in turbulent markets. However, incorporating such model-specific elements introduces a range of additional issues that move the analysis away from the generality of the presented results.

Let us denote a two-dimensional vector by $\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, \epsilon_{2,t})'$. The DCC model assumes that:

$$\boldsymbol{\epsilon}_t | \Omega_{t-1} \sim N(\mathbf{0}, \boldsymbol{H}_t), \boldsymbol{H}_t = \boldsymbol{D}_t \boldsymbol{R}_t \boldsymbol{D}_t, \tag{7}$$

where $\boldsymbol{D}_t = diag(\sqrt{h_{11,t}}, \sqrt{h_{22,t}})$ and conditional variance $h_{ii,t}$ is modelled using the GARCH-type model. The conditional correlation matrix $\boldsymbol{R}_t$ is expressed by

$$\boldsymbol{R}_t = (diag(\boldsymbol{Q}_t))^{-\frac{1}{2}} \boldsymbol{Q}_t (diag(\boldsymbol{Q}_t))^{-\frac{1}{2}}, \tag{8}$$

with

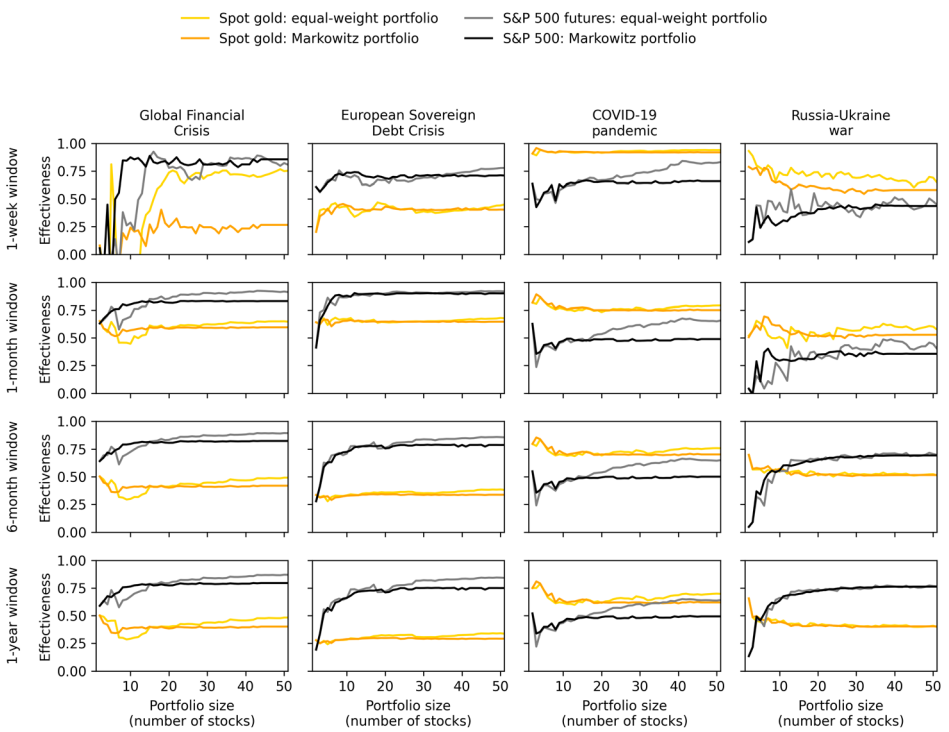
$$\boldsymbol{Q}_t = (1 - a - b)\boldsymbol{Q}^* + a\boldsymbol{z}_{t-1}\boldsymbol{z}'_{t-1} + b\boldsymbol{Q}_{t-1}, \tag{9}$$

where $\boldsymbol{Q}^*$ is the unconditional covariance matrix of z_t $\left(z_{i,t} = \epsilon_{i,t}/\sqrt{h_{ii,t}}\right)$, and a, b are parameters such that $a, b \geq 0$ and $a + b < 1$.

