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

Przegląd Statystyczny
Statistical Review

Vol. 69 | No. 4 | 2022

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 manuscripts considered for publication should significantly contribute to the theoretical aspects of the aforementioned fields or shed new light on the practical applications of these aspects. Manuscripts reporting 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 submission on the Editor-in-Chief's request.

Since 1 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 work 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. 69 No. 4 2022

EDITORIAL OFFICE'S ADDRESS

University of Gdańsk, ul. Armii Krajowej 101, 81-824 Sopot, Poland

Language editing: Scientific Journals Division, Statistics Poland
Technical editing and typesetting: Statistical Publishing Establishment – team supervised by Maciej Adamowicz

Printed and bound: Statistical Publishing Establishment
al. Niepodległości 208, 00-925 Warsaw, Poland, zws.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 Publishing Establishment
phone no. +48 22 608 32 10, +48 22 608 38 10

Order no. 51/2022 – 235 printed copies

CONTENTS

Piotr Szczepocki
Estimation of Yu and Meyer bivariate stochastic volatility model by iterated filtering **1**

Andrzej Tomski, Barbara Gorzawska
Sample size in clinical trials – challenges and approaches **20**

Honorata Sosnowska, Michał Ramsza, Paweł Zawiślak
Impact of a priori positive information on the results of voting methods **29**

Jakub Morkowski
Comparison of the accuracy of forecasts based on neural networks before and after the outbreak of the COVID-19 pandemic on the example of selected exchange rates **41**

Przegląd Statystyczny. Statistical Review, 2022, vol. 69, 4, 1–19 https://doi.org/10.59139/ps.2022.04.1
Received: 11.02.2023, revised: 17.03.2023, accepted: 20.03.2023

Estimation of Yu and Meyer bivariate stochastic volatility model by iterated filtering

Piotr Szczepocki[a]

Abstract. In financial applications, understanding the asset correlation structure is crucial to tasks such as asset pricing, portfolio optimisation, risk management, and asset allocation. Thus, modelling the volatilities and correlations of multivariate stock market returns is of great importance.

This paper proposes the iterated filtering algorithm for estimating the bivariate stochastic volatility model of Yu and Meyer. The iterated filtering method is a frequentist-based approach that utilises particle filters and can be applied to estimating the parameters of non-linear or non-Gaussian state-space models.

The paper presents an empirical example that demonstrates the way in which the proposed estimation method might be used to estimate the correlation between the returns of two assets: Standard and Poor's 500 index and the price of gold in US dollars. This is accompanied by a simulation study that proves the validity of the approach.

Keywords: multivariate stochastic volatility, iterated filtering, particle filters

JEL: C32, C58, G15

1. Introduction

The knowledge of correlation structures is vital in many financial applications, because it provides a measurement of the relationship between different financial assets or variables. This information can be used to make informed investment decisions, assess risk, and design and evaluate financial products. In a portfolio construction, knowledge of the correlation structure between assets can help investors create a well-diversified portfolio. If the assets are highly correlated, their returns are likely to move in the same direction, which can lead to a higher portfolio risk. On the other hand, if the assets are uncorrelated or negatively correlated, their returns may offset each other, reducing the portfolio risk. In risk management, understanding the correlation structure between different financial variables can help financial institutions assess the potential for losses and determine how to allocate capital to manage risk. This is particularly important for complex financial products, such as derivatives, which can have non-linear and highly interconnected risk profiles. The knowledge of the correlation structure can also be used to price

[a] University of Łódź, Department of Statistical Methods, ul. POW 3/5, 90-255 Łódź, e-mail: piotr.szczepocki@uni.lodz.pl, ORCID: https://orcid.org/0000-0001-8377-3831.

financial products and to develop financial models, e.g. value at risk (VaR) models, which are widely used to measure the risk of financial portfolios.

Numerous applications of the correlation structure generate the need for modelling the volatilities of multivariate stock market returns. Over several past decades, there has been significant progress in the estimation of multivariate volatility models. Nowadays, we can distinguish three main approaches to this problem: multivariate generalised autoregressive conditional heteroskedasticity (GARCH) models (represented by e.g. the BEKK models of Engle and Kroner (1995) and the DCC models of Engle (2002)), multivariate stochastic volatility (MSV) models (see Chib et al., 2009), and realised covariance models (see e.g. Bollerslev et al., 2018; Jin & Maheu, 2013). Recently, machine learning (ML) algorithms have also become increasingly popular for the forecasting of multivariate financial time series (Bejger & Fiszeder, 2021; Fiszeder & Orzeszko, 2021).

The estimation of multivariate stochastic volatility models generates significant difficulties due to both their high-dimensional parameter space resulting from their multidimensional nature, and the absence of a closed-form likelihood in stochastic volatility models. Moreover, the estimation process has to account for the positive semidefiniteness of the covariance matrix.

Despite these challenges, various methods for estimating MSV models have been developed. The first approach involved the application of Kalman-based filtering to the evaluation and maximisation of the quasi-loglikelihood function (Harvey et al., 1994; So et al., 1997). Soon afterwards, the Bayesian approach began to dominate the estimation of MSV models. It was followed by the multi-move sampler proposed by Shephard and Pitt (1997) and modified by Watanabe and Omori (2004), which became a standard technique in the early 2000s. This method was used, among other authors, by Ishihara and Omori (2012) for the estimation of the MSV model with cross-leverage and heavy-tailed errors, Ishihara et al. (2016) for the matrix exponential MSV model with cross-leverage, and Kastner et al. (2017) for the multivariate factor SV model. In the case of MSV models, it is necessary to construct an efficient MCMCM sampler separately in each model (Chib et al., 2009).

Although the Bayesian inferences are very effective, they have some limitations that can restrict their usefulness in certain applications. Firstly, the inferences depend on the choice of prior distributions. Choosing appropriate prior distributions is therefore of key importance in the case of Bayesian methods. Otherwise, estimates might turn out biased, and results incorrect. It is especially advisable to use methods independent of a priori beliefs if the task is to identify the information in the studied data. Secondly, in some cases, the interpretation of Bayesian models can be challenging due to the complexity of the posterior

distributions. In addition, Bayesian methods are more difficult to implement and require more computing power than frequentist approaches.

Frequentist approaches, such as maximum likelihood estimation, are often easier to interpret and are computationally simpler, which renders them more accessible than Bayesian methods, and thus widely used. However, they also have limitations, such as problems with taking into account a priori knowledge. In addition, using them, one cannot get a full picture of the uncertainty of estimates. The choice between the Bayesian and the frequentist approaches is often very subjective. Both have strengths and weaknesses, and the preference of one over the other depends on the specific context and the type of problem addressed.

Frequentist-based statistical inference for MSV is very limited compared to the Bayesian analysis. The quasi-maximum likelihood method of Harvey et al. (1994) is restricted only to models with constant correlation. Jungbacker and Koopman (2006) proposed importance-sampling Monte Carlo techniques for the maximum likelihood estimation of the SV of three specific multivariate extensions of the basic SV model.

In this paper, we propose to use an iterated filtering algorithm (Ionides et al. 2006, 2015) for estimating the bivariate SV model of Yu and Meyer (2006). Iterated filtering is a frequentist-based method based on particle filters that can be used to estimate parameters for general non-linear or non-Gaussian state-space models (SSM). Despite being limited to only a bivariate relationship, the Yu and Meyer model has gained popularity due to its ability to model the dynamic correlation between a pair of assets.

Johansson (2010b) used the Yu and Meyer model to study the systematic risk of sovereign bonds in four East Asian countries, and the relationship between stocks and bonds in nine Asian countries (Johansson, 2010a). Du et al. (2011) applied it to the investigation of volatility spillovers in the crude oil and agricultural commodity markets. Hui and Zheng (2012a, 2012b) examined the correlations between housing and retail property markets in Hong Kong by means of it. Gębka and Karoglou (2013) used the Yu and Meyer model to explore the integration of the European peripheral financial markets with Germany, France, and the UK. Kliber (2011) applied it to the study of the correlation between selected sovereign Central European credit default swaps. Recently, Kliber et al. (2019) used the Yu and Meyer model to check whether Bitcoin can act as a hedge, diversifier or safe haven in five countries (Japan, Venezuela, China, Estonia and Sweden), whereas Będowska-Sójka and Kliber (2021) examined by means of it the safe-haven properties of gold and two cryptocurrencies, Bitcoin and Ether, for four main stock indices (S&P500, FTSE, DAX and STOXX600). All the above-mentioned applications of the Yu and Meyer model utilised the Bayesian approach for parameter estimation.

The most valuable contribution of this paper to the existing body of research is the proposed frequentist-based estimation method for the Yu and Meyer model. The frequentist-based approach may also be applied to the estimation of the filtering distribution of log-volatilities and the dynamic correlation using the standard bootstrap particle filter. The paper further consists of: Section 2 introducing the Yu and Meyer model and categorising it as a special case within the broader class of state-space models, Section 3 presenting the estimation methodology, Section 4 featuring an empirical example, Section 5 where a simulation study is conducted, and Section 6 presenting the conclusions of the study.

2. Yu and Meyer model

Yu and Meyer (2006) proposed a bivariate SV model for which not only volatilities but also correlation coefficients are time-varying. This model describes the evolution of two asset returns through time. Let us denote the observed mean-centred log-returns at time t as $y_t = (y_{1t}, y_{2t})'$ for $t = 1, \dots, T$. Let $h_t = (h_{1t}, h_{2t})'$ be a vector of log-volatiles, $\mu = (\mu_1, \mu_2)'$ a vector of long-term means of log-volatiles, $\Omega_t = diag(\exp(h_{1t}/2), \exp(h_{2t}/2))$ a diagonal matrix of log-return standard deviations, and $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t})'$ and $\eta_t = (\eta_{1t}, \eta_{2t})'$ two vectors of error terms. The model might then be written as:

$$\begin{cases} y_t = \Omega_t \varepsilon_t, \qquad \varepsilon_t | \Omega \sim iid\ N\left(0, \Sigma_{\varepsilon,t}\right), \\ \Sigma_{\varepsilon,t} = \begin{pmatrix} 1 & \rho_t \\ \rho_t & 1 \end{pmatrix}, \\ h_{t+1} = \mu + diag(\phi_1, \phi_2)(h_t - \mu) + \eta_t, \eta_t \sim iid\ N\left(0, diag\left(\sigma_{\eta 1}^2, \sigma_{\eta 2}^2\right)\right), \\ q_{t+1} = \psi_0 + \psi_1(q_t - \psi_0) + \sigma_\eta \upsilon_t, \qquad \upsilon_t \sim iid\ N(0,1), \ \ \rho_t = \dfrac{\exp(q_t) - 1}{\exp(q_t) + 1}, \end{cases} \tag{1}$$

with initial conditions: $h_0 = \mu$ and $q_0 = \psi_0$. Error terms ε_t, η_t and υ_t are independent. Correlation matrix $\Sigma_{\varepsilon,t}$ is well-defined, because the inverse Fisher transformation for q_t constrains ρ_t to interval (–1,1). Log-volatiles h_t follow reverting autoregressive processes to the order one mean. The model is equivalent to the bivariate GARCH model with dynamic conditional correlation (DCC-MGARCH). One significant limitation of this model is that it is difficult to extend it to higher dimensions. The main challenge is to ensure the positive definiteness of correlation matrix $\Sigma_{\varepsilon,t}$. Asai et al. (2006) suggested the following solutions for situations where the dimension of log-returns is larger than two: the Cholesky decomposition (Tsay, 2005), the matrix exponential (Chiu et al., 1996), and the

Wishart models (Gouriéroux, 2006). However, in many practical situations, the principal goal is to examine the temporal correlation between a pair of assets.

There are nine parameters to be estimated in the Yu and Meyer model, namely $\mu_1, \mu_2, \phi_1, \phi_2, \sigma_1, \sigma_2, \psi_0, \psi_1$ and σ_η. The authors employed a Bayesian approach, defining separate prior distributions for each of the considered parameters. They used the WinBUGS, which enables a convenient and efficient implementation of the single-move Gibbs sampler. All the examples of the application of the Yu and Mayer model mentioned in the introduction also use the Bayesian approach through WinBUGS[1] or OpenBUGS (often using the R2WinBUGS and R2OpenBUGS R packages, respectively (Sturtz et al., 2005)). To the author's best knowledge, there has been no attempt to estimate the Yu and Meyer model in the classical inference paradigm so far.

In fact, the Yu and Meyer model is, like most (multivariate) SV models, an example of a broader class of statistical models called state-space models (SSMs). This class provides a general framework for analysing a hidden stochastic process that is measured or observed through another stochastic process. SSMs are very flexible and have been widely applied in economics, ecology, epidemiology, medicine (mainly neuroscience), signal processing and mechanical system monitoring (see Chapter 1 in Cappé et al. (2005) and Chapter 2 in Chopin and Papaspiliopoulos (2020) for details).

More specifically, an SSM consists of a pair of discrete-time processes: $\mathbb{Y}_t = = (Y_t)_{t\geq 0}$, i.e. the measurement process, and $\mathbb{X}_t = (X_t)_{t\geq 0}$, i.e. the latent state process. The observable random variables $\mathbb{Y}_t$ are assumed to be conditionally independent given $\mathbb{X}_t$. According to the definition of SSMs, the latent process model is determined by the set of densities $\left(f_t(x_{t+1}|x_t;\theta)\right)_{t\geq 0}$ and the initial density $f_0(x_0;\theta)$ (i.e. the state process, $\mathbb{X}_t$, is Markovian). The measurement process is determined by the set of densities $\left(g_t(y_t|x_t;\theta)\right)_{t\geq 0}$ (see Chapter 2 of Chopin and Papaspiliopoulos (2020) for a detailed examination of this definition of SSMs and two alternative ones).

In the case of the Yu and Meyer model at time *t*, the latent state process is $X_t = (h_{1t}, h_{2t}, q_t)'$, and the measurement process is $Y_t = (y_{1t}, y_{2t})'$. The latent process density (at time *t*) $f_t(x_{t+1}|x_t;\theta)$, due to independence of error terms η_t and υ_t, may be decomposed as:

$$h_{t+1}|h_t \sim N\left(\mu + diag(\phi_1, \phi_2)(h_t - \mu), diag\left(\sigma_{\eta 1}^2, \sigma_{\eta 2}^2\right)\right), \tag{2}$$

[1] WinBUGS 1.4.3 is available for routine use, but is no longer being developed (Lunn et al., 2009).

$$q_{t+1}|q_t \sim N\left(\psi_0 + \psi_1(q_t - \psi_0), \sigma_\eta^2\right), \tag{3}$$

with the initial densities:

$$h_0 \sim \delta(h_0 - \mu), q_0 \sim \delta(q_0 - \psi_0), \tag{4}$$

where δ is delta Dirac function (i.e. $P(h_0 = \mu) = 1$, $P(q_0 = \psi_0) = 1$). The measurement process density (at time t) may be written as

$$y_t|h_t, q_t \sim N\left(0, \Omega\Sigma_{\varepsilon,t}\Omega'\right), \tag{5}$$

where:

$$\Omega_t = diag(\exp(h_{1t}/2),\ \exp(h_{2t}/2)),\ \Sigma_{\varepsilon,t} = \begin{pmatrix} 1 & \rho_t \\ \rho_t & 1 \end{pmatrix}, \rho_t = \frac{\exp(q_t) - 1}{\exp(q_t) + 1}. \tag{6}$$

The SSM resulting from the Yu and Meyer model is non-linear (exponential transformation for h_t and the inverse Fisher transformation for q_t) and Gaussian. Strong non-linearity excludes the direct use of the Kalman filter. Particle filters are an established method of estimating the latent state in a nonlinear/non-Gaussian SSM when the parameters are fixed. Particle filters also provide an unbiased likelihood estimator, but they cannot be used directly to estimate likelihood, because the estimator of likelihood is not continuous as a function of parameters θ (Malik & Pitt, 2011). Xu and Jasra (2019) proposed a particle filter technique for the inference of high-dimensional SV models with a constant correlation matrix, but the parameter estimation is based on the Bayesian particle marginal Metropolis-Hasting algorithm (Andrieu et al., 2010).

3. Estimation method

Estimating parameters for state-space models (SSMs) is a complex task. The primary obstacle is that the exact computation of likelihood functions is not possible, as it requires evaluating multiple integrals. Another significant difficulty is that SSMs often generate log-likelihoods that are awkward to optimise numerically, for example non-concave, multi-modal, or flat (in certain directions) ones (Chopin & Papaspiliopoulos, 2020, p. 260).

Iterated filtering was pioneered by Ionides et al. (2006), and theoretically substantiated by Ionides et al. (2011). The second generation of iterated filtering, IF2, was introduced by Ionides et al. (2015) and developed by a theoretical study of

Nguyen (2016). Even though both generations of iterated filtering employ a recursive filtering approach through an augmented model, their theoretical foundations differ. The first generation (IF1) approximates the Fisher score function, while the second one (IF2) combines the concept of data cloning (Lele et al., 2007) with the convergence of an iterated Bayes map (Nguyen, 2016). Empirical studies showed that the IF2 outperforms the IF1 (Ionides et al., 2015). The calculations in this article used exclusively the second generation of the algorithm.

Iterated filtering has been successfully applied to many SSMs, mostly in the context of epidemiology (Bhadra et al., 2011; He et al., 2009; King et al., 2008; Stocks et al., 2020; You et al., 2020), but also in economic modelling, especially for univariate SV models (Bretó, 2014; Szczepocki, 2020).