We use the standard GARCH(1,1) approach proposed by Bollerslev (1986) with Gaussian innovations to model the conditional volatility. We extend the estimation window to two years to provide a sufficient amount of data for this model (Hafner & Reznikova, 2012). For the starting sample (two years prior to the onset of each crash), we estimate the parameters of the model and compute one-day-ahead forecasts of the conditional covariance matrix, which are necessary for calculating the hedge ratios and the weights of gold. Subsequently, the estimation sample is updated by including a new observation and removing the oldest one. Next, we re-estimate the model and generate forecasts using the updated estimation sample. This algorithm is repeated iteratively until forecasts are obtained for the last day of the evaluation window.

The entire procedure enables us to compute the forecast of returns of the protected portfolio. Similarly to the baseline approach, portfolio variances and, consequently, effectiveness are computed using the same weekly, monthly, semi-annual and annual horizons starting with the crash. The final results are presented in Figure 7. The effectiveness shown in these plots behaves in the same way as in the static approach during the first two crises and the last one. The hedging strategy outperforms the safe-haven strategy, with minor exceptions already discussed in Subsection 4.2. During the COVID-19 pandemic, investing in gold outperformed the hedging strategy in all cases. However, the differences in the efficiency between the strategies became less pronounced as the number of portfolio constituents was growing.

Figure 7. Effectiveness (dynamic approach)



Note. This figure illustrates the effectiveness of the protection strategies calculated according to Equation (6).
Source: authors' work.

The results presented in the right panel of Table 2 for the dynamic strategy are similar to those obtained for the static strategy. Excluding the weekly horizon, a longer protection horizon is associated with a gradual decline in effectiveness. However, when comparing the static and dynamic approaches, the latter exhibits higher effectiveness in the majority of cases. Rebalancing the optimal protection strategy is

associated with a greater ability to follow market trends and to adjust the investment allocation to changing market conditions. However, in practice such a strategy would also generate transaction costs, which in turn would reduce its net effectiveness.

5.3. Maximum drawdown

Although variance is the most widely used risk measure in the context of this study, it represents only the volatility dimension of the overall risk. Therefore, in this section we analyse the level of risk reduction, understood as the portfolio loss mitigated through the implementation of a protection strategy. Maximum drawdown (MDD) was selected as the risk measure, as it represents the largest peak-to-trough decline in portfolio value over a given period. In practice, this measure represents the worst-case scenario for an investment over a specific time horizon:

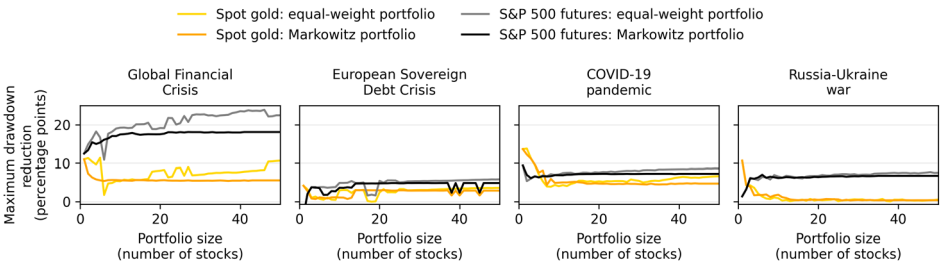
$$\mathrm{MDD} = \frac{V_{\mathrm{peak}} - V_{\mathrm{trough}}}{V_{\mathrm{peak}}}. \tag{10}$$

In our study, peak value V_{peak} corresponds to the last day of the construction period, while the specific time horizon refers to the full evaluation period, namely one year. We introduce the effectiveness of the protection strategy in the following way:

$$Effectiveness = \text{MDD of a stock portfolio} - \text{MDD of a protected portfolio}. \tag{11}$$

The final results showing the effectiveness of the protection strategies are shown in Figure 8. The maximum drawdown duration (the number of days from peak to trough) differed only slightly across the portfolios, regardless of the portfolio size. For the unprotected portfolios, the most frequent (modal) maximum drawdown duration was 171 days during the GFC, 9 days during the ESDC, 12 days during COVID-19 and 163 days during the war. The plots presented in this figure strongly resemble those in Figure 5 for variance. The similarity is particularly visible for the semi-annual/annual periods during the GFC and the war, as well as for the weekly/monthly periods during COVID-19, which corresponds to the mode of the MDD duration. Regardless of the type of crisis, hedging effectiveness exceeded that of investing in gold for portfolios containing at least several assets. During the ESDC, hedging with index futures generally outperformed the safe-haven strategy, although the differences between the two were less pronounced than under the variance-minimisation criterion.