Iterated filtering is a technique that utilises particle filters and involves replacing the model we are interested in with a similar model, but with parameters that take a random walk in time. This extra variability smooths the likelihood surface (which is the main impediment for particle filters in parameter estimation) and counteracts particle depletion. Over multiple repetitions of the filtering procedure (each made using a particle filter), the variance of this random walk goes to zero and the augmented model approaches the original one. As a result, iterated filtering provides a sequence of iteratively updated parameter estimates that converge to the maximum likelihood estimate (see Ionides et al., 2015; Nguyen, 2016, for details). Thus the algorithm is *likelihood-based*.

In practical applications, the convergence of the algorithm is often assessed via diagnostic plots (see e.g. Bretó, 2014; King et al., 2008; Szczepocki, 2020). Iterated filtering uses only a basic bootstrap particle filter (Gordon et al., 1993), and thus it does not have to evaluate the transition density $f_t(x_{t+1}|x_t;\theta)$. It only requires the capability to simulate from this density (*simulation-based*). This *simulation-based* methodology has developed fast because of the relatively non-restrictive requirements, but its main representatives follow the Bayesian paradigm, i.e. the Approximate Bayesian Computation (Toni et al., 2009) and the Particle Markov Chain Monte Carlo (Andrieu et al., 2010), SMC^2 (Chopin et al., 2013).

To sum up, iterated filtering is one of few, if not the only method for the maximum likelihood inference in general state-space models (SSMs) that satisfy the three following conditions: it is *likelihood-based* (applies full data-likelihood inference), *simulation-based* (captures dynamics of the model only via the simulation of $f_t(x_{t+1}|x_t;\theta)$), and *frequentist-based* (based on a frequency interpretation of probability).

The necessary conditions for the application of the iteration algorithm to a specific SSM include the ability to:

- simulate from the initial density $f_0(x_0;\theta)$;
- simulate from the transition density $f_t(x_{t+1}|x_t;\theta)$;
- evaluate the measurement density $g_t(y_t|x_t;\theta)$.

In the case of the Yu and Meyer model, all the above conditions are fulfilled. The initial conditions are with probability one, which in practical implementations are treated as not random. However, there is also a possibility to initially draw conditions from stationary distributions:

$$h_{0i}\sim N\left(\mu_i,\frac{\sigma_i^2}{1-\rho_i^2}\right), i = 1,2, \tag{7}$$

$$q_0\sim N\left(\psi_0,\frac{\sigma_\eta^2}{1-\psi_1^2}\right) \tag{8}$$

Simulating from the transition density is straightforward because it requires drawing from a normal distribution (Equations (2) and (3)). Similarly, evaluating the measurement density is immediate, as it is based on the bivariate normal distribution (Equations (5) and (6)).

All the computations presented in the article were made using the *POMP* package (Partially Observed Markov Processes, King et al., 2016) written for the R statistical computing environment (R Development Core Team, 2010). To make the calculations faster, the code of initial, transition and measurement density was written in the C programming language.

4. Empirical example

In this section, we will apply our Yu and Meyer model-estimation method to the estimation of the correlation between the returns of two assets, i.e. Standard and Poor's 500 index (S&P500) and the price of gold in US dollars. The sample period from 22 July 2014 to 3 March 2022 yielded a total of 1,751 observations of logarithmic returns multiplied by 100 (we excluded those days when at least one of two observations was not reported). The data came from Eikon Refinitiv Database. Figure 1 presents time series of assets prices (top row) and returns (bottom row). Table 1 shows summary statistics of returns which demonstrate that the S&P500 returns are more leptokurtic and more left-skewed than the gold returns. As in Yu and Meyer (2006), the returns were mean-corrected before the estimation.

Gold is often considered a safe haven asset, which means that it is expected to retain or increase its value during periods of economic or political uncertainty. This is because for centuries gold has been the carrier of value and was able to retain purchasing power over time. During crises, investors tend to resort to gold as a way to protect their assets from inflation or currency devaluation.

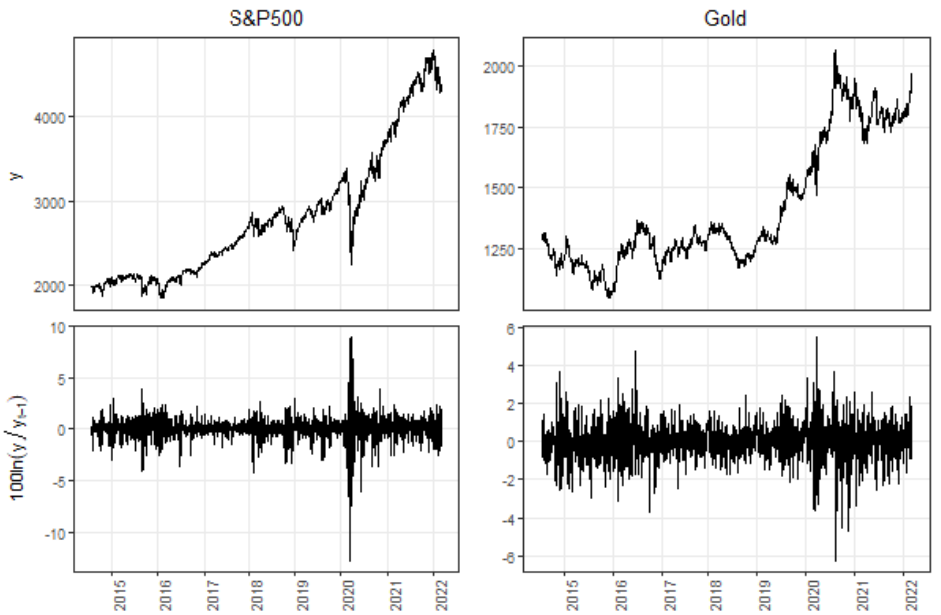
Baur and Lucey (2009) introduced a precise conceptual distinction between a 'safe haven' and a 'hedge'. A safe haven asset is defined as a security that is uncorrelated with stock market returns in the case of market crash, and a hedge as a security that is uncorrelated with the stock market on average. Baur and McDermott (2010) distinguished between strong and weak safe haven effects. A strong safe haven is an asset that is negatively correlated, and a weak safe haven is one that is uncorrelated with another asset or portfolio at a time of the falling stock prices.

Table 1. Summary statistics of logarithmic returns multiplied by 100 calculated for Standard and Poor's 500 index (S&P500) and the price of gold in US dollars in the period from 22 July 2014 to 3 March 2022

Asset	Summary statistics								
	Mean	St. Dev.	Skewness	Kurtosis	Minimum	Q1	Median	Q3	Maximum
S&P500	0.045	1.170	−1.168	22.808	−12.765	−0.342	0.069	0.550	8.968
Gold	0.023	0.928	-0.207	7.157	−6.254	-0.452	0.033	0.500	5.477

Source: author's work.

Figure 1. Time series of Standard and Poor's 500 index (S&P500) and the price of gold (top row) and corresponding logarithmic returns multiplied by 100 (bottom row) in the period from 22 July 2014 to 3 March 2022



Source: author's work.

Table 2 presents the maximum likelihood estimation results for the analysed data. We used the following algorithmic settings: 200 iterations, 1,000 particles, random-walk perturbations with the initial 0.01 perturbations, and a geometric decay of perturbations of $\alpha = 0.5$ (the perturbations at the end of 50 iterations are a fraction α smaller than they are at first) for all the parameters. The estimation of the log-likelihood results from taking the average of nine likelihood evaluations of a bootstrap particle filter with 1,000 particles, from which we also calculate the Monte Carlo standard error. Sandard errors of the parameters were calculated via numerical approximation to the Hessian (see supporting text by Ionides et al. (2006) for details of the procedure). Figures A1 and A2 in the Appendix show diagnostics plots.

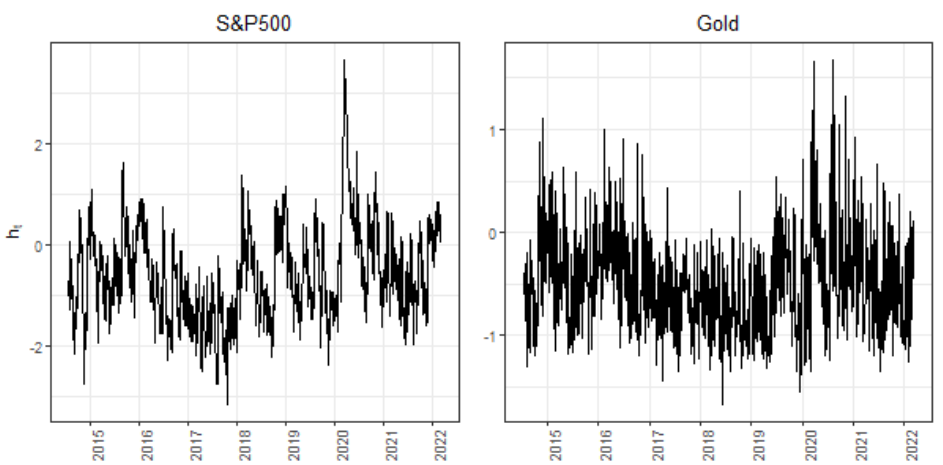
Table 2. Parameter and log-likelihood estimates and standard errors (in parentheses) of the Yu and Meyer model obtained using iterated filtering for the analysed data

Log-likelihood	S&P500			Gold			Conditional correlation		
	μ_1	ϕ_1	σ_1	μ_2	ϕ_2	σ_2	ψ_0	ψ_1	σ_η
−4344.984	−0. 6408	0.9313	0.3939	−0. 5003	0.7668	0.5236	−0.3769	0.9746	0.1425
(1.139)	(0.0767)	(0.0067)	(0.0085)	(0.0512)	(0.0090)	(0.0089)	(0.0768)	(0.0055)	(0.0089)

Note. Algorithm settings: 200 iterations, 1,000 particles, random-walk perturbations with the initial 0.01 perturbations, and a geometric decay of perturbations of α = 0.5 for all the parameters.
Source: author's work.

Comparing the estimation results of log-volatilities for S&P500 and gold, we can see that the stock index has the value of persistency parameter ϕ closer to one and a smaller value of conditional volatility parameter σ. The former parameter controls the persistency of log-volatility (1- ϕ is the strength of a mean-reversion towards the unconditional mean μ after a shock in log-volatility), while the latter regulates its variability. Consequently, the volatility of the S&P500 index is more clustered and less time-varying than the volatility of gold. These differences are visible on the plots of logarithmic returns (bottom row of Figure 1) and log-volatilities (Figure 2).

Figure 2. The mean of the filtering distributions of the log-volatilities.



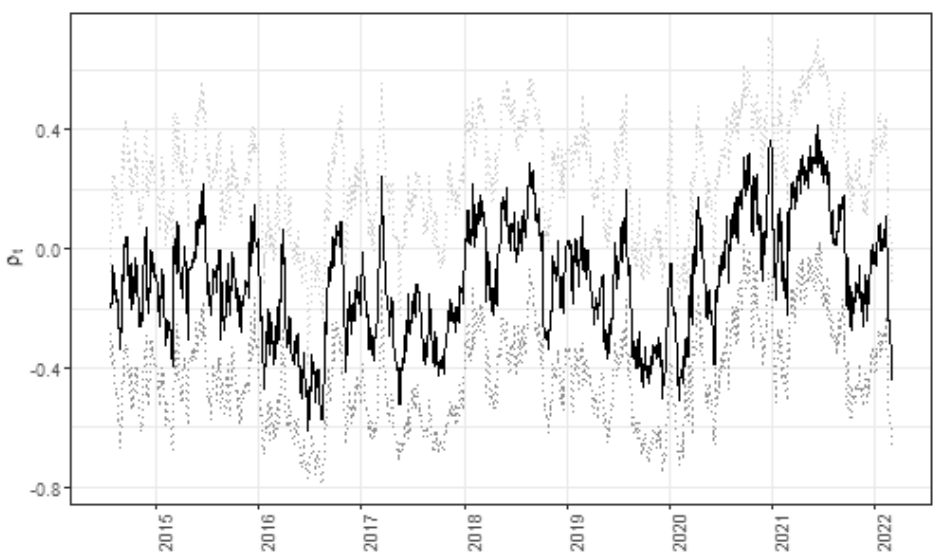
Note. The results were obtained by using a bootstrap particle filter with 1,000 particles.
Source: author's work.

Parameters ψ_0, ψ_1 and σ_η control the Fisher-transformed conditional correlation process q_t. Similarly to log-volatilities, this is also the mean-reverted autoregressive process of order 1, and consequently, the parameters have similar interpretations: ψ_0

is the long-term mean, and ψ_1 and σ_η regulate the persistency and the variability of the transformed conditional correlations, respectively. On the basis of the estimated value of ψ_0, after the inverse Fisher transformation, we obtain the long-term value of correlation –0.186. This value can be interpreted as a slightly or moderately negative long-term correlation. The strength of the mean-reversion is very high – higher than in the case of both log-volatilises.

Figure 3 presents a plot of conditional correlation process ρ_t. For most of the sample period, the conditional correlation had a negative value, but the trajectory was volatile with several picks above zero. For the period of the COVID-19 pandemic (2020–2022), the relationship between the S&P500 index and gold was mainly positive. Generally, in the sample period, gold acted as a hedge for the S&P500 index, but failed to be a safe haven asset during the COVID-19 crisis. Będowska-Sójka and Kliber (2021) arrived at a similar conclusion, as they analysed, among other things, the safe haven characteristics of gold versus the S&P500 index: 'gold tends to take the role of a safe haven asset in relatively short periods, yet the recent COVID crisis does not belong to them'.

Figure 3. The mean (solid black line), the 0.05 quantile and the 0.95 quantile (dotted lines) of the filtering distributions of the dynamic conditional correlation



Note. The results were obtained by using a bootstrap particle filter with 1,000 particles.
Source: author's work.

5. Simulation study

As the convergence of the iterated filtering algorithm to the maximum likelihood estimates is difficult to demonstrate analytically, we assessed its performance through a simple simulation study. In order to evaluate the precision of the parameter estimation, we conducted 100 time-series simulations of the Yu and Meyer model, using identical parameter values as those derived from the empirical study detailed in Section 4 (see Table 2). Each simulation has the same length as the analysed time series of the S&P500 index and gold (1,751 observations). We followed the same estimation procedure as in the empirical study for each simulation. The results of the study are presented in Table 3, which consists of the mean errors (ME) and the root of mean square errors (RMSE) of parameter estimates. The obtained results show that there is no such type of a parameter (long-term mean, persistency, volatility) that would be biased in one direction. The obtained RMSE are comparable with standard errors from the empirical study. The results indicate that the proposed method is reliable to a sufficient degree.

Table 3. Results of the simulation study based on a 100-time-series simulation of the Yu and Meyer model with the same parameter values as those obtained in the empirical study and the same length of time series as the analysed time series of the S&P500 index and gold

Measure of errors	S&P500			Gold			Conditional correlation		
	μ_1	ϕ_1	σ_1	μ_2	ϕ_2	σ_2	ψ_0	ψ_1	σ_η
ME	0.0294	0.0154	–0.0446	–0.0326	–0.0403	0.0851	0.0025	0.0182	–0.0175
RMSE	0.0399	0.0009	0.0049	0.0104	0.0039	0.0101	0.1006	0.0057	0.0134

Note. Algorithm settings: 200 iterations, 1,000 particles, the random-walk perturbations with initial 0.01 perturbations, and geometric decay of perturbations of $\alpha = 0.5$ for all parameters.

Source: author's work.

6. Conclusions

The main contribution of this paper to the is the existing body of research is the proposed iterated filtering algorithm for estimating the bivariate SV model of Yu and Meyer. It allows the estimation of parameters in the frequentist-based approach. Consequently, the filtering distribution of log-volatilities and the dynamic correlation might be estimated in the frequentist-based approach using a standard bootstrap particle filter. In order to show the effectiveness of the proposed estimation method, an empirical example was presented, in which the Yu and Meyer model was used to estimate the dynamic correlation between the returns of two assets, i.e. Standard and Poor's 500 index (S&P500) and the price of gold in US

dollars. The results indicated that gold acted as a hedge within the observed period, but during the COVID-19 pandemic, it failed to perform as a safe haven. The obtained parameter estimates from the empirical study were confirmed in the simulation experiment. One of the limitations of our study was its restriction to a bivariate model of stochastic volatility. Further research in this area should determine whether the proposed approach proves effective in multivariate models of stochastic volatility based on the Cholesky decomposition or a matrix exponential.

References

Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *72*(3), 269–342. https://doi.org/10.1111/j.1467-9868.2009.00736.x.

Asai, M., McAleer, M., & Yu, J. (2006). Multivariate Stochastic Volatility: A Review. *Econometric Reviews*, *25*(2–3), 145–175. https://doi.org/10.1080/07474930600713564.

Baur, D. G., & Lucey, B. M. (2009). Flights and contagion – An empirical analysis of stock-bond correlations. *Journal of Financial Stability*, *5*(4), 339–352. https://doi.org/10.1016/j.jfs.2008.08.001.

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

Bejger, S., & Fiszeder, P. (2021). Forecasting currency covariances using machine learning tree-based algorithms with low and high prices. *Przegląd Statystyczny. Statistical Review*, *68*(3), 1–15. https://doi.org/10.5604/01.3001.0015.5582.

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*. https://doi.org/10.1016/j.najef.2021.101390.

Bhadra, A., Ionides, E. L., Laneri, K., Pascual, M., Bouma, M., & Dhiman, R. C. (2011). Malaria in Northwest India: Data Analysis via Partially Observed Stochastic Differential Equation Models Driven by Lévy Noise. *Journal of the American Statistical Association*, *106*(494), 440–451. https://doi.org/10.1198/jasa.2011.ap10323.

Bollerslev, T., Patton, A. J., & Quaedvlieg, R. (2018). Modeling and forecasting (un)reliable realized covariances for more reliable financial decisions. *Journal of Econometrics*, *207*(1), 71–91. https://doi.org/10.1016/j.jeconom.2018.05.004.

Bretó, C. (2014). On idiosyncratic stochasticity of financial leverage effects. *Statistics & Probability Letters*, *91*, 20–26. https://doi.org/10.1016/j.spl.2014.04.003.

Cappé, O., Moulines, E., & Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer Science & Business Media. https://doi.org/10.1007/0-387-28982-8.

Chib, S., Omori, Y., & Asai, M. (2009). Multivariate Stochastic Volatility. In T. G. Andersen, R. A. Davis, J.-P. Kreiß & T. Mikosch (Eds.), *Handbook of Financial Time Series* (pp. 365–400). Springer. https://doi.org/10.1007/978-3-540-71297-8_16.

Chiu, T. Y. M., Leonard, T., & Tsui, K.-W. (1996). The Matrix-Logarithmic Covariance Model. *Journal of the American Statistical Association*, *91*(433), 198–210. https://doi.org/10.1080/01621459.1996.10476677.

Chopin, N., Jacob, P. E., & Papaspiliopoulos, O. (2013). SMC^2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *75*(3), 397–426. https://doi.org/10.1111/j.1467-9868.2012.01046.x.

Chopin, N., & Papaspiliopoulos, O. (2020). *An Introduction to Sequential Monte Carlo*. Springer Nature. https://doi.org/10.1007/978-3-030-47845-2.

Du, X., Yu, C. L., & Hayes, D. J. (2011). Speculation and volatility spillover in the crude oil and agricultural commodity markets: A Bayesian analysis. *Energy Economics*, *33*(3), 497–503. https://doi.org/10.1016/j.eneco.2010.12.015.

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

Engle, R. F., & Kroner, K. F. (1995). Multivariate Simultaneous Generalized ARCH. *Econometric Theory*, *11*(1), 122–150. https://doi.org/10.1017/S0266466600009063.

Fiszeder, P., & Orzeszko, W. (2021). Covariance matrix forecasting using support vector regression. *Applied Intelligence*, *51*(10), 7029–7042. https://doi.org/10.1007/s10489-021-02217-5.

Gębka, B., & Karoglou, M. (2013). Have the GIPSI settled down? Breaks and multivariate stochastic volatility models for, and not against, the European financial integration. *Journal of Banking and Finance*, *37*(9), 3639–3653. https://doi.org/10.1016/j.jbankfin.2013.04.035.

Gordon, N. J., Salmond, D .J., & Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F. Radar and Signal Processing*, *140*(2), 107–113. https://doi.org/10.1049/ip-f-2.1993.0015.

Gouriéroux, C. (2006). Continuous Time Wishart Process for Stochastic Risk. *Econometric Reviews*, *25*(2–3), 177–217. https://doi.org/10.1080/07474930600713234.

Harvey, A., Ruiz, E., & Shephard, N. (1994). Multivariate Stochastic Variance Models. *The Review of Economic Studies*, *61*(2), 247–264. https://doi.org/10.2307/2297980.

He, D., Ionides, E. L., & King, A. A. (2009). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface*, *7*(43), 271–283. https://doi.org/10.1098/rsif.2009.0151.

Hui, E. C. M., & Zheng, X. (2012a). Exploring the dynamic relationship between housing and retail property markets: An empirical study of Hong Kong. *Journal of Property Research*, *29*(2), 85–102. https://doi.org/10.1080/09599916.2012.674968.

Hui, E. C., & Zheng, X. (2012b). The dynamic correlation and volatility of real estate price and rental: An application of MSV model. *Applied Economics*, *44*(23), 2985–2995. https://doi.org/10.1080/00036846.2011.568409.

Ionides, E. L., Bhadra, A., Atchadé, Y., & King, A. (2011). Iterated filtering. *The Annals of Statistics*, *39*(3), 1776–1802. https://doi.org/10.1214/11-AOS886.

Ionides, E. L., Bretó, C., & King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(49), 18438–18443. https://doi.org/10.1073/pnas.0603181103.

Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., & King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences*, *112*(3), 719–724. https://doi.org/10.1073/pnas.1410597112.

Ishihara, T., & Omori, Y. (2012). Efficient Bayesian estimation of a multivariate stochastic volatility model with cross leverage and heavy-tailed errors. *Computational Statistics and Data Analysis*, *56*(11), 3674–3689. https://doi.org/10.1016/j.csda.2010.07.015.

Ishihara, T., Omori, Y., & Asai, M. (2016). Matrix exponential stochastic volatility with cross leverage. *Computational Statistics and Data Analysis*, *100*, 331–350. https://doi.org/10.1016/j.csda.2014.10.012.

Jin, X., & Maheu, J. M. (2013). Modeling Realized Covariances and Returns. *Journal of Financial Econometrics*, *11*(2), 335–369. https://doi.org/10.1093/jjfinec/nbs022.

Johansson, A. C. (2010a). Asian sovereign debt and country risk. *Pacific-Basin Finance Journal*, *18*(4), 335–350. https://doi.org/10.1016/j.pacfin.2010.02.002.

Johansson, A. C. (2010b). *Stock and Bond Relationships in Asia* (CERC Working Paper Series No. 14). https://ideas.repec.org//p/hhs/hacerc/2010-014.html.

Jungbacker, B., & Koopman, S. J. (2006). Monte Carlo Likelihood Estimation for Three Multivariate Stochastic Volatility Models. *Econometric Reviews*, *25*(2–3), 385–408. https://doi.org/10.1080/07474930600712848.

Kastner, G., Frühwirth-Schnatter, S., & Lopes, H. F. (2017). Efficient Bayesian Inference for Multivariate Factor Stochastic Volatility Models. *Journal of Computational and Graphical Statistics*, *26*(4), 905–917. https://doi.org/10.1080/10618600.2017.1322091.

King, A. A., Ionides, E. L., Pascual, M., & Bouma, M. J. (2008). Inapparent infections and cholera dynamics. *Nature*, *454*(7206), 877–880. https://doi.org/10.1038/nature07084.

King, A. A., Nguyen, D., & Ionides, E. L. (2016). Statistical Inference for Partially Observed Markov Processes via the R Package pomp. *Journal of Statistical Software*, *69*(12), 1–43. https://doi.org/10.18637/jss.v069.i12.

Kliber, A. (2011). Sovereign CDS Instruments in Central Europe – Linkages and Interdependence. *Dynamic Econometric Models*, *11*, 111–128. https://doi.org/10.12775/DEM.2011.008.

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. https://doi.org/10.1016/j.physa.2019.04.145.

Lele, S. R., Dennis, B., & Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, *10*(7), 551–563. https://doi.org/10.1111/j.1461-0248.2007.01047.x.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067. https://doi.org/10.1002/sim.3680.

Malik, S., & Pitt, M. K. (2011). Particle filters for continuous likelihood evaluation and maximisation. *Journal of Econometrics*, *165*(2), 190–209. https://doi.org/10.1016/j.jeconom.2011.07.006.

Nguyen, D. (2016). Another look at Bayes map iterated filtering. *Statistics and Probability Letters*, *118*, 32–36. https://doi.org/10.1016/j.spl.2016.05.013.

R Development Core Team. (2010). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. http://www.r-project.org.

Shephard, N., & Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, *84*(3), 653–667. https://doi.org/10.1093/biomet/84.3.653.

So, M. K. P., Li, W. K., & Lam, K. (1997). Multivariate modelling of the autoregressive random variance process. *Journal of Time Series Analysis*, *18*(4), 429–446. https://doi.org/10.1111/1467-9892.00060.

Stocks, T., Britton, T., & Höhle, M. (2020). Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in Germany. *Biostatistics*, *21*(3),400–416. https://doi.org/10.1093/biostatistics/kxy057.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, *12*(3), 1–16. https://doi.org/10.18637/jss.v012.i03.

Szczepocki, P. (2020). Application of iterated filtering to stochastic volatility models based on non-Gaussian Ornstein-Uhlenbeck process. *Statistics in Transition new series*, *21*(2), 173–187. https://doi.org/10.21307/stattrans-2020-019.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, *6*(31), 187–202. https://doi.org/10.1098/rsif.2008.0172.

Tsay, R. S. (2005). *Analysis of Financial Time Series*. John Wiley & Sons.

Watanabe, T., & Omori, Y. (2004). A multi-move sampler for estimating non-Gaussian time series models: Comments on Shephard & Pitt (1997). *Biometrika*, *91*(1), 246–248. https://doi.org/10.1093/biomet/91.1.246.

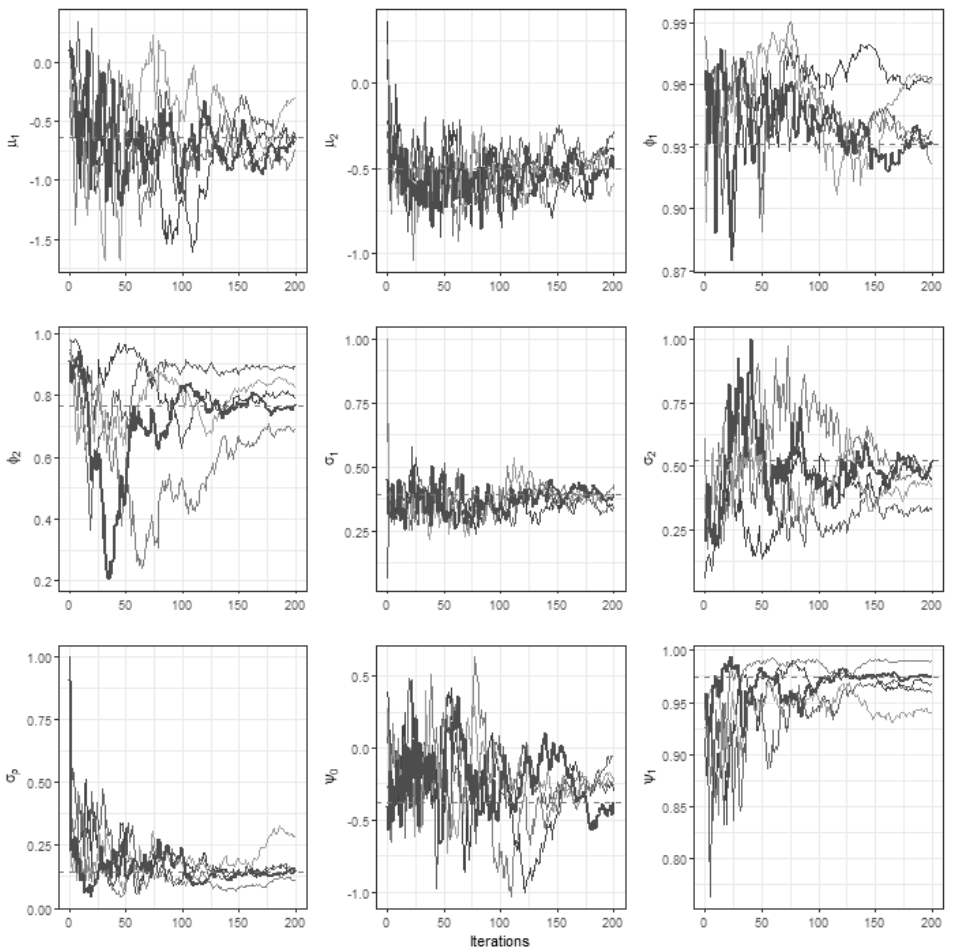
Xu, Y., & Jasra, A. (2019). Particle filters for inference of high-dimensional multivariate stochastic volatility models with cross-leverage effects. *Foundations of Data Science*, *1*(1), 61–85. https://doi.org/10.3934/fods.2019003.

You, C., Deng, Y., Hu, W., Sun, J., Lin, Q., Zhou, F., Pang, C. H., Zhang, Y., Chen, Z., & Zhou, X.-H. (2020). Estimation of the time-varying reproduction number of COVID-19 outbreak in China. *International Journal of Hygiene and Environmental Health*, *228*, 1–7. https://doi.org/10.1016/j.ijheh.2020.113555.

Yu, J., & Meyer, R. (2006). Multivariate Stochastic Volatility Models: Bayesian Estimation and Model Comparison. *Econometric Reviews*, *25*(2–3), 361–384. https://doi.org/10.1080/07474930600713465.

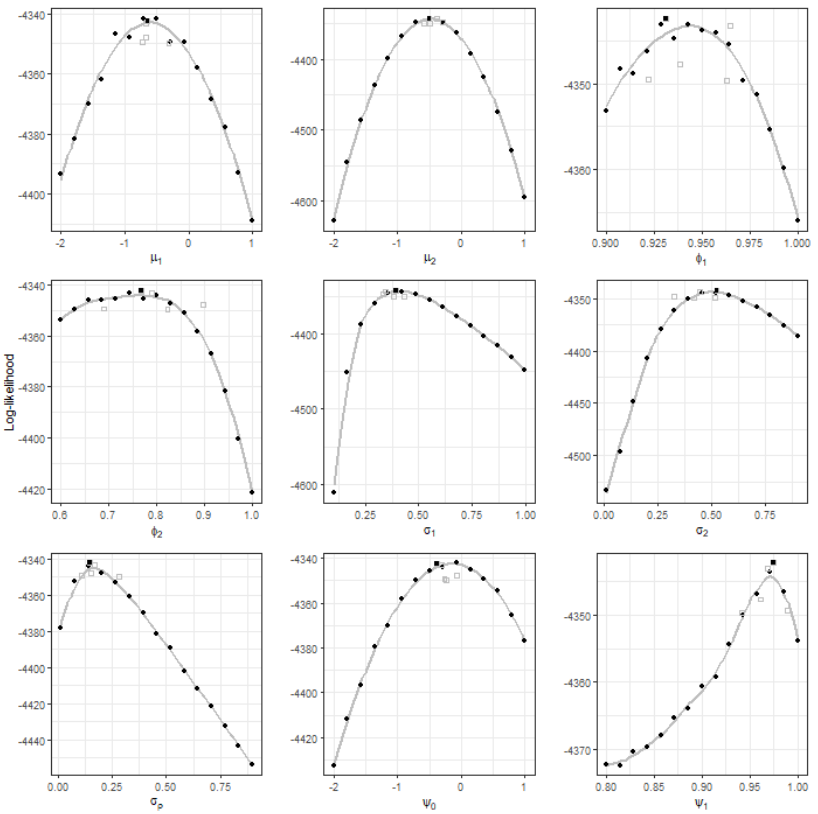
Appendix

Figure A1. Five trajectories of the algorithm with random start points (updates of estimation values during iteration steps)



Note. The bolded trajectory converges to the highest value of log-likelihood.
Source: author's work.

Figure A2. Sliced log-likelihoods for the analysed parameters



Note. The log-likelihood surface is explored along one of the parameters, keeping the others fixed at the point to which iterated filtering converges (see Table 2 for values). Each black circle shows the log-likelihood estimation obtained with 2,000 particles. The grey curve results from smoothing the log-likelihood evaluations with local quadratic regression (R base function *loess*). The empty squares correspond to the ends of the five trajectories of the iterated filtering algorithm (see Figure A1 for the trajectory plots). The black squares correspond to the estimated values.

Source: author's work.

Przegląd Statystyczny. Statistical Review, 2022, vol. 69, 4, 20–28 https://doi.org/10.59139/ps.2022.04.2
Received: 4.01.2023, revised: 31.03.2023, accepted: 15.04.2023

Sample size in clinical trials – challenges and approaches

Andrzej Tomski,[a] Barbara Gorzawska[b]

Abstract. Sample size estimation is a necessary and crucial step in clinical trial research. Statistical requirements, limited patient availability and high financial risk of a clinical trial necessitate the proper calculation of this measure. The aim of this paper is to discuss the reasons why the estimation of the sample size is important and, based on the obtained results, to show how this process may be completed in selected cases. Stochastic simulations based on the Monte Carlo methods approach are applied. Therefore, new challenges facing this area of research are mentioned.

Keywords: sample size, clinical trial, Monte Carlo methods, stochastic simulations
JEL: C13, C15, C18

1. Introduction

Estimating the sample size is an important issue, relating in the recent years particularly to statistics for clinical trials. Clinical trials are prospective biomedical research studies on humans designed to answer crucial questions about biomedical interventions, including new treatments (National Institutes of Health, n.d.). While scientists in some disciplines have easy access to all of the data representing their research topic, it is not the case in the field of clinical trials. In biology, scientists need to obtain a sufficient portion of research material, while in medicine, they have to select patients who are suitable for therapy and only then begin the research on the treatment. Clinical trials are even more problematic, as they are expensive, with a high level of formal requirements and involve a complex patient recruitment process. Sample size estimation is a problematic procedure, as it is costly, time-consuming, the number of subjects fit for the process tends to be limited, while the statistical requirements are very specific. On the other hand, estimating the sample size is mandatory in these studies before any patients are even recruited.

At the same time, with the increasing number of instances requiring the estimation of a sample size before a given study is initiated, no data relating to it is available. Therefore, a problem arises, for example, when a researcher wants to apply

[a] University of Silesia in Katowice, Institute of Mathematics, ul. Bankowa 12, 40-007 Katowice, e-mail: andrzej.tomski@us.edu.pl, ORCID: https://orcid.org/0000-0001-6944-0600.
[b] Parexel FSP, Parexel International, Warsaw, Poland, e-mail: gorzawska.barbara@gmail.com, ORCID: https://orcid.org/0000-0002-4263-8489.

for a grant and needs to provide the precise costs and details of a study, in particular the number of samples needed. Another issue is the fact that researchers tend not to consider such concepts as the test power, errors or the effect size. Researchers pay great attention mainly to the concept of the p-value, while ignoring other statistical indicators (Amrhein et al., 2019). Consequently, the results of many studies are concluded mainly on the basis of low p-value results considered as significant. On the other hand, the non-random components (Szreder, 2022) of the total survey error do not have to decrease as the sample size increases. Non-random errors such as coverage errors or measurement errors generate a bias which does not depend on the sample size (Chin, 2012).