Figure 8. Effectiveness measured using the maximum drawdown metric



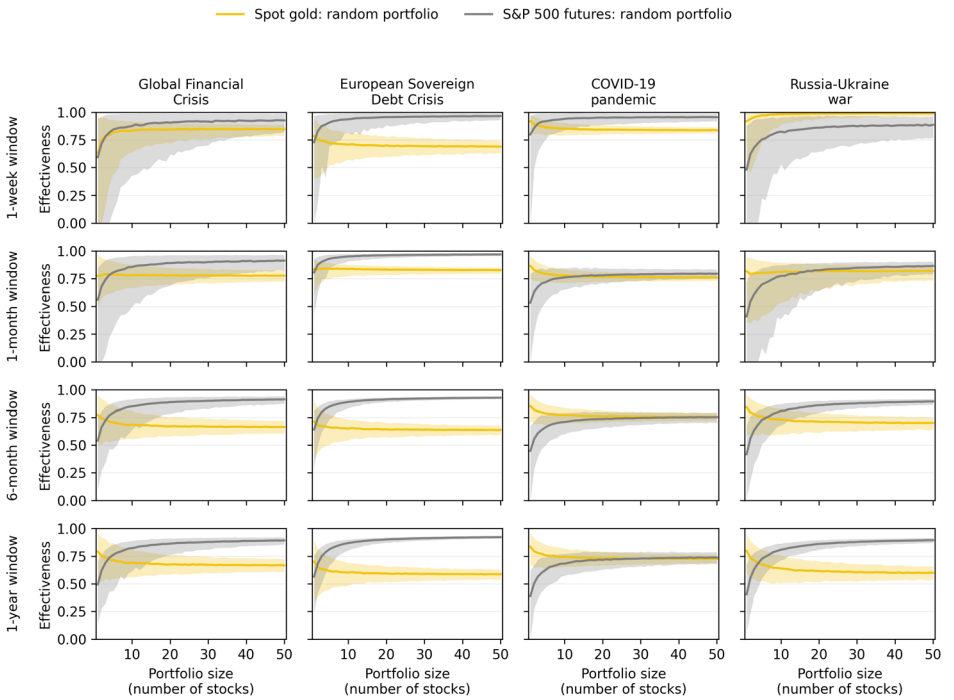
Note. The figure illustrates the effectiveness of the protection strategies calculated according to Equation (11).
Source: authors' work.

5.4. Random portfolios

The main objective of this study is to compare the effectiveness of hedging strategies with investment in a safe-haven asset. The vast majority of previous studies address this problem from the perspective of variance minimisation; therefore, equity portfolios are constructed following the same criterion. In this subsection, we compare the effectiveness of these strategies relative to equally-weighted portfolios generated using Monte Carlo simulations. Portfolios of varying sizes were randomly selected from all constituents of the S&P 500 index and their risk is mitigated using either index futures or gold. This process is repeated 1,000 times to ensure robust statistical results. The results are indicated in Figure 9 as shaded bands representing the 95% confidence interval, with the central line denoting the median.

Overall, the shape of the plots closely resembles those obtained previously for minimum variance portfolios, suggesting that the observed relationships are robust to the portfolio construction method. As shown in Figure 9, the confidence intervals for effectiveness do not overlap for the semi-annual and annual evaluation horizons and for larger portfolios, indicating that the difference between the two protection strategies is statistically significant. Hedging effectiveness increases as the number of assets in the portfolio grows, while the effectiveness of gold as a safe haven gradually declines. The exception is the pandemic period, where the differences between the effectiveness of the two strategies are not statistically significant (the confidence intervals overlap), except for the weekly horizon. In contrast, during ESDC, hedging outperforms investing in a safe haven in all evaluation periods for portfolios containing more than a few stocks. An important finding is that there is no case, at the 5% significance level, in which investing in gold is more effective for risk mitigation than hedging with futures contracts.

Figure 9. Effectiveness for random portfolios



Note. This figure illustrates the effectiveness of the protection strategies calculated according to Equation (6).
Source: authors' work.