Relying only on the p-value has been widely criticised in the recent years (Platek and Särndal, 2001) and researchers have trouble distinguishing between statistical significance and practical significance. It should be noted that statistical significance does not mean significance in general. For example, the term 'clinical relevance' refers to the practical significance of a treatment effect. Researchers focusing solely on statistical significance or the lack of it may report results that are not significant in practice. This can occur especially when a large sample size is considered. A large sample size is crucial when determining statistical significance and in such a case confusion relates to its interpretation. However, an article published by the American Statistical Association (ASA) presents a formal statement explaining several commonly agreed upon principles underlying the correct use and interpretation of p-values (Wasserstein & Lazar, 2016).

In this paper, we will briefly describe the requirements and standards of clinical trials to show the importance of sample size estimation. Researchers of other disciplines interested in this issue may also consider adopting a similar approach in their field of study.

2. Clinical trial guidelines

Medicine and pharmacy are disciplines most interested in sample size estimation, although the group of people involved in those fields also greatly benefit from sample size estimation as it saves them time, money and effort. This is definitely important, especially when gathering data requires significantly more work from the researcher than just downloading them from the internet.

Recent years have shown that clinical trials are a very dynamically developing field of study, thus the demand for its corresponding statistical methods is constantly growing. In the initial phase of a study, a clinical study protocol[1] and a statistical

[1] A protocol is a document that describes how a clinical trial will be conducted and ensures both the safety of the subjects and the integrity of the collected data.

analysis plan[2] for the research are developed. Both require the sample size of the patients recruited for the study. The principles for conducting such surveys do not clearly indicate how the sample size should be estimated. However, there are some requirements in terms of the statistical and medical demand. Guidelines for statisticians are provided by the European Medicines Agency (EMA, n.d.) and the U.S. Food and Drug Administration (n.d.). Statistical analysis, therefore, does not start with the receipt of data and the selection of specific tests. In practice, the researcher first formulates the objectives of the study and determines the acceptable level of significance α. According to the EMA guidelines, it is set at the level of 5%. The Sponsor[3] expects a test to be constructed so that its power is as high as possible for the fixed α level. This is where the interest of the Sponsor (who wants the most powerful test) and the patients (who just want effective treatment) may clash. As a result, the statistician recommends to both the Sponsor and the Investigator[4] that the minimum sample size for the test power exceeds the acceptable level to at least 80% in accordance with the EMA guidelines (EMA, n.d). Thus, we assume that

$$\alpha = 5\% \; and \; 1 - \beta \geq 80\%, \tag{1}$$

where β is a type II error rate.

The next step involves gathering a sufficiently large sample size by the Sponsor and Investigator by inviting the required number of subjects to participate in the study until its completion. After that, the main statistical analysis can start. In practice, interim analyses[5] (Hayes and Patterson, 1921) are performed in such cases, i.e. research is conducted after only a part of the recommended sample is collected. Performing interim analyses ensures the safety of the study so that the risk of complications, in particular serious adverse events[6] is minimised. Additionally, if the study shows that the administered treatment is ineffective, it may be discontinued, thus saving any further effort. While the study can be stopped at any moment due to patient safety concerns or unsatisfactory results indicated by the interim analyses, their status cannot be in any way considered as confirmatory. Therefore, in order to pronounce the treatment effective, the study has to be continued. This means that it is not possible to perform a statistical analysis by collecting a smaller sample to

[2] A statistical analysis plan (SAP) outlines the analytical approach of the data collected in a clinical trial.
[3] Legal person who funds the research.
[4] It is a person who is involved in running a clinical trial.
[5] It is an analysis of data that is conducted before data collection is completed.
[6] An adverse event is any undesirable experience associated with the use of a medical product on a patient. The event is considered serious when its outcome is life-threatening or it leads to the patient's death, hospitalisation or permanent health impairment.

confirm the investigated hypotheses, namely the effectiveness of the drug. The issue of sample size estimation must also take into account the results of the recently introduced non-inferiority tests, which assume a certain margin of error compared to reference objects, i.e. existing drugs.

3. Sample size calculation – selected computational tools

Sample size estimation is a topic that is growing in popularity along with the big data sector, the increase of the number of clinical trials (Delgado et al., 2018) and technological advances. Numerous reference books (Chow et al., 2007) and software tools for estimating sample size in the simplest cases are widely available. These tools are in most cases very user-friendly and publicly provided, nevertheless, they do have certain disadvantages.

Many sample size calculators are available on the Internet, although here we will provide examples of two of them to describe the nature of their work.

The G*Power (Faul et al., 2009) is a tool used to perform statistical power analyses for many different tests, including t-tests, F tests, χ^2 tests, z-tests and some exact tests. The G*Power can also be used to compute effect sizes and to depict the results of power analyses. In order to calculate the sample size, it only requires the user to select a test from a list, provide a measure of the effect size and enter the test power along with its significance level. However, the program has some disadvantages: the test selection is limited to a list, there is no possibility to specify the exact form of the statistical model or link the results to confidence intervals. Moreover, the lack of access to its source code raises questions as to how these estimates were obtained.

Similarly, the Sample Size Calculator (Raosoft, 2004) does not refer to the type of the investigated variable. It does not even offer a choice of the test. This tool requires from the user to enter the α level, the upper limit of the sample size and a confidence level without any explanation. Its overall applicability seems to be quite limited and thus it may not be able to estimate sample size in certain models accurately enough. What is more, the graphical interface of the application shows a lot of additional windows, which may surprise and confuse the user.

Online sample size estimation tools tend to offer a narrow range of possibilities. It is sometimes difficult to clearly identify which models they refer to or they are intended for a very simple and one specific statistical model. Another serious disadvantage of these models is that, in general, they do not refer to scientific results obtained in papers presenting this type of research. In conclusion, a more universal method needs to be devised offering an efficient approach for a wide variety of statistical models, but which would also refer to the results provided in the literature.

As in the case of the numerous online tools, there are many books and research articles extensively discussing the issue of sample size estimation.

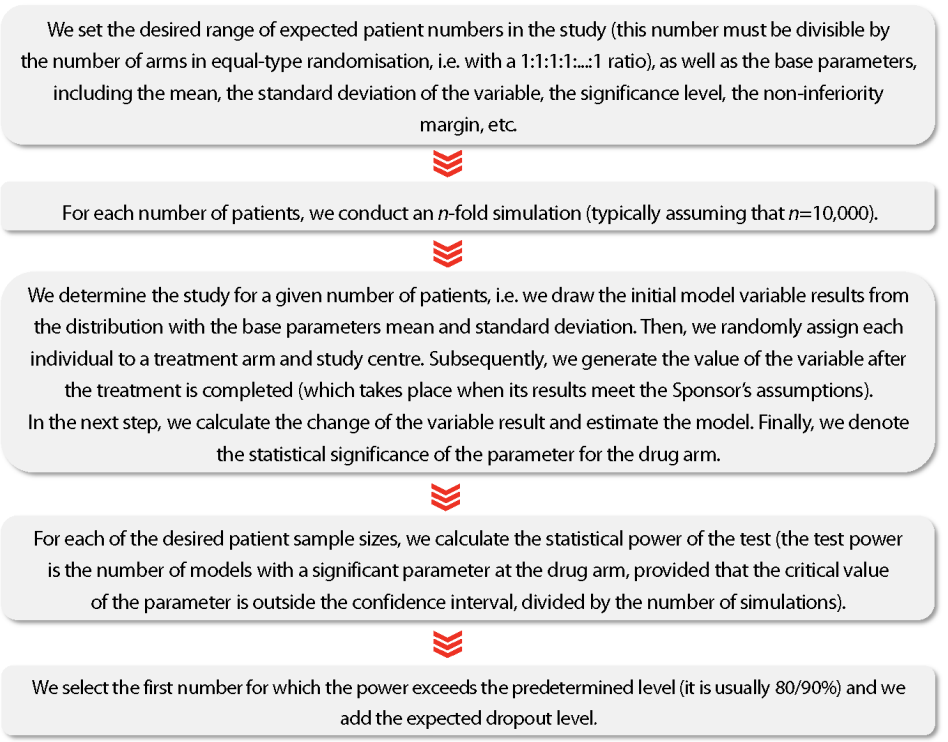
3.1. Monte Carlo methods – a brief review

Monte Carlo methods constitute a large class of computational algorithms that are based on repeated random sampling aiming to obtain some numerical results. In general, the application of a Monte Carlo method involves the limitation of the power of the test in order to obtain the sample size, i.e. it allows the evaluation of the parameter bias and the power of the test based on computer-generated population data. The most popular statistical methods include using parameter values determined in a study from the past, meta-analysis techniques or self-estimated statistical parameters. However, if the research concerns newly discovered diseases, there is basically only a third way: using self-estimated statistical parameters. When the population data is obtained, samples of a given size from a certain range are generated. This enables the calculation of model parameters and the power of the test is able to satisfy the expected demands. It must be noted that the parameters' bias is strongly associated with the deviation of the experimentally estimated value from the set of the replicated estimates. The details of the approach estimating the sample size with many successful experimental applications is presented in the next section.

3.2. Monte Carlo-based approach

This part of the paper focuses on an approach to sample size estimation that combines the concepts of p-values with the test power, the effect size and the use of confidence intervals. As a result, the estimate is not based on the debatable p-value alone, but also on other relevant significance criteria stated in the study. Monte Carlo methods have been present in statistics for quite some time (see, for example, papers like Jiang et al., 2012). However, this paper attempts to systematise this approach in the form of a fairly simple scheme with reference to several important aspects. It must be emphasised here that the sample size is calculated for a specific research hypothesis and for a specific statistical test which was selected on the basis of this hypothesis. The adopted approach is based on the scheme illustrated in the following Figure.

Figure. A structured approach to sample size estimation in a typical clinical trial



Source: authors' work.

The presented approach strengthens the role of not only the effect size, but also the confidence intervals, because the result is not considered significant if the confidence interval includes a critical value suggesting that there is no difference for the estimated parameters.

This part of the paper provides an example illustrating the discussed approach. The Visual Analogue Scale (VAS) is a pain rating scale used for the first time in 1921 by Hayes and Patterson (1921). We consider a study with a primary endpoint[7] measured by a decrease in VAS between the pre-treatment and the 30-day treatment. The study specifies that randomisation[8] assigns the patient to one of the four arms:[9] placebo, low dose, medium dose and high dose. Three medical centres[10] participated in the study. The relevant literature indicates that patients with the examined disease evaluate their pain on the VAS scale, e.g. at an average of 7.5 with a standard

[7] Main hypothesis in a study.
[8] Patients are randomly assigned to the control group and to the treatment group.
[9] Arm in a clinical trial refers to each group or subgroup of participants that receives specific interventions (or placebo) according to the protocol.
[10] A place where an experiment is conducted.

deviation of 1.0. The Sponsor expects the patients receiving placebo to have an average VAS score of 5.5 with a standard deviation of 1.5 after 30 days, while the patients receiving the study drug an average VAS of 4 with a standard deviation of 1.0. We carry out 10,000 study simulations in order to estimate certain parameters in the study, i.e. from the distribution with base parameters m (mean) and s (standard deviation) and for the appropriate number of patients, we draw their initial VAS result, then we randomly assign them to an arm and a centre; in the next step we generate a VAS result after 30 days of treatment based on the Sponsor's assumptions; we then calculate the change on the VAS scale and estimate the ANCOVA model, which takes the following form:

$$\Delta VAS_{ijk} = \beta_0 + \tau_i + \beta_1 x_{ijk} + \varphi_k + \epsilon_{ijk}, \tag{2}$$

where:

ΔVAS_{ijk} is the difference between the outcome and the baseline of the VAS grade for the j-th patient under the i-th treatment in the k-th centre
x_{ijk} is the baseline VAS grade for the j-th patient under the i-th treatment in the k-th centre,
τ_i are the fixed treatment effects,
φ_k are the fixed centre effects,
β_0 and β_1 are regression coefficients,
ϵ_{ijk} are independent random variables with the $N(0, \sigma)$ distribution.

In the final step, we record the significance of the parameter at the treatment arm. The percentage of statistically significant results (provided that the confidence interval for this parameter does not contain zero) constitutes the estimate of the power of the test. We select the smallest number n for which the power exceeds 80%. The full implementation of our single Monte Carlo experiment in the R programming language for a sample dataset is provided in the supplementary material (available upon request submitted to the authors via e-mail).

4. Limitations of the study

This section outlines certain limitations of the approach presented above. Besides the previously mentioned lack of literature on the new issues, the problem is the lack of extensive information in the literature on the proper distribution of data. In many cases, only the basic central tendency measures are calculated. In order to perform a simulation, however, we need to specify the distribution precisely. In the proposed

solution, we use a normal distribution, but the question is how precisely this distribution approximates the values of the studied parameter. Therefore, some assumptions have to be made at this stage as well. Sometimes we can use a chart from an article, although it still tends to provide quite scarce information. Whether a given value has any constraints (i.e. discrete variables, possible minimums and maximums) should also be taken into account. These problems may escalate when more advanced statistical methods (e.g. mixed models, survival analysis) or a more complicated study design (e.g. crossover arms as described in Yeh et al., 2020) are used in the clinical trial. It also raises questions as to the validation method and the strictly numerical accuracy of the model. In these cases the number of uncertainties resulting from the lack of information can increase in the future.

5. Conclusion

This paper discussed the issue of sample size estimation in a clinical trial. Various approaches used to calculate this value were described. We proposed an approach in the form of a diagram based on the Monte Carlo methods. Our aim was to draw researchers' attention not only to the problem itself, but also to the role of effect sizes and confidence intervals in such estimates. This is a serious challenge for the further development of statistics, involving almost all the key concepts of modern statistics. The inclusion of non-inferiority tests or meta-analyses as new directions of change give hope that the raised aspect will be further discussed and developed.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have occurred and influenced the work reported in this paper.

References

Amrhein, V., Greenland, S., & McShane, B. (2019, March 20). *Scientists rise up against statistical significance*. https://www.nature.com/articles/d41586-019-00857-9.

Chin, R. (2012). *Adaptive and Flexible Clinical Trials*. CRC Press.

Chow, S.-C., Wang, H., & Shao, J. (2007). *Sample Size Calculations in Clinical Research* (2nd edition). CRC Press. https://doi.org/10.1201/9781584889830.

Delgado, D. A., Lambert, B. S., Boutris, N., McCulloch, P. C., Robbins, A. B., Moreno, M. R., & Harris, J. D. (2018). Validation of Digital Visual Analog Scale Pain Scoring With a Traditional Paper-based Visual Analog Scale in Adults. *Journal of the American Academy of Orthopaedic Surgeons. Global Research & Reviews*, *2*(3), 1–6. https://doi.org/10.5435/JAAOSGlobal-D-17-00088.

European Medicines Agency. (n.d.). *ICH E9 statistical principles for clinical trials – Scientific guideline*. Retrieved October 14, 2019, from https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149.

Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin*, *18*(2), 98–99.

Jiang, Z., Wang, L., Li, C., Xia, J., & Jia, H. (2012). A Practical Simulation Method to Calculate Sample Size of Group Sequential Trials for Time-to-Event Data under Exponential and Weibull Distribution. *PLOS ONE*, *7*(9), 1–10. https://doi.org/10.1371/journal.pone.0044013.

National Institutes of Health. (n.d.). *Clinical Research Trials and You: The Basics*. Retrieved October 3, 2022, from https://www.nih.gov/health-information/nih-clinical-research-trials-you/basics.

Platek, R., & Särndal, C. E. (2001). Czy statystyk może dostarczyć dane wysokiej jakości?. *Wiadomości Statystyczne*, *46*(4), 1–21.

Raosoft. (2004). *Raosoft Sample Size Calculator*. http://www.raosoft.com/samplesize.html.

Szreder, M. (2022). Szanse i iluzje dotyczące korzystania z dużych prób we wnioskowaniu statystycznym. *Wiadomości Statystyczne. The Polish Statistician*, *67*(8), 1–16. https://doi.org/10.5604/01.3001.0015.9704.

U.S. Food and Drug Administration. (n.d.). *Guidance for Industry. E9 Statistical Principles for Clinical Trials*. Retrieved September 11, 2019, from https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9-statistical-principles-clinical-trials.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on *p*-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. http://dx.doi.org/10.1080/00031305.2016.1154108.

Yeh, J., Gupta, S., Patel, S. J., Kota, V., & Guddati, A. K. (2020). Trends in the crossover of patients in phase III oncology clinical trials in the USA. *Ecancermedicalscience*, *14*, 1–8. https://doi.org/10.3332/ecancer.2020.1142.

Przegląd Statystyczny. Statistical Review, 2022, vol. 69, 4, 29–40 https://doi.org/10.59139/ps.2022.04.3
Received: 27.02.2023, revised: 18.04.2023, accepted: 25.04.2023

Impact of a priori positive information on the results of voting methods

Honorata Sosnowska,[a] Michał Ramsza,[b] Paweł Zawiślak[c]

Abstract. The aim of this paper is to present the results of experiments relating to voting methods based on the bounded rationality theory. The research demonstrated that a positive nudge changes the voting results. The study focused on three methods of voting: the Borda Count method, the Condorcet winner method and the anti-manipulation method. In a laboratory experiment, the subjects were asked to select the best musician. They were to manipulate their voting so that a predetermined winner is chosen. In the first voting, the subjects did not receive any a priori information, while in the second voting, some a priori information was provided, i.e. the true, objective ranking of the musicians. What followed was another voting. It was initially assumed that the participants would manipulate their voting the same way as in the first voting. The results, however, were different. The obtained second ranking of musicians was closest to the true, objective ranking, thus proving that the manipulation effect was neutralised by the a priori positive information about the true, objective order.

Keywords: a priori information, strategic voting, voting methods

JEL: D71, D83, D91

1. Introduction

The idea of bounded rationality was introduced by Herbert Simon in 1955. It concerns limited rationality of individuals when making decisions. These limitations may be caused by the cognitive capacity of the mind. In a perfectly defined world, the agents would be perfect, i.e. economically rational; however, in the real world, the understanding of dilemmas is affected by subjective concepts, not necessarily consistent with the expected value maximisation principle. The most popular examples of situations where bounded rationality occurs is framing and nudge, both of which relate to micro and macroeconomics (Kowalski, 2002). The experiment presented in this paper is connected with the latter, i.e. the nudge concept.

[a] SGH Warsaw School of Economics, Institute of Mathematical Economics, al. Niepodległości 162, 02-554 Warszawa, e-mail: honorata@sgh.waw.pl, ORCID: http://orcid.org/0000-0001-8249-2368.

[b] SGH Warsaw School of Economics, Institute of Mathematical Economics, al. Niepodległości 162, 02-554 Warszawa, e-mail: mramsz@sgh.waw.pl, ORCID: http://orcid.org/0000-0002-0728-5315.

[c] SGH Warsaw School of Economics, Institute of Mathematical Economics, al. Niepodległości 162, 02-554 Warszawa, e-mail: pzawis@sgh.waw.pl, ORCID: http://orcid.org/0000-0002-5297-7644.

Framing (Tversky & Kahneman, 1981) occurs when the way problems are formulated affects decision-making. Experiments based on the 'Asian Disease' are very often studied to exemplify the reasoning behind the idea.

'The Asian Disease' assumes that a deadly illness is threatening to the lives of 600 inhabitants of a town. The task is to choose between two alternative rescue programmes: one is certain, while the other is risky. They are described (framed) either positively or negatively, but equal in their expected value. The positively framed subjects were to select between:

(A) saving 200 people for certain or

(B) saving 600 people with a one-third probability and a two-thirds probability that no one will survive.

The negatively framed subjects were to choose between:

(A') a certain death of 400 people and

(B') a one-third probability that nobody will die and a two-thirds probability that all 600 people will die.

The framing effect itself shows the violation of the invariance principle by choosing a risky gamble (B') rather than something certain (A') when the descriptors are negative (78% chose B' in the loss domain), and a certain option (A) rather than a gamble (B) when the descriptors are positive (72% selected A in the gain domain).

The nudge is a kind of indirect suggestion used to affect decision-making. It was introduced by Thaler and Sunstein (2008). Its most popular example refers to children's choice of healthy food. If we want children to decide on healthy food, we place it at the level of their eyes on the store shelves. The nudge theory has many applications in economics, healthcare and politics.

A priori information is a special type of nudge. We will analyse a case involving a priori information in the context of classical music competitions. A jury consisting of a dozen jurors using a given voting method creates a ranking of contestants from the best to the worst. A lot of nudges may be used. For example, information about the contestant, their achievements (halo effect) or the teachers. The order according to which the performances are presented is important. Whether the performance preceding the given contestant was weak or very good proves significant. The information about the contestant and the previous performances is a priori information.

The impact of a priori information has been analysed in many papers. One of them is Chmurzyńska (2015), where she quotes the research conducted by Manturzewska (1970a, 1970b), repeated in Chmurzyńska and Kamińska (2006). Papers by Manturzewska (1970a, 1970b), and Chmurzyńska and Kamińska (2006) are in Polish and are not widely known. Manturzewska presented five performances of one of Chopin's works to a team of experts. The performances ranged from very

poor, through average to very good. One very good performance and one average were presented twice, preceded by performances of different quality. The result depended on the quality of the preceding performance. The same performance received two different scores. Manturzewska's research results were confirmed and generalised in a number of independently written papers: the impact of the halo effect (Duerksen, 1972; Hunter & Russ, 1996; Radocy, 1976), the influence of a position in a sequence (Flôres & Ginsburgh, 1996; Ginsburgh & van Ours, 2003), and differences between various assessments of the same performance (Fiske, 1977, 1979; Wapnick et al., 1993). There are also papers which present investigations in sports competitions: Bertini et al. (2010), Gambarelli (2008), Gambarelli et al. (2012), and Tyszka and Wielochowski (1991).

The observed results may be found not only in classical music competitions but also in sports competitions, evaluation projects by experts, and other.

In this paper, we attempt to determine whether positive a priori information is likely to reduce the differences in the voting results where strategic voting is applied. By 'positive' we mean that information is provided about which particular alternatives might be highly evaluated.

The source of the experiment presented in this paper is a scenario featuring a university board. In Poland, students are members of such boards. The board has to select its chairperson. There are two main candidates: A and B. The students are divided into two groups. The first group supports candidate A, while the second group supports candidate B. Candidate A promises more funds allocated to sports, while candidate B promises the renovation of student housing. The students want to vote in such a way as to mark their favorite as the best candidate and the opponent as the worst. Before voting, they receive some additional information from the school graduates that C is the best candidate. The students are sure that the graduates correctly evaluate candidates and do not ignore their opinions, but they promise to support their favorite. The experiment described in this paper aims to answer the question of how would the students vote in the presented circumstances. In this scenario, the students receive objective, positive a priori information about the best candidate. It is a positive nudge. Therefore, a situation is created where a bounded rationality effect may occur. We attempt to answer the question above by conducting an experiment in the area of music competitions. Such a scenario was chosen as it has a relatively simple construction. Similar methods may be used in cases where expert opinions relate to sports, science and economic issues.

The paper is constructed as follows: Section 2 presents the used voting methods, while Section 3 introduces the formulated hypotheses. Section 4 describes the whole experiment and Section 5 its results. The conclusions are discussed in Section 6.

2. Voting methods

Three voting methods are used in this paper: the Borda Count method, the Condorcet winner method and the anti-manipulation method.

2.1. The Borda Count

This method was introduced by Jean-Charles de Borda in the 18th century. There are n alternatives. Each voter neutrally ranks the alternatives from the best to the worst. The best alternative is granted n points, the next one n–1 points, then n–2 points, etc. The worst alternative scores 1 point. Each alternative receives the sum of points obtained from all the voters. The alternative with the highest score wins. Let us consider the following example.

Example 1

The example relates to the voting in the final of the 15th International Henryk Wieniawski Violin Competition, presented in Table 1. The competition was held in Poznań, Poland in 2016. There were seven violinists (A…, G) and 11 jurors (J1…, J11). The Borda Count was used. Violinist A won, while violinist B came second. In fact, the inverted Borda Count was used. The best alternative obtained one point, the worst seven and the alternative with the lowest score won. The methods are isomorphic and lead to the same results.

Table 1. The final of the 15th International Henryk Wieniawski Violin Competition

Jurors → Contestants ↓	J1	J2	J3	J4	J5	J6	J7	J8	J9	J10	J11
A	7	3	2	7	7	4	3	7	7	7	7
B	4	7	7	2	2	7	7	2	5	6	5
C	5	5	5	3	6	6	5	5	6	1	6
D	3	6	4	5	1	5	4	4	3	5	1
E	1	4	6	1	3	3	6	3	4	3	4
F	6	2	1	6	4	2	1	6	1	2	2
G	2	1	3	4	5	1	2	1	2	4	3

Source: Kontek and Sosnowska (2020).

This method is used in sports or music competitions and in the assessment of projects. It should be noted that this approach is sensitive to manipulation. The highest number of points is granted to the favorite and the lowest to the opponent. Such manipulation can be observed during classical music competitions (see Table 1). If A has the best note of seven, then some jurors award B only two points. And conversely, if B is allotted seven points, then some jurors give two or three to A. This situation may result from a significant difference in music tastes, but it may also

be the effect of manipulation. Borda said that his method is designed for honest people. The application of voting methods in the context of music competitions is possible since the data are widely available. More about the Borda Count can be found in Gaertner (2013).

2.2. The Condorcet winner

The concept of the Condorcet winner was introduced by Nicolas de Condorcet in the 18th century. It is assumed that each voter has a preference relation which they use to compare alternatives. Let us consider alternatives A and B. Alternative A wins in comparison with B if more than 50% of voters (the majority) prefer A to B. All alternatives are compared. Alternative A is the Condorcet winner if it wins in comparison with each alternative. The Condorcet winner does not always exist. Let us consider the following example.

Example 2

The example is presented in Table 2. There are three voters: J1, J2, J3 and three alternatives: A, B, C. Voter 1 ranks alternatives A, B, C. Voter 2 – B, C, A. Voter 3 – C, A, B. The first alternative is the best, the third is the worst.

Table 2. The Condorcet paradox

Voters → Alternatives ↓	J1	J2	J3
A	3	1	2
B	2	3	1
C	1	2	3

Source: authors' work.

It should be noted that alternative X is before alternative Y in someone's order when X gets a higher score than Y in this order. Table 2 shows that A is twice before B, so A wins in comparison with B. C, on the other hand, is twice before A, so it wins in comparison with A. Therefore, C wins in comparison with A. A does not win in comparison with any alternative, so it is not the Condorcet winner. C loses in comparison with B, thus C is not the Condorcet winner. B loses in comparison with A, so B is not the Condorcet winner. Therefore, for these preferences, the Condorcet winner does not exist. These preferences are called the Condorcet paradox. The preferences are scattered. If the Condorcet winner exists, we deal with preferences that may lead directly to a common decision. In the case of the final of the 15th International Henryk Wieniawski Violin Competition, the Condorcet winner is A. More about the Condorcet winner can be found in Gaertner (2013).

2.3. The anti-manipulation method

No voting method is immune to manipulation except for the dictatorship method (Gibbard, 1974; Satherwaitte, 1975). Dictatorship is the only method which fulfils the conditions of the Arrow Impossibility Theorem (Arrow, 1951), while others meet these conditions only partly. Thus, there is no universally effective method, but we can choose a method with special properties. Kontek and Sosnowska (2020) proposed an approach that may reduce the possibility of manipulation. This method was devised on the basis of the observation of the final of the 15th International Henryk Wieniawski Violin Competition. It involves n voters and is constructed as follows:

- all voters use the Borda Count. Each voter has his or her vector of scores;
- the mean of the obtained scores is computed for every alternative. The vector of means is formed;
- there is a computed distance between each vector of the scores and each vector of the means;
- the whole part of $n/5$, $[n/5]$ is computed: $[n/5]=r$;
- the r voters with the highest distance from the mean are removed. If the distance of more than one voter who is in the r place is the same, the relevant parts are computed. For example, if there are two such voters, their scores are computed with a weight of 1/2;
- the Borda Count is applied to the rest of the voters.

It is assumed that a juror will consider carefully whether to apply manipulation in fear of obtaining extreme results and being removed from the group of jurors whose evaluations are taken into account. According to the anti-manipulation method, violinist B is the winner of the 15th International Henryk Wieniawski Violin Competition. This violinist is different from the Borda winner. Let us note that we indicated that the Condorcet winner might be different from the anti-manipulation method winner.

Some versions of the anti-manipulation method were used for the 5th International Fryderyk Chopin Competition for Amateur Pianists in Warsaw in 2021. The method was mentioned in the special edition of the main Polish classical music journal 'Ruch Muzyczny' devoted to the 18th International Fryderyk Chopin Piano Competition, which was held in Warsaw in October 2021 (Miklaszewski, 2021). In the case of many voters and many alternatives, the calculation of the anti-manipulation method determining the winner is complicated. The computer program may be found in Ramsza and Sosnowska (2020).

3. Hypotheses

In the following parts of the paper, the voting methods presented in Section 2 are discussed. The aim of the paper is to determine how the results of voting change when positive a priori information about the objective ranking is provided. A voting is considered where there is no common favourite for the whole group of voters. Voters are divided into two subgroups, which manipulate voting to get their favourite winner. The hypothesis is formulated as follows:

H.1. The voting result preceded by positive a priori information about the objective ranking is highly positively correlated with this objective ranking when two subgroups of voters are involved, each with a favorite. Favorites do not coincide.

In some experiments, alternatives are labelled with letters, starting with A. In this situation, we can consider the alphabetical order A, B, C... The alphabetical ranking, as the most popular kind, holds a special place in the minds of voters, which implies specific results of the voting. The following hypothesis is formulated:

H.2. When no nudges or common favorites of subgroups occur and alternatives are presented alphabetically, the results of the voting are positively correlated with the alphabetical order.

4. Experiment

The experiment was conducted in May 2021 on a group of undergraduate students of the SGH Warsaw School of Economics, majoring in quantitative methods in economics. The group consisted of 20 students and was divided into two 10-person subgroups. Each subgroup was to select the best musician among musicians A, B, C, D, E, F, G, H. They did not listen to the music, but relied on the provided information only. Each subgroup had its favorite and was told to manipulate the voting so that this person wins. Musician D was the favorite of the first subgroup, while E of the second. Both subgroups knew that the other subgroup was also manipulating the voting. Moreover, they were aware of who the opponent's favorite was. They were also informed that the anti-manipulation method would be used. Each subgroup's voting results were established. Moreover, the votes of the members of both subgroups were combined and a data simulation was performed which determined the voting results of the whole group involved in the research.

The anti-manipulation method based on the Borda Count was used. Therefore, we obtained results for the Borda Count and by applying profiles of preferences for the Borda Count, we were also able to determine the Condorcet winner. Voting based on these three methods was thus analysed.

There were two votings. The first one was conducted using the above knowledge about favorites and manipulation. Before the second vote, some additional a priori information was provided: that the true, objective ranking of musicians was B, C, F, E, D, G, A, H. Then, both subgroups voted for the second time taking into account the same information as in the first voting and the additional information about the true, objective ranking. Again, the anti-manipulation method was used. The anti-manipulation method winner, the Borda winner, as well as the Condorcet winner were determined for both subgroups. The results of the voting in both subgroups were established. Using these data, the voting of the whole group was determined, combing the individual preferences of both subgroups and computing the result of the voting for the group formed in this way. It should be noted that this group had no common favorite as the two subgroups had different favorites.

To summarise, it can be said that the experiment is based on a 3×2 plan. There are 3 voting methods (Borda, Condorcet and anti-manipulation) and 2 votings (without a priori information and with a priori information). The manipulation involving a priori information is a within-subject study (in the first voting each participant is not provided with a priori information and in the second voting each participant receives a priori information), while manipulation by information about favorites is a between-subject study design (each participant receives information only about the favorite of his or her subgroup).

5. Results

The results of both votings in each subgroup and the whole group are presented in Table 3.

Table 3. Winners of the voting determined on the basis of the experiment

Number	Group or subgroup (favorite)	Borda Count	Anti-manipulation method	Condorcet winner
1.	Subgroup 1 (D) without a priori information	D	D	D
2.	Subgroup 2 (E) without a priori information	E	E	E
3.	The whole group without a priori information	A	A	does not exist
4.	Subgroup 1 (D) with a priori information	D	B	D
5.	Subgroup 2 (E) with a priori information	E	E	E
6.	The whole group with a priori information	B	B	B

Source: authors' calculations.

Let us determine whether the winner in the case with no a priori information is A considering the voting of the whole group. The fact that the alphabetical ranking: A, B, C, D, E, F, G, H is the one that comes to people's minds first is no surprise.

When the group is provided with additional a priori information about the objective ranking, the winner changes to B. Thus, B is the best alternative in this objective ranking and additional information changes the winner. Moreover, B, who is the winner, is the best alternative in the ranking obtained due to this information. The results hold for all the considered voting methods. Let us compare the alphabetical ranking and the objective ranking by Kendall Tau. The Kendall Tau (τ) rank correlation coefficient is used to measure the ordinal association between measured quantities (Kendall, 1938). Kendall Tau for these two rankings is equal to 0.35, thus, there is no high or medium correlation between these two rankings. In addition, in one of the subgroups, B (the first one in the voting with additional information) is the winner in the anti-manipulation method. Therefore, the additional information about the objective ranking neutralises the manipulation effect and allows the true, best alternative to win.

Now, let us compare the obtained rankings: the true, objective ranking and the alphabetical ranking using the Kendall Tau. The following notation will be used: r_1 – ranking obtained in voting without additional a priori information, r_2 – ranking obtained in voting with additional a priori information, r_3 – objective ranking, and r_4 – alphabetical ranking. The Kendall Tau correlations are presented in Table 4.

Table 4. Kendall Tau correlations between rankings

Number	Group	Voting method	$\tau(r_1, r_2)$	$\tau(r_1, r_3)$	$\tau(r_1, r_4)$	$\tau(r_2. r_3)$
1.	Whole group	Borda	0.50	0.28	0.92	0.78
2.	Whole group	Anti-manipulation	0.33	0.28	0.92	0.92
3.	Subgroup 1	Borda	0.63	0.21	0.57	0.57
4.	Subgroup 1	Anti-manipulation	0.50	0.21	0.42	0.71
5.	Subgroup 2	Borda	0.50	0.14	0.35	0.64
6.	Subgroup 2	Anti-manipulation	0.35	0.07	0.28	0.71

Source: authors' calculations.