6. Conclusions

This study compares the effectiveness of a hedging strategy relying on S&P 500 index futures with a safe-haven strategy based on gold, according to the definition of a safe haven proposed by Baur and Lucey (2010). Therefore, risk minimisation is the only criterion common to all of the considered strategies.

The empirical results indicate that hedging with index futures generally provided more effective protection, particularly for larger and equally-weighted portfolios. Gold as a safe-haven outperformed hedging with futures only for small portfolios and over short time horizons, particularly in the early stages of the GFC and during the Russia-Ukraine war; however, this result is not statistically significant. In general, if risk minimisation is the only criterion considered by investors, hedging with index futures is recommended as a more effective strategy. This result is particularly robust over longer horizons (six months to a year) and becomes stronger as the level of diversification increases. The evidence remains largely consistent across the range of the conducted robustness tests and carries important implications for investors and

portfolio managers. These findings align with the traditional hedging theory, which states that derivative instruments such as index futures provide direct and efficient risk reduction due to their strong linkage with the underlying asset and their ability to precisely replicate market exposure. The results also suggest that theoretical definitions based solely on correlation may be insufficient, as they fail to fully capture the economic effectiveness of hedging strategies under realistic portfolio constraints and they do not reflect investors' true expectations toward safe-haven assets. Analysing Bitcoin's ability to serve as a safe haven, Baur et al. (2022) highlight the limitations of a correlation-based definition by showing that extreme volatility can destroy its hedging properties even when the correlation is negative. Conlon and McGee (2020) confirm this finding, showing that including highly volatile instruments in stock portfolios increases downside risk exposure despite their negative correlation with equities. Incorporating potential profits and evaluating the utility of a protective strategy may better reflect investors' expectations, in which case the assessment presented in this study would likely be different. It is worth noting that in the recent literature, in addition to the traditional approach based on risk minimisation, an alternative definition of a safe-haven asset has emerged, which is based on the maximisation of the expected utility in the context of the prospect theory (Echaust et al., 2026). This perspective will constitute an important direction for our further research.

References

Ali, S., Bouri, E., Czudaj, R. L., & Shahzad, S. J. H. (2020). Revisiting the valuable roles of commodities for international stock markets. *Resources Policy*, *66*, 101603. http://dx.doi.org/10.1016/j.resourpol.2020.101603.

Anas, M., Bouri, E., & Shahzad, S. J. H. (2024). A bibliometric analysis of literature on hedge and safe haven assets. *Journal of Economic Surveys*, *39*(5), 1852–1882. http://dx.doi.org/10.1111/joes.12677.

Banerjee, A. K., & Pradhan, H. (2024). Did precious metals serve as hedge and safe-haven alternatives to equity during the COVID-19 pandemic: New insights using a copula-based approach. *Journal of Emerging Market Finance*, *23*(4), 399–423. https://doi.org/10.1177/09726527241251515.

Batten, J. A., Kinateder, H., Szilagyi, P .G., & Wagner, N. F. (2021). Hedging stocks with oil. *Energy Economics*, *93*, 104422. http://dx.doi.org/10.1016/j.eneco.2019.06.007.

Baur, D. G., Hoang, L. T., & Hossain, M. Z. (2022). Is Bitcoin a hedge? How extreme volatility can destroy the hedge property. *Finance Research Letters*, *47*(B), 102655. https://doi.org/10.1016/j.frl.2021.102655.

Baur, D. G., & Lucey, B. M. (2010). Is Gold a Hedge or a Safe Haven? An Analysis of Stocks, Bonds and Gold. *The Financial Review*, *45*(2), 217–229. https://doi.org/10.1111/j.1540-6288.2010.00244.x.

Baur, D. G., & McDermott, T. K. (2010). Is gold a safe haven? International evidence. *Journal of Banking & Finance*, *34*(8), 1886–1898. http://dx.doi.org/10.1016/j. jbankfin.2009.12.008.

Beckmann, J., Berger, T., & Czudaj, R. (2015). Does gold act as a hedge or a safe haven for stocks? A smooth transition approach. *Economic Modelling*, *48*, 16–24. http://dx.doi.org/10.1016/j .econmod.2014.10.044.