It is worth noting that for all the groups and both methods, the obtained ranking where voters had additional information r_2 was highly correlated with objective ranking r_3. This correlation was especially high for the voting of the whole group. Hypothesis H.1 has therefore been confirmed. The ranking where voters had no additional information r_1 was correlated with alphabetical ranking r_4, especially for the whole group. This indicates that the alphabetical ranking, where the alternatives were listed alphabetically, is a kind of natural ranking. Hypothesis H.2 is thus confirmed. The correlation between ranking r_1 with no additional information for the voters and ranking r_2, i.e. the one with the a priori information provided, is low or medium. There is a low correlation between r_1, where voters have no additional information, and true, objective ranking r_3. Therefore, the result of the voting with

the additional information given is the closest to the true, objective ranking. The additional information about the true, objective ranking makes the voters vote in such a way that the results are very close to those presented by this ranking.

6. Conclusions

It was shown that the anti-manipulation method, the Borda Count and the Condorcet winner methods are sensitive to positive a priori information, thus proving that reducing manipulation is possible. We can use a positive nudge as a weak suggestion, not an evaluation.

During the 18th International Chopin Piano Competition, the jurors did not know how the others voted. On the one hand, it is a positive feature, as it may be assumed that jurors did not influence each other. On the other hand, however, jurors are not the same and some of them may be recognised authorities for others. The knowledge of how these authorities vote may form a positive nudge and reduce manipulation. The knowledge which jurors have about students is another issue – the jurors do not participate in the voting concerning their own learners, but it is common knowledge which jurors taught which students. If there is a student of a juror who is an authority and teaches only very good students, this knowledge may act as positive a priori information and contribute to a higher score than in the case of someone not taught by a recognised authority. Thus, there are many practical questions connected with positive frames or nudges caused by the sensitivities of the voting methods.

Acknowledgments

The authors would like to express their gratitude to their colleagues from the National Science Centre, Poland (UMO-2018/31/B/HS4/01005 Opus 16) grant seminar for their valuable remarks and comments.

The research was supported by grant no. KAE/S21 awarded by the SGH Warsaw School of Economics and grant no. UMO-2018/31/B/HS4/01005 Opus 16 of the National Science Centre, Poland.

References

Arrow, K. J. (1951). *Social Choice and Individual Values*. John Wiley & Sons.

Bertini, C., Gambarelli, G., & Uristani, A. (2010). Indices of Collusion among Judges and Anti-collusion Average. In S. Greco, R. A. Marques Pereira, M. Squillante, R. R. Yager & J. Kacprzyk (Eds.), *Preferences and Decisions. Models and Applications* (pp. 195–210). Springer-Verlag. https://doi.org/10.1007/978-3-642-15976-3_12.

Chmurzyńska, M. (2015). *Influence of* a priori *information on music performance assessment*. https://www.academia.edu/23461055/Influence_of_a_priori_information_on_music_performance_assessment.

Chmurzyńska, M., & Kamińska, B. (Eds.). (2006). *Ocenianie wykonań muzycznych*. Wydawnictwo Akademii Muzycznej im. Fryderyka Chopina.

Duerksen, G. L. (1972). Some effects of expectation on evaluation of recorded musical performance. *Journal of Research in Music Education*, *20*(2), 268–272. https://doi.org/10.2307/3344093.

Fiske, H. E. (1977). Relationship of selected factors in trumpet performance adjudication reliability. *Journal of Research in Music Education*, *25*(4), 256–263. https://doi.org/10.2307/3345266.

Fiske, H. E. (1979). Musical performance evaluation ability: Toward a model of specificity. *Bulletin of the Council for Research in Music Education*, (59), 27–31.

Flôres, R. G., & Ginsburgh, V. A. (1996). The Queen Elisabeth musical competition: how fair is the final ranking?. *Journal of the Royal Statistical Society (the Statistician)*, *45*(1), 97–104. https://doi.org/10.2307/2348415.

Gaertner, W. (2013). *A Primer in Social Choice Theory*. Oxford University Press.

Gambarelli, G. (2008). The 'coherent majority averages' for juries' evaluation processes. *Journal of Sport Sciences*, *26*(10), 1091–1095. https://doi.org/10.1080/02640410801930135.

Gambarelli, G., Iaquinta, G., & Piazza, M. (2012). Anti-collusion indices and averages for the evaluation of performances and judges. *Journal of Sport Sciences*, *30*(4), 411–417.

Gibbard, A. (1974). A Pareto-Consistent Libertarian Claim. *Journal of Economic Theory*, *7*(4), 388–410. https://doi.org/10.1016/0022-0531(74)90111-2.

Ginsburgh, V. A., & van Ours, J. C. (2003). Expert Opinion and Compensation: Evidence from a Musical Competition. *American Economic Review*, *93*(1), 289–296. https://www.aeaweb.org/articles?id=10.1257/000282803321455296.

Hunter, D., & Russ, M. (1996). Peer assessment in performance studies. *British Journal of Music Education*, *13*(1), 67–78. https://doi.org/10.1017/S0265051700002953.

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrica*, *30*(1-2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81.

Kontek, K., & Sosnowska, H. (2020). Specific Tastes or Cliques of Jurors? How to Reduce the Level of Manipulation in Group Decisions?. *Group Decision and Negotiation*, *19*(6), 1057–1084. https://doi.org/10.1007/s10726-020-09694-y.

Kowalski, T. (2002). The Simonian bounded rationality hypothesis and the expectation formation mechanism. *Economic and Business Review*, *2*(1), 5–24. https://www.ebr.edu.pl/pub/2002_1_5.pdf.

Manturzewska, M. (1970a). Rzetelność ocen wykonawstwa wydawanych przez ekspertów muzycznych. *Ruch Muzyczny*, (21), 3–8.

Manturzewska, M. (1970b). Rzetelność ocen wykonawstwa wydawanych przez ekspertów muzycznych. *Ruch Muzyczny*, (23), 13–15.

Miklaszewski, K. (2021, October 2). *Dlaczego nie lubię konkursów?*. http://ww.ruchmuzyczny.pl/article/1456-dlaczego-nie-lubie-konkursow.

Radocy, R. E. (1976). Effects of authority figure biases on changing judgements on musical events. *Journal of Research in Music Education*, *24*(3), 119–128. https://doi.org/10.2307/3345155.

Ramsza, M., & Sosnowska, H. (2020). Trials of Characterization of Anti-manipulation Method. In N. T. Nguyen, R. Kowalczyk, J. Mercik, A. Motylska-Kuźma (Eds.), *Transaction on Computational Collective Intelligence XXXV* (pp. 21–37). Springer-Verlag. https://doi.org/10.1007/978-3-662-62245-2_2.

Satherwaitte, M. A. (1975). Strategy-proofness and Arrow's condition: existence and correspondence for voting procedures and social welfare functions. *Journal of Economic Theory*, *10*(2), 187–217. https://doi.org/10.1016/0022-0531(75)90050-2.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge. Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683.

Tyszka, T., & Wielochowski, M. (1991). Must boxing verdicts be biased. *Journal of Behavioral Decision Making*, *4*(4), 283–295. https://doi.org/10.1002/bdm.3960040406.

Wapnick, J., Flowers, P., Alegant, M., & Jasinskas, L. (1993). Consistency in Piano Performance Evaluation. *Journal of Research in Music Education*, *41*(4), 282–292. https://doi.org/10.2307/3345504.

Przegląd Statystyczny. Statistical Review, 2022, vol. 69, 4, 41–60 https://doi.org/10.59139/ps.2022.04.4
Received: 31.01.2023, revised: 14.04.2023, accepted: 24.04.2023

Comparison of the accuracy of forecasts based on neural networks before and after the outbreak of the COVID-19 pandemic on the example of selected exchange rates

Jakub Morkowski[a]

Abstract. This article examines the impact of the COVID-19 pandemic on the accuracy of forecasts for three currency pairs before and after its outbreak based on neural networks (ELM, MLP and LSTM) in terms of three factors: the forecast horizon, hyper parameterisation and network type.

Keywords: neural network, currency market, forecasts, COVID-19 pandemic
JEL: C45, C53, E44

1. Introduction

The outbreak of the COVID-19 pandemic has been one of the most important events of recent years. It has had a significant impact on many aspects of life including demographics, through a recorded increase in mortality and a decrease in the birth rate (Balbo et al., 2020), the level of education, as schools were closed to prevent the spread of the virus (Daniel, 2020), and growing domestic violence (Boserup et al., 2020). Along with the increase in domestic crimes and the isolation of children from their peers, the pandemic has also had an impact on people's mental health (Cullen et al., 2020; Pfefferbaum & North, 2020; Usher et al., 2020). Cultural life also suffered since cultural events were either postponed or cancelled (Akser, 2020).

This article focuses on another aspect of the phenomenon, i.e. on the impact of the economic uncertainty on stock markets and the closure of specific industries and bankruptcies of companies.

The impact of COVID-19 was also very quickly and extensively recognised in the case of the broadly understood economy. On 24 February 2020, considerable drops were recorded on stock exchanges worldwide caused by an increasing number of infections (mainly in China). Prices of other assets, such as crude oil, gas, cryptocurrencies and corporate bonds also decreased. It is estimated that for the first

[a] Poznań University of Economics and Business, Applied Mathematics Department, al. Niepodległości 10, 61-875, Poznań, e-mail: jakub.morkowski@ue.poznan.pl, ORCID: https://orcid.org/0000-0001-5727-5089.

nine days of March 2020, listed companies lost a total of USD 9 trillion in value (Raifu et al., 2021). Transport limitations caused further economic problems as did the frequent excessive purchases of necessities done by the population as they deregulated the logistics sector. The purpose of this study, however, is to check the impact of the pandemic on the accuracy of forecasts generated by neural networks.

The influence that COVID-19 has had on various aspects of economic life has been widely analysed in the latest scientific articles. The impact of the pandemic on the stock market on the example of indices such as IBEX35, FTSE100, DAX30, CAC40, and others are examined in Zeren & Hizarci (2020). The research in this article has been conducted over a short period, from January to March 2020, and indicated that investing in stocks after the COVID-19 outbreak was very risky and safer forms of investment should have been sought. After the virus appeared, information on the number of deaths and infections on a given day was provided frequently and on a regular basis. In the study by Ashraf (2020), where stock markets in as many as 64 countries were analysed, it was demonstrated that the information about the increase in the number of deaths was detrimental to stock markets and that their reaction was immediate. It should be emphasized that COVID-19 significantly affected other markets, e.g. those relating to oil: the impact of the pandemic on the volatility of the oil markets exceeded the consequences caused by the global financial crisis of 2008. The sharp drop in oil prices created an unprecedentedly high level of risk, causing investors to suffer from major losses in the short term (Zhang & Hamori, 2021). Countless articles also describe how the COVID-19 pandemic influenced currency markets. A vast linear relationship was observed between the number of confirmed deaths and the stability of the US and Chinese currencies (Li et al., 2022). A risk analysis of six currency pairs (USD/EUR, USD/GBP, USD/JPY, USD/CNY, USD/BRL, USD/TRY) shows that in the early months of the pandemic, the movements in the currency markets were not as intense as during the 2008 crisis. However, the Diebold-Yilmaz spillover index demonstrated that in the long run, the shock wave of the COVID-19 pandemic was about eight times greater than in 2008 (Gunay, 2021).

The second key element of this article is neural networks. The literature on forecasting based on neural networks states that the ones often used are ELM (Extreme Learning Machines), MLP (Multilayer Perceptrons) and LSTM (Long short-term memory) (Das et al., 2021; Sun et al., 2018; Wu & Gao, 2018). Therefore, the aforementioned networks were applied to make predictions in the empirical part of this article. The main objective of this study was to verify the accuracy of exchange rate forecasts generated using neural networks before and after the outbreak of the COVID-19 pandemic. The study is based on three neural networks with different hyperparameters. In their article, Abedin et al. (2021) indicated that the ensemble

deep learning method based on the LSTM network achieved better results than the proposed benchmark regarding committed forecast errors for exchange rates. However, it was noted that the differences in the size of the prediction errors before and after the COVID-19 outbreak were substantial. When investing in currency markets, it is crucial to identify and measure the size of the error made in forecasting and to be aware of the correctly forecasted direction of change. Thus, this article also analyses neural networks which are less complex than LSTM, ELM and MLP. They were examined in terms of the accuracy obtained in forecasting the direction of change in the exchange rate.

The choice of exchange rates as the forecast asset was dictated by the fact that exchange rate fluctuations have a great impact on individual countries' economies. In the era of globalisation, exchange rates directly affect the operation of corporations, enterprises and individual investors (Markova, 2019). Many consider the exchange rate as a factor reflecting the current situation and condition of a given country's economy (Kartono et al., 2020). In this context, the most significant currency pairs are USD/EUR, GBP/EUR and CHF/EUR.

2. Methodology

Forecasting currency prices using neural networks is made at price levels. Three different neural networks were used in the study, namely:

- MLP (Rosenblatt, 1958);
- ELM (Rumelhart et al., 1986);
- LSTM (Hochreiter & Schmidhuber, 1997).

The MLP network is the simplest network used in the empirical studies presented in this article. It is built of at least three different layers, each performing its characteristic tasks:

- input layer – receives the signal for processing;
- hidden layers – responsible for processing signals from the input layer so as to generate auxiliary data. These data form the basis for determining the final solution through the output layers. Hidden layers mediate between the input and output layers, and their effect is visible indirectly through the output layer results;
- output layer – it returns the result of the calculations made in the hidden layers.

The way MLP networks operate is simple. Each neuron from each layer calculates a weighted sum of its inputs. The calculated activation level is an argument passed to

the activation function which calculates the neuron's output. Each node, in addition to the input nodes, has a non-linear activation function which can take different forms. Examples include the sigmoidal function. MLP is a feedforward network, which means that the signal between the input layer and output layer runs only one way, i.e. from the input nodes, through the hidden nodes to the output nodes. Despite its simplicity, MLP networks can approximate any continuous function and solve problems that are not linear (Abirami & Chitra, 2020). An essential step in designing MLP networks is determining the appropriate number of layers and neurons in the layers.

ELM was initially developed for single hidden layer feedforward neural networks (SLFNs), i.e. as a single layer network. It uses a continuous and differentiable activation function to activate the hidden neurons, with sigmoidal and Gaussian functions used most often (Jastrzębski et al., 2015). Neural network learning occurs in two stages. The ELM network learning paradigm assumes the random generation of hidden layer parameters to map the input data into the feature space. These parameters remain constant (independent of the learning process). The second stage is based on minimising the squared error present in the first stage. As a result, the weights connecting the hidden layer and the output layer are determined. The output layer weights are obtained using the generalised inverse of the output matrix of the hidden layer.

LSTM belongs to the class of recurrent networks. A characteristic feature of LSTM networks is their structure designed to remember short patterns (hence 'short-term' in the name). LSTM was initially developed for sequence analysis and used in text sequence analysis. It contributed to the development of a number of applications, including Google Translate, Siri and Google's voice assistant. In the later stages of development, it was also adapted into time series (Smagulova & James, 2019). The LSTM network consists of connected multiple recurrent memory blocks. Each block includes three gates: an input, output and memory gate. What distinguishes LSTM from other recurrent networks are memory blocks. In classical recurrent networks, there is a data flow within the network. However, LSTM additionally has what is called a long memory: thanks to the gates, the LSTM network can store data for more than one period (Van Houdt et al., 2020). Another important difference between classical recurrent networks and LSTM worth mentioning here is the fact that the memory cell receives information from three input sources.

For each of the neural networks, 49 different combinations of hyperparameters describing the network were used. The assessed hyperparameters are the number of lags and the number of hidden nodes. Both hyperparameters can take seven different values. The number of hidden nodes took the values of 2, 5, 10, 15, 20, 25, 50, and the number of lags used in the neural network took the values of 1, 2, 3, 4, 5, 6, 7.

The number of hidden nodes did not take the form of consecutive natural numbers but at certain intervals so as to search for the optimal order of the magnitude of nodes and to examine the forecast results' dependence on this parameter. Other neural network settings, particularly the LSTM network are fine-tuned in the training process. The empirical part of the study considered three currency pairs: CHF/EUR, GBP/EUR and USD/EUR.

This paper describes the accuracy of individual neural networks. The forecast accuracy in this study is understood as the ratio of correctly forecasted directions of changes in relation to all of the produced forecasts. Each day, a forecast is made for 10 different horizons (from 1 to 10 days) using all combinations of two hyperparameters, which allows the determination of not only one relevance for a given network, but also the maximum, minimum and average relevance for a given network in each horizon. The study compares these values before and after the outbreak of the COVID-19 pandemic.

3. Data description

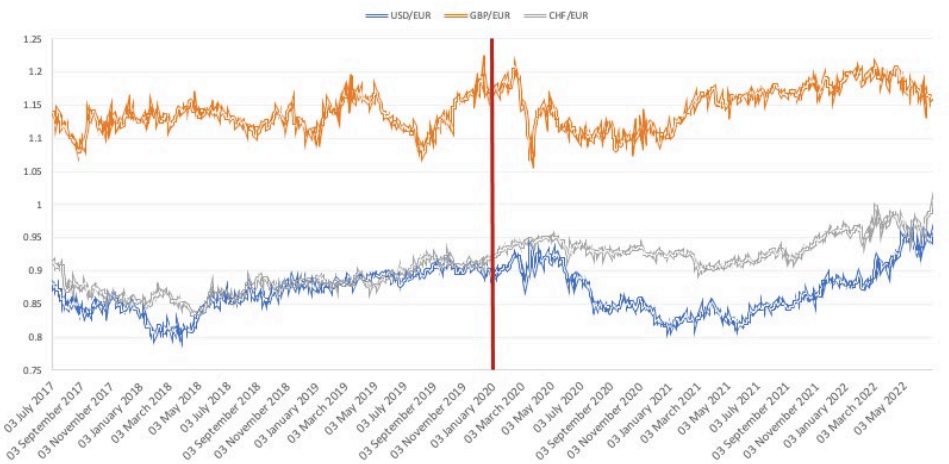
As previously mentioned, the currency pairs used in this study were USD/EUR, GBP/EUR and CHF/EUR. Forecasts with three neural networks for all currency pairs were made from 1 July 2017 to 30 June 2022, and the data were daily. It was assumed that the data before the COVID-19 outbreak included 10 quarters from 1 July 2017 to 31 December 2019; after the pandemic outbreak, we also considered 10 quarters: from 1 Jan 2020 to 30 June 2022. When generating forecasts, the data were divided into training and test sets. Therefore, the training data extended the data adopted in the study to 100 days before 1 July 2017. The study was carried out at price levels and the horizon length extended from 1 to 10 days. For forecasts more than one day ahead, a direct forecast was used. The training set was always composed of data one hundred days before the day the forecast was computed, which means that the training set always contained an equal number of observations, but their scope depended on the day of the forecast. Table 1 shows the descriptive statistics of the rates of return studied in two periods and Figure 1 shows the relevant exchange rates of the same period. The red line indicates the COVID-19 pandemic outbreak date adopted in the study, i.e. 1 January 2020.

Table 1. Descriptive statistics of the currency pairs

Currency pair	Period	Mean	Standard deviation	Skewness	Kurtosis
USD/EUR	01/07/2017–31/12/2019	0.00002	0.00401	0.16324	0.78288
	01/01/2020–30/06/2022	0.00011	0.00446	0.18710	1.61366
CHF/EUR	01/07/2017–31/12/2019	0.00007	0.00440	0.03802	1.82360
	01/01/2020–30/06/2022	–0.00002	0.00445	–0.65019	3.33893
GBP/EUR	01/07/2017–31/12/2019	0.00002	0.00301	–0.00539	0.89935
	01/01/2020–30/06/2022	0.00013	0.00302	0.30309	4.07684

Source: author's work based on stooq.com.

Figure 1. Currency rate chart – USD/EUR, CHF/EUR and GBP/EUR



Source: author's work based on stooq.com.

The information presented in Table 1 indicated little difference between the average rate of return, which for all currency pairs and regardless of the period studied was close to zero. Similar conclusions could be drawn from the standard deviation. However, the skewness and kurtosis show differences for all currency pairs. Particularly for the GBP/EUR pair, the skewness changed from positive for the 2.5 years before the pandemic outbreak to negative for the 2.5 years after the COVID-19 pandemic outbreak. There was also a change in skewness for the CHF/EUR pair, although in this case from negative to positive. Significant changes in kurtosis could also be seen for these pairs. The skewness change was small for the USD/EUR pair, while the kurtosis change was smaller than in the case of the other currency pairs. Figure 1 shows the price levels of the three currency pairs under study. Price levels are intended to illustrate the trends in the currency pairs during the period under study. Figure 1 indicates that just before the end of 2019, all the currencies studied were in an uptrend. After the pandemic outbreak, the largest

losses were recorded for the GBP/EUR pair. For USD/EUR, there was also a decline at the same time as for GBP/EUR, but the change was less abrupt, and the exchange rate of this currency returned more quickly to the levels before the decline. The slightest fluctuations in the period after the outbreak of the COVID-19 pandemic were recorded for the CHF/EUR pair.

4. Research

The empirical study predicted three currency pairs using 49 different combinations of hyperparameters for three neural networks, making it a total of 147 models analysed. Many of the obtained forecasts were aggregated in search of answers on the impact of the COVID-19 pandemic on the accuracy of forecasts generated using neural networks. The first part of the empirical study focused on more general conclusions. Then a more in-depth analysis was performed to search for more detailed conclusions. As mentioned in Section 3, the data were divided into 20 quarters. Forecasts were made for individual days for each of the quarters. Accordingly, the data in Tables 2–7 refer to the accuracy of the forecasts for the days included in each quarter. Accuracy in this case describes the number of correctly forecasted directions of change in relation to all of the produced forecasts. If the table does not contain a division by neural network, it should be assumed that the data include accuracy across all neural networks.

Tables 2–4 present the average accuracy for all 147 models, broken down by quarter and forecast horizon but not by neural network types. These tables show that for the majority of forecast horizons and throughout most of the studied period, neural networks achieved accuracy exceeding 50%.

Table 2. The average prediction accuracy for all neural networks for the USD/EUR currency pair

Period	*t*+1	*t*+2	*t*+3	*t*+4	*t*+5	*t*+6	*t*+7	*t*+8	*t*+9	*t*+10
	%									
Q3 2017	**52.49**	44.89	45.72	46.06	**51.19**	46.72	45.39	47.53	47.70	43.90
Q4 2017	**58.41**	**54.21**	**54.52**	**55.31**	**55.57**	**55.73**	**55.47**	**63.91**	**66.43**	**67.32**
Q1 2018	49.52	**51.71**	**56.48**	**50.54**	**52.82**	49.17	**50.88**	48.33	46.09	47.38
Q2 2018	47.83	45.60	43.32	43.30	42.13	41.73	45.13	46.98	47.05	47.96
Q3 2018	**55.55**	**59.54**	**63.84**	**59.90**	**56.86**	**55.79**	**57.88**	**61.13**	**62.80**	**61.40**
Q4 2018	**52.63**	**54.89**	**53.19**	49.38	49.21	**61.22**	**59.85**	**55.11**	**53.90**	**50.28**
Q1 2019	**58.81**	**56.56**	**60.45**	**64.60**	**69.06**	**69.38**	**71.55**	**70.69**	**68.54**	**67.06**
Q2 2019	49.43	**51.12**	**53.21**	**59.86**	**56.05**	**56.16**	**58.04**	**54.27**	**52.25**	**50.80**
Q3 2019	48.61	**55.71**	**57.79**	**58.48**	**55.94**	**56.95**	**55.80**	**55.36**	**58.79**	**54.07**
Q4 2019	**51.91**	**51.91**	**55.81**	**60.00**	**63.70**	**66.95**	**63.08**	**67.56**	**64.13**	**62.00**
Q1 2020	47.49	44.90	46.18	**50.80**	**52.77**	**54.10**	**55.67**	**51.39**	**57.75**	**55.53**
Q2 2020	46.03	**50.62**	49.97	49.85	**52.30**	**51.42**	**55.91**	**55.72**	**57.94**	**54.25**
Q3 2020	43.65	49.29	45.80	48.44	46.85	47.39	48.49	43.93	45.81	41.99
Q4 2020	**51.26**	**54.64**	**51.97**	**53.49**	**51.65**	49.71	**53.20**	48.43	**54.89**	49.08
Q1 2021	**50.81**	**53.77**	**50.91**	**52.25**	**52.33**	**52.70**	**54.30**	**52.55**	**60.76**	**53.02**
Q2 2021	**58.84**	**58.76**	**59.55**	**59.68**	**60.74**	**65.82**	**62.17**	**57.58**	**62.79**	**62.85**
Q3 2021	47.20	**53.93**	48.29	44.67	46.17	49.34	48.52	47.44	**53.76**	**54.32**
Q4 2021	**56.35**	**54.19**	**50.02**	43.88	48.92	47.81	**50.80**	47.35	**53.30**	**50.01**
Q1 2022	**59.95**	**58.83**	**56.37**	**58.72**	**61.15**	**62.38**	**63.24**	**59.57**	**60.93**	**61.07**
Q2 2022	**57.14**	**53.13**	45.26	43.74	44.28	46.65	48.47	46.80	47.19	42.58

Note. *t*+1, *t*+2, ..., *t*+10 – forecast horizons. The numbers in bold indicate accuracy above 50%.

Source: author's work.

Table 3. The average accuracy of forecasts for all neural networks for the CHF/EUR currency pair

Period	*t*+1	*t*+2	*t*+3	*t*+4	*t*+5	*t*+6	*t*+7	*t*+8	*t*+9	*t*+10
	%									
Q3 2017	**54.71**	49.58	**58.42**	**59.93**	**61.12**	**66.69**	**64.72**	**64.31**	**63.10**	**62.18**
Q4 2017	**50.65**	44.66	42.67	44.83	42.64	40.13	36.80	32.87	35.66	34.52
Q1 2018	46.51	**50.98**	48.83	49.67	45.93	43.36	43.92	43.29	41.95	38.77
Q2 2018	46.18	40.66	38.83	41.55	39.35	38.17	36.23	37.80	35.20	37.08
Q3 2018	**54.42**	**55.46**	**53.60**	**55.68**	**57.06**	**56.18**	**55.68**	**58.22**	**58.53**	**54.88**
Q4 2018	**51.80**	**52.93**	48.94	**53.04**	**53.81**	**53.17**	**51.27**	**51.57**	**54.82**	**56.19**
Q1 2019	**50.66**	**54.83**	**55.85**	**54.55**	**55.41**	**57.95**	**53.70**	**55.47**	**57.02**	**54.45**
Q2 2019	48.22	**51.88**	47.34	45.54	46.71	46.04	46.12	46.01	43.29	38.76
Q3 2019	**50.59**	**54.95**	**57.75**	**58.29**	**56.35**	**57.96**	**59.61**	**61.85**	**68.08**	**63.92**
Q4 2019	**50.63**	**58.16**	**52.80**	**56.63**	**61.58**	**63.84**	**60.94**	**60.30**	**55.48**	**52.26**
Q1 2020	44.35	48.54	40.18	42.50	38.65	38.67	37.82	37.18	40.99	35.67
Q2 2020	**52.16**	**54.62**	48.24	**54.30**	**52.97**	**50.87**	48.45	47.83	45.59	43.01
Q3 2020	**52.21**	**52.88**	**51.52**	**55.94**	**51.11**	**51.75**	**51.77**	**51.72**	**53.82**	49.92
Q4 2020	**52.85**	**53.59**	**56.44**	**59.15**	**50.17**	49.51	**55.15**	**54.66**	**58.50**	**54.19**
Q1 2021	**54.78**	**56.42**	**55.71**	**54.85**	**53.31**	49.40	**58.51**	**57.22**	**60.94**	**56.36**
Q2 2021	**55.13**	**53.05**	**55.11**	**58.92**	**52.87**	49.48	**55.82**	**50.91**	**57.20**	49.39
Q3 2021	**54.14**	49.60	49.44	**52.15**	**51.56**	47.46	**51.98**	**54.25**	**54.81**	**54.23**
Q4 2021	**50.55**	45.73	43.32	38.72	36.98	33.20	38.61	38.06	41.49	35.97
Q1 2022	**52.62**	**50.29**	42.80	47.83	48.77	45.10	**52.98**	**51.65**	**56.36**	**52.61**
Q2 2022	**54.29**	**51.52**	48.05	**56.16**	**51.06**	**51.35**	**62.37**	**62.05**	**65.19**	**63.95**

Note. As in Table 2.

Source: author's work.

Table 4. The average prediction accuracy for all neural networks for the GBP/EUR currency pair

Period	*t*+1	*t*+2	*t*+3	*t*+4	*t*+5	*t*+6	*t*+7	*t*+8	*t*+9	*t*+10
	%									
Q3 2017	46.28	40.81	37.33	39.74	38.77	39.83	37.68	34.81	32.45	30.90
Q4 2017	**56.73**	**61.80**	**60.83**	**60.45**	**62.43**	**65.38**	**65.16**	**69.44**	**66.96**	**68.97**
Q1 2018	**61.69**	**62.58**	**64.34**	**70.04**	**66.51**	**69.16**	**69.48**	**70.04**	**69.67**	**72.06**
Q2 2018	**55.94**	**53.89**	**61.90**	**60.57**	**62.39**	**55.87**	**52.54**	**55.18**	**57.04**	**60.75**
Q3 2018	47.32	**51.82**	48.57	47.08	47.90	45.39	46.44	44.30	44.72	43.69
Q4 2018	48.23	48.39	**51.10**	**59.52**	**59.90**	**57.52**	**59.09**	**55.97**	**58.20**	**55.48**
Q1 2019	**50.63**	**51.51**	**52.71**	**53.81**	**53.09**	**54.82**	48.48	**52.09**	**56.04**	**54.28**
Q2 2019	46.91	43.08	43.03	46.01	45.39	46.65	45.87	48.86	**51.28**	46.99
Q3 2019	47.22	42.36	42.52	34.50	34.71	31.77	36.46	36.53	34.24	37.12
Q4 2019	49.67	**50.61**	**52.58**	48.17	49.64	**50.14**	49.15	48.45	49.26	49.21
Q1 2020	47.12	**51.27**	48.86	**52.43**	47.48	**50.86**	**50.80**	49.08	48.13	48.73
Q2 2020	49.23	**50.34**	**51.03**	49.85	44.21	**52.36**	**53.32**	**53.89**	**54.87**	**52.74**
Q3 2020	49.25	48.67	**57.02**	**57.06**	**52.03**	**58.03**	**59.61**	**59.08**	**59.38**	**61.23**
Q4 2020	**57.35**	**55.82**	**64.42**	**62.47**	**58.10**	**60.97**	**63.61**	**62.53**	**62.64**	**58.63**
Q1 2021	44.54	44.50	46.85	48.76	42.66	47.61	48.48	46.91	45.71	46.04
Q2 2021	**57.38**	**58.26**	**64.15**	**60.97**	**57.13**	**64.97**	**65.45**	**63.67**	**61.87**	**59.40**
Q3 2021	**56.47**	**61.80**	**65.16**	**62.77**	**57.39**	**63.19**	**62.05**	**64.97**	**62.27**	**59.72**
Q4 2021	**51.72**	**52.46**	**51.71**	**51.48**	45.40	**53.24**	49.93	**58.38**	**53.32**	**55.23**
Q1 2022	**53.08**	**59.86**	**61.80**	**58.62**	**56.03**	**62.50**	**63.70**	**65.57**	**57.93**	**59.55**
Q2 2022	**51.30**	49.84	**53.72**	**51.48**	44.09	49.31	45.27	**53.44**	47.52	49.16

Note. As in Table 2.

Source: author's work.

Tables 2–4 above indicate a situation in which the accuracy of forecasts grew as the forecast horizon increased. This situation can have two potential causes. The first is the characteristics of the LSTM network and the consistent flow of gradients through the network. The gradient in a recurrent neural network is responsible for remembering how many errors the network makes in the successive iterations. The second reason for higher accuracy over longer horizons may be the influence of trends. If the network correctly recognizes a trend, it makes fewer errors in multiple-day forecasts as it follows the trend.

To illustrate the results, Figure 2 presents the changes in the accuracy of forecasts in individual quarters for 1-, 2-, 3-, and 4-day-ahead forecasts for the USD/EUR currency pair.

Figure 2. The average accuracy of forecasts for the USD/EUR network in a 1- to 4-day forecast horizon

Source: author's work.

When comparing the last period before the pandemic (Q4 2019) and the first period of the pandemic (Q1 2020) shown in Figure 2, a decrease in the accuracy of forecasts was observed. However, it should be objectively noted that the decrease in the accuracy of forecasts at the beginning of 2020 was not considerably lower than the accuracy of forecasts for Q3 2017 with 1-, 2- or 3-day forecasts or with 1-, 2-, 3- and 4-day forecasts in Q2 2018.

Not all currency pairs showed significant drops compared to the period prior to the COVID-19 outbreak. The average accuracy of forecasts, despite a temporary decrease, very often returned to the levels recorded before the pandemic.

In the next stage of the study, only the accuracy of forecasts obtained in the quarter preceding the start of the pandemic and Q1 2020 were compared. In this case, however, the data were broken down into individual types of neural networks.

Table 5. Comparison of forecast accuracy in Q4 2019 and Q1 2020 for the USD/EUR currency pair

NN	Period	USD/EUR									
		t+1	*t*+2	*t*+3	*t*+4	*t*+5	*t*+6	*t*+7	*t*+8	*t*+9	*t*+10
		%									
ELM	Q4 2019	**53.85**	**52.72**	**58.05**	**62.61**	**67.06**	**68.57**	**64.33**	**68.10**	**63.39**	**62.92**
	Q1 2020	**50.64**	**50.13**	**52.20**	**58.26**	**61.32**	**63.52**	**67.67**	**58.77**	**66.74**	**63.74**
	Difference ...	–3.21	–2.59	–5.85	–4.35	–5.74	–5.05	3.33	–9.33	3.35	0.82
MLP	Q4 2019	**52.24**	**51.15**	**53.66**	**57.24**	**62.29**	**67.82**	**62.07**	**65.87**	**60.28**	**56.30**
	Q1 2020	44.96	36.93	41.36	42.35	43.94	46.43	**51.40**	45.63	49.11	49.36
	Difference ...	–7.28	–14.22	–12.30	–14.89	–18.35	–21.39	–10.67	–20.24	–11.18	–6.93
LSTM	Q4 2019	49.63	**51.87**	**55.71**	**60.15**	**61.76**	**64.47**	**62.82**	**68.72**	**68.72**	**66.78**
	Q1 2020	46.88	47.66	44.98	**51.79**	**53.03**	**52.34**	47.95	49.78	**57.41**	**53.48**
	Difference ...	–2.76	–4.21	–10.73	–8.36	–8.72	–12.13	–14.87	–18.94	–11.31	–13.30

Note. As in Table 2. The numbers are marked red (green) when the forecast accuracy for Q4 2019 is higher (lower) than that for Q1 2020.

Source: author's work.