Będowska-Sójka, B., & Kliber, A. (2021). Is there one safe-haven for various turbulences? The evidence from gold, Bitcoin and Ether. *The North American Journal of Economics and Finance*, *56*, 101390. https://doi.org/10.1016/j.najef.2021.101390.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. http://dx.doi.org/10.1016/0304-4076(86) 90063-1.

Boubaker, H., Cunado, J., Gil-Alana, L. A., & Gupta, R. (2020). Global crises and gold as a safe haven: Evidence from over seven and a half centuries of data. *Physica A: Statistical Mechanics and its Applications*, *540*, 123093. http://dx.doi.org/10.1016/j. physa.2019.123093.

Cho, D., & Han, H. (2021). The tail behavior of safe haven currencies: A cross-quantilogram analysis. *Journal of International Financial Markets, Institutions and Money*, *70*, 101257. http://dx.doi.org/10.1016/j.intfin.2020.101257.

Conlon, T., & McGee, R. (2020). Safe haven or risky hazard? Bitcoin during the COVID-19 bear market. *Finance Research Letters*, *35*, 1–5. http://dx.doi.org/10. 1016/j.frl.2020.101607.

Disli, M., Nagayev, R., Salim, K., Rizkiah, S .K., & Aysan, A. F. (2021). In search of safe haven assets during COVID-19 pandemic: An empirical analysis of different investor types. *Research in International Business and Finance*, *58*, 1–24. http://dx.doi.org/10.1016/j.ribaf.2021.101461.

Dong, X., Li, C., & Yoon, S. M. (2021). How can investors build a better portfolio in small open economies? Evidence from Asia's Four Little Dragons. *The North American Journal of Economics and Finance*, *58*, 101500. http://dx.doi.org/10.1016/ j.najef.2021.101500.

Drobetz, W., Schröder, H., & Tegtmeier, L. (2020). The role of catastrophe bonds in an international multi-asset portfolio: Diversifier, hedge, or safe haven?. *Finance Research Letters*, *33*, 101198. http://dx.doi.org/10.1016/j.frl.2019.05.016.

Echaust, K., & Just, M. (2022). Is gold still a safe haven for stock markets? New insights through the tail thickness of portfolio return distributions. *Research in International Business and Finance*, *63*, 1–19. http://dx.doi.org/10.1016/j.ribaf.2022.101788.

Echaust, K., Just, M., & Kliber, A. (2024). To hedge or not to hedge? Cryptocurrencies, gold and oil against stock market risk. *International Review of Financial Analysis*, *94*, 103292. http://dx.doi.org/10.1016/j.irfa.2024.103292.

Echaust, K., Just, M., & Musti, S. (2026). Hedge and safe-haven assets for stock markets revisited: Evidence from the prospect theory. *International Review of Economics and Finance*, *108*, 1–21. https://doi.org/10.1016/j.iref.2026.105300.

Ederington, L. H. (1979). The Hedging Performance of the New Futures Markets. *The Journal of Finance*, *34*(1), 157–170. http://dx.doi.org/10.1111/j.1540-6261.1979.tb02077.x.

Elton, E. J., & Gruber, M. J. (1977). Risk Reduction and Portfolio Size: An Analytical Solution. *The Journal of Business*, *50*(4), 415–437. https://doi.org/10.1086/295964.

Engle, R. (2002). Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models. *Journal of Business & Economic Statistics*, *20*(3), 339–350. https://doi.org/10.1198/073500102288618487.

Eom, C., Kaizoji, T., Livan, G., & Scalas, E. (2021). Limitations of portfolio diversification through fat tails of the return distributions: Some empirical evidence. *The North American Journal of Economics and Finance*, *56*, 101358. http://dx.doi.org/10.1016/j.najef.2020.101358.

Fałdziński, M., & Pietrzak, M. B. (2015). The multivariate DCC-GARCH model with interdependence among markets in conditional variances' equations. *Przegląd Statystyczny. Statistical Review*, *62*(4), 397–413. https://doi.org/10.5604/01.3001.0014.1763.

Forbes, K. J., & Rigobon, R. (2002). No contagion, only interdependence: Measuring stock market comovements. *The Journal of Finance*, *57*(5), 2223–2261. https://doi.org/10.1111/0022-1082.00494.

Goodell, J. W., & Goutte, S. (2021). Diversifying equity with cryptocurrencies during COVID-19. *International Review of Financial Analysis*, 76, 101781. http://dx.doi.org/10.1016/j.irfa.2021.101781.

Hafner, C. M., & Reznikova, O. (2012). On the estimation of dynamic conditional correlation models. *Computational Statistics & Data Analysis*, *56*(11), 3533–3545. http://dx.doi.org/10.1016/j.csda.2010.09.022.

Ji, Q., Zhang, D., & Zhao, Y. (2020). Searching for safe-haven assets during the COVID-19 pandemic. *International Review of Financial Analysis*, *71*, 1–10. https://doi.org/10.1016/j.irfa.2020.101526.

Just, M., & Echaust, K. (2024). Cryptocurrencies against stock market risk: New insights into hedging effectiveness. *Research in International Business and Finance*, *67*(A), 1–26. http://dx.doi.org/10.1016/j.ribaf.2023.102134.

Klein, T. (2017). Dynamic correlation of precious metals and flight-to-quality in developed markets. *Finance Research Letters*, *23*, 283–290. https://doi.org/10.1016/j.frl.2017.05.002.

Kliber, A., Marszałek, P., Musiałkowska, I., & Świerczyńska, K. (2019). Bitcoin: Safe haven, hedge or diversifier? Perception of bitcoin in the context of a country's economic situation – A stochastic volatility approach. *Physica A: Statistical Mechanics and its Applications*, *524*, 246–257. http://dx.doi.org/10.1016/j.physa.2019.04.145.

Latoszek, M., & Ślepaczuk, R. (2020). Does the inclusion of exposure to volatility into diversified portfolio improve the investment results? Portfolio construction from the perspective of a Polish investor. *Economics and Business Review*, *6*(1), 46–81. https://doi.org/10.18559/ebr.2020.1.3.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91. http://dx.doi.org/10.1111/j.1540-6261.1952.tb01525.x.

Michańków, J., Sakowski, P., & Ślepaczuk, R. (2023). *Hedging Properties of Algorithmic Investment Strategies using Long Short-Term Memory and Time Series models for Equity Indices*. https://arxiv.org/abs/2309.15640.

Mizerka, J., Stróżyńska-Szajek, A., & Mizerka, P. (2020). The role of Bitcoin on developed and emerging markets – on the basis of a Bitcoin users graph analysis. *Finance Research Letters, 35*, 1–8. https://doi.org/10.1016/j.frl.2020.101489.

Pisedtasalasai, A. (2021). Hedging stocks in crises and market downturns with gold and bonds: Industry analysis. *Asian Economic and Financial Review*, *11*(1), 1–16. http://dx.doi.org/10.18488/JOURNAL.AEFR.2021.111.1.16.

Ryan, M., Corbet, S., & Oxley, L. (2024). Is gold always a safe haven?. *Finance Research Letters*, *64*, 1–6. https://doi.org/10.1016/j.frl.2024.105438.

Sandoval, L., & Franca, I. D. P. (2012). Correlation of financial markets in times of crisis. *Physica A: Statistical Mechanics and its Applications*, *391*(1–2), 187–208. https://doi.org/10.1016/j.physa.2011.07.023.

Siemaszkiewicz, K. (2023). Alternative investments during turbulent times – a comparison of dynamic relationship. *Przegląd Statystyczny. Statistical Review*, *69*(3), 23–43. http://dx.doi.org/10.5604/01.3001.0016.2377.

Siemaszkiewicz, K., & Let, B. (2020). Looking for alternatives in times of market stress: A tail dependence between the european stock markets and bitcoin, gold and fine wine market. *Czech Journal of Economics and Finance*, *70*(5), 407–430. https://doi.org/10.32065/cjef.2020.05.02.

Wen, X., & Cheng, H. (2018). Which is the safe haven for emerging stock markets, gold or the US dollar?. *Emerging Markets Review*, *35*, 69–90. https://doi.org/10.1016/j.ememar.2017.12.006.