Table 6. Comparison of forecast accuracy in Q4 2019 and Q1 2020 for the CHF/EUR currency pair

NN	Period	CHF/EUR									
		t+1	*t*+2	*t*+3	*t*+4	*t*+5	*t*+6	*t*+7	*t*+8	*t*+9	*t*+10
		%									
ELM	Q4 2019	**52.18**	**54.19**	49.80	49.61	**55.67**	**58.30**	**55.38**	**53.75**	**50.61**	**54.91**
	Q1 2020	36.26	40.21	31.31	29.08	25.16	23.95	21.17	19.61	19.36	15.27
	Difference ...	–15.93	–13.98	–18.48	–20.53	–30.51	–34.36	–34.21	–34.14	–31.26	–39.64
MLP	Q4 2019	48.63	**57.83**	**51.65**	**60.72**	**66.81**	**68.82**	**65.78**	**64.21**	**58.15**	**57.77**
	Q1 2020	43.97	**50.77**	41.14	47.90	44.83	47.19	43.27	44.39	45.18	44.87
	Difference ...	–4.66	–7.07	–10.51	–12.83	–21.98	–21.63	–22.51	–19.82	–12.96	–12.90
LSTM	Q4 2019	**51.06**	**62.45**	**56.96**	**59.56**	**62.27**	**64.40**	**61.65**	**62.93**	**57.69**	44.10
	Q1 2020	**52.83**	**54.65**	48.10	**50.52**	45.97	44.86	49.01	47.55	**58.43**	46.86
	Difference ...	1.77	–7.81	–8.86	–9.04	–16.30	–19.54	–12.64	–15.38	0.74	2.76

Note. As in Table 5.

Source: author's work.

Table 7. Comparison of forecast accuracy in Q4 2019 and Q1 2020 for the GBP/EUR currency pair

NN	Period	GBP/EUR									
		t+1	*t*+2	*t*+3	*t*+4	*t*+5	*t*+6	*t*+7	*t*+8	*t*+9	*t*+10
		%									
ELM	Q4 2019	47.03	45.90	43.80	34.07	36.04	31.65	30.80	28.76	25.71	25.49
	Q1 2020	43.27	46.88	45.25	44.90	46.91	46.08	48.05	44.04	40.88	37.85
	Difference ...	–3.76	0.97	1.45	10.83	10.86	14.43	17.25	15.28	15.17	12.36
MLP	Q4 2019	**52.94**	**53.59**	**53.78**	**53.03**	**52.18**	**55.23**	**51.15**	**52.09**	**54.79**	**52.28**
	Q1 2020	**55.93**	**52.61**	**50.73**	**54.72**	**53.95**	49.43	**51.72**	**51.88**	**51.02**	**55.01**
	Difference ...	3.00	–0.98	–3.05	1.69	1.77	–5.80	0.58	–0.21	–3.77	2.73
LSTM	Q4 2019	49.05	**52.34**	**60.15**	**57.40**	**60.70**	**63.55**	**65.49**	**64.51**	**67.29**	**69.85**
	Q1 2020	42.15	**54.31**	**50.59**	**57.66**	41.59	**57.07**	**52.64**	**51.32**	**52.49**	**53.33**
	Difference ...	–6.90	1.96	–9.56	0.26	–19.11	–6.49	–12.86	–13.19	–14.80	–16.53

Note. As in Table 5.

Source: author's work.

When summarising the values in the 'difference' rows, it can be observed that in 68 cases, a decrease in the accuracy of forecasts was recorded in Q1 2020 compared to Q4 2019. In 22 cases, on the other hand, an increase appeared in the accuracy of forecasts. However, two crucial features of the presented data should be noted. As many as 16 out of the 22 cases concern the GBP/EUR currency pair. For this pair, the lowest differences between the examined accuracy in the surveyed quarters were recorded. Therefore, only six times higher accuracy was noted in Q1 2020 than in Q4 2019 for the USD/EUR and CHF/EUR currency pairs (three each). For these two currency pairs, the differences between Q1 2020 and Q4 2019 in several cases oscillated within a dozen or so percentage points of difference. At this point, the conjecture about the negative impact of the outbreak of the COVID-19 pandemic on the accuracy of forecasts generated using neural networks has been confirmed.

In order to look for more complex relationships between the accuracy of individual neural network forecasts and currency pairs, tables were created to present the average accuracy of forecasts across all quarters of the studied period jointly with forecast horizons. In addition to the average accuracy, the tables also focused on the maximum and minimum forecasts within the forecasts of the same neural network but with different hyperparameters. Table 8 shows an example of such a table, with the average relevance for the ELM network and, in parentheses, the maximum relevance for this network forecasting the GBP/EUR currency pair.

Table 8. The average accuracy of forecasts for the ELM network and the GBP/EUR currency pair

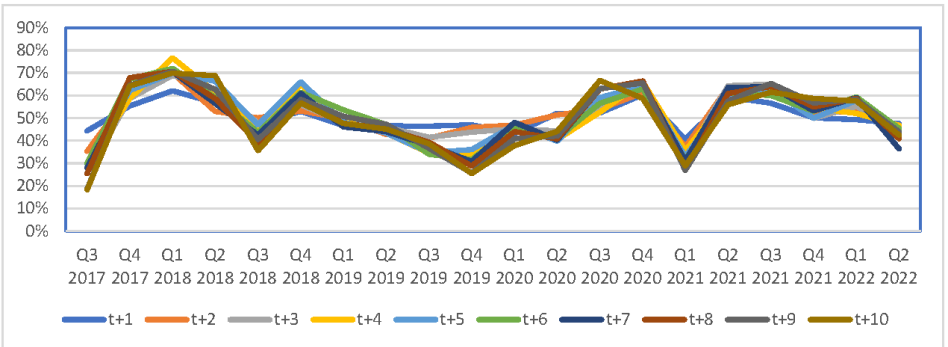
Period	*t*+1	*t*+2	*t*+3	*t*+4	*t*+5	*t*+6	*t*+7	*t*+8	*t*+9	*t*+10
	%									
Q3 2017	44.36	35.32	25.71	29.92	29.70	29.95	28.01	25.43	18.59	18.27
	(47.69)	(40.00)	(29.23)	(32.31)	(33.85)	(33.85)	(35.38)	(29.23)	(23.08)	(23.08)
Q4 2017	55.42	60.65	58.71	58.74	62.18	67.32	64.25	67.83	64.54	64.19
	(62.50)	(65.63)	(67.19)	(65.63)	(68.75)	(73.44)	(70.31)	(75.00)	(76.56)	(78.13)
Q1 2018	62.09	69.96	68.91	76.63	69.90	71.97	70.44	70.66	70.63	69.93
	(65.63)	(71.88)	(71.88)	(79.69)	(73.44)	(73.44)	(73.44)	(73.44)	(73.44)	(71.88)
Q2 2018	56.70	52.97	62.76	61.48	66.31	59.03	56.51	58.62	62.70	68.73
	(60.00)	(56.92)	(67.69)	(63.08)	(70.77)	(61.54)	(58.46)	(61.54)	(64.62)	(70.77)
Q3 2018	47.94	50.05	47.82	46.28	46.97	43.64	42.48	39.06	40.53	35.60
	(53.85)	(55.38)	(53.85)	(49.23)	(52.31)	(46.15)	(46.15)	(44.62)	(43.08)	(40.00)
Q4 2018	52.90	53.12	55.79	64.71	65.93	61.35	60.85	57.46	58.27	56.73
	(61.54)	(63.08)	(61.54)	(69.23)	(69.23)	(64.62)	(64.62)	(63.08)	(61.54)	(64.62)
Q1 2019	46.61	49.76	48.07	48.62	48.20	53.61	46.10	47.68	50.79	47.62
	(52.38)	(53.97)	(50.79)	(52.38)	(49.21)	(53.97)	(47.62)	(49.21)	(50.79)	(47.62)
Q2 2019	46.59	42.83	46.01	43.40	43.37	47.13	44.04	45.31	47.26	45.15
	(51.56)	(48.44)	(51.56)	(46.88)	(48.44)	(53.13)	(48.44)	(46.88)	(50.00)	(46.88)
Q3 2019	46.39	41.15	41.46	34.48	34.69	34.02	37.95	39.23	36.58	38.59
	(53.73)	(49.25)	(50.75)	(43.28)	(41.79)	(35.82)	(44.78)	(44.78)	(40.30)	(41.79)
Q4 2019	47.03	45.90	43.80	34.07	36.04	31.65	30.80	28.76	25.71	25.49
	(49.23)	(49.23)	(46.15)	(35.38)	(36.92)	(33.85)	(32.31)	(30.77)	(26.15)	(26.15)
Q1 2020	43.27	46.88	45.25	44.90	46.91	46.08	48.05	44.04	40.88	37.85
	(46.88)	(51.56)	(50.00)	(50.00)	(54.69)	(51.56)	(53.13)	(50.00)	(46.88)	(43.75)
Q2 2020	52.01	51.43	44.32	40.53	39.83	41.20	40.21	40.82	42.06	44.10
	(60.94)	(56.25)	(50.00)	(45.31)	(45.31)	(46.88)	(43.75)	(48.44)	(46.88)	(45.31)
Q3 2020	52.29	53.99	54.64	52.63	58.97	56.52	62.96	62.83	62.96	66.57
	(54.55)	(57.58)	(57.58)	(54.55)	(62.12)	(60.61)	(65.15)	(65.15)	(66.67)	(68.18)
Q4 2020	60.31	61.22	66.22	66.00	64.46	62.89	65.53	66.41	65.49	58.71
	(67.69)	(66.15)	(67.69)	(67.69)	(66.15)	(64.62)	(66.15)	(67.69)	(66.15)	(61.54)
Q1 2021	40.56	37.16	32.56	35.67	32.82	31.55	31.10	27.05	26.92	28.93
	(46.03)	(44.44)	(41.27)	(44.44)	(41.27)	(39.68)	(39.68)	(38.10)	(36.51)	(38.10)
Q2 2021	59.38	63.62	64.35	59.66	59.60	61.00	63.49	60.87	57.53	55.93
	(65.63)	(70.31)	(67.19)	(65.63)	(68.75)	(68.75)	(73.44)	(68.75)	(65.63)	(62.50)
Q3 2021	56.62	63.30	64.94	63.30	62.52	60.02	63.88	63.98	65.28	61.41
	(60.61)	(68.18)	(66.67)	(68.18)	(68.18)	(66.67)	(69.70)	(68.18)	(69.70)	(65.15)
Q4 2021	50.06	51.95	49.88	54.48	50.00	53.09	53.18	54.89	56.74	58.72
	(54.55)	(56.06)	(53.03)	(57.58)	(53.03)	(54.55)	(54.55)	(56.06)	(59.09)	(62.12)
Q1 2022	49.27	53.28	55.19	52.10	57.82	59.34	59.06	58.72	58.29	57.61
	(51.56)	(56.06)	(59.09)	(56.06)	(60.61)	(63.64)	(63.64)	(63.64)	(62.12)	(60.61)
Q2 2022	47.51	45.79	45.45	46.82	42.52	45.45	36.36	40.88	43.94	42.42
	(50.00)	(46.97)	(46.97)	(46.97)	(43.94)	(45.45)	(36.36)	(40.91)	(43.94)	(42.42)

Note. as in Table 2.

Source: author's work.

A graphic representation of the table above for the average accuracy of forecasts is shown in Figure 3.

Figure 3. The average forecasted accuracy for the ELM network and the GBP/EUR currency pair



Note. *t*+1, *t*+2, ..., *t*+10 – forecast horizons.
Source: authors' work.

Figure 3 shows three quarters for which the forecasts were considerably lower: Q3 2017 (the first period covered by the study), Q4 2019 (just at the start of the pandemic) and Q1 2021. The figure also shows slight differences between the accuracy of forecasts depending on the forecast horizon. For different quarters, different forecast horizons demonstrated the highest accuracy. It is difficult to determine for which horizons the forecasts were most accurate due to the fluctuations in accuracy occurring in different quarters. It should be noted that the most negligible differences in accuracy between various quarters can be observed in the forecast for the following day.

Table 9 shows the aggregate results for the period before and after the COVID-19 outbreak for the GBP/EUR currency pair predicted by the ELM network. The average accuracy of the period preceding the outbreak of COVID-19 (from Q3 2017 to Q4 2019) was lower than that observed in the period following the outbreak of the pandemic (from Q1 2020 to Q2 2022).

Table 9. The average accuracy for the period before and after the COVID-19 outbreak for the GBP/EUR currency pair predicted by the ELM network

Period in relation to COVID-19 outbreak	*t*+1	*t*+2	*t*+3	*t*+4	*t*+5	*t*+6	*t*+7	*t*+8	*t*+9	*t*+10
	%									
Before	**50.60**	**50.17**	49.90	49.83	**50.33**	49.97	48.14	48.00	47.56	47.03
After	**51.13**	**52.86**	**52.28**	**51.61**	**51.54**	**51.71**	**52.38**	**52.05**	**52.01**	**51.23**

Note. As in Table 2.
Source: author's work.

However, when making such a comparison for all neural networks and currency pairs, this situation is not a dominant one. The average accuracy for individual neural networks and currency pairs broken down into periods before and after the COVID-19 outbreak is presented in Table 10.

Table 10. Average accuracy for the period before and after the COVID-19 outbreak for all currency pairs and types of neural networks

Currency pair	NN	Period in relation to COVID-19 outbreak	*t*+1	*t*+2	*t*+3	*t*+4	*t*+5	*t*+6	*t*+7	*t*+8	*t*+9	*t*+10
			%									
USD/EUR	ELM	Before	**52.89**	**50.63**	**52.97**	**52.00**	**53.08**	**53.47**	**53.04**	**53.09**	**53.06**	**50.48**
		After	48.41	49.11	48.29	48.49	48.79	48.64	**50.18**	47.50	49.13	47.98
	MLP	Before	**54.15**	**53.92**	**54.65**	**54.87**	**55.49**	**56.38**	**57.06**	**56.89**	**56.42**	**54.51**
		After	49.83	**51.29**	**50.29**	**51.48**	**52.71**	**53.03**	**53.68**	**52.76**	**52.58**	**52.27**
	LSTM	Before	**50.52**	**53.30**	**55.67**	**57.35**	**57.19**	**58.08**	**58.83**	**61.28**	**60.82**	**60.66**
		After	**57.38**	**59.23**	**52.72**	**51.69**	**53.64**	**56.52**	**58.38**	**52.97**	**64.83**	**57.17**
CHF/EUR	ELM	Before	**51.68**	**51.49**	**51.95**	**52.13**	**52.76**	**53.30**	**51.79**	**51.23**	**50.70**	49.74
		After	47.96	48.67	46.12	46.51	45.90	45.76	44.38	45.12	45.52	44.92
	MLP	Before	49.74	48.99	45.91	48.81	47.84	47.22	46.43	45.65	46.72	44.98
		After	**50.45**	49.34	48.78	**50.76**	49.97	**51.54**	48.85	**50.17**	48.45	48.82
	LSTM	Before	49.89	**53.74**	**53.65**	**54.97**	**55.40**	**56.52**	**54.47**	**56.62**	**56.53**	**53.19**
		After	**58.52**	**56.86**	**52.34**	**58.89**	**50.35**	42.74	**60.81**	**56.37**	**66.50**	**54.85**
GBP/EUR	ELM	Before	**50.60**	**50.17**	49.90	49.83	**50.33**	49.97	48.14	48.00	47.56	47.03
		After	**51.13**	**52.86**	**52.28**	**51.61**	**51.54**	**51.71**	**52.38**	**52.05**	**52.01**	**51.23**
	MLP	Before	**50.83**	49.90	**50.51**	**51.21**	**50.50**	49.25	48.63	49.10	49.93	48.76
		After	**51.90**	**52.75**	**54.11**	**55.21**	**55.24**	**55.51**	**56.27**	**56.93**	**57.12**	**56.77**
	LSTM	Before	**51.75**	**51.99**	**54.05**	**54.92**	**55.39**	**55.74**	**56.33**	**57.60**	**58.47**	**60.05**
		After	**52.21**	**54.23**	**63.03**	**59.95**	44.57	**61.69**	**60.02**	**64.28**	**56.97**	**57.13**

Note. As in Table 2.

Source: author's work.

When comparing the average accuracy before and after the COVID-19 outbreak, the following conclusions can be drawn:

- higher accuracy was obtained before the COVID-19 outbreak period for the USD/EUR currency pair forecasted by the ELM and MLP networks and for CHF/EUR forecasted by ELM networks;
- higher accuracy was obtained after the COVID-19 outbreak for the GBP/EUR currency pair forecasted by the ELM and MLP networks, and CHF/EUR forecasted by MLP networks;
- the results for the LSTM network vary across all of the tested currency pairs.

After having compared the average accuracy of forecasts obtained by all neural networks with 49 different hyper-parameterisations, the next step was to conduct a study comparing the highest and the lowest accuracy of forecasts. The difference between these validities was obtained within the same neural network but using different combinations of hyperparameters. The results are presented in Table 11.

Table 11. Differences between the minimum and maximum forecast accuracy

Currency	NN	Period in relation to COVID-19 outbreak	*t*+1	*t*+2	*t*+3	*t*+4	*t*+5	*t*+6	*t*+7	*t*+8	*t*+9	*t*+10
			%									
USD/EUR	ELM	Before	8.82	8.83	7.90	7.90	7.90	9.30	9.46	9.45	9.44	8.97
		After	10.22	9.38	8.62	8.73	9.18	8.73	8.89	9.64	8.74	8.57
	MLP	Before	12.37	12.85	12.84	10.99	11.28	10.04	11.28	10.19	9.90	9.74
		After	15.45	13.21	14.45	14.28	14.76	14.75	14.77	15.03	14.59	15.50
	LSTM	Before	15.48	20.99	20.69	22.68	23.26	25.44	26.69	27.75	29.97	29.46
		After	32.50	31.43	27.69	23.59	26.61	21.59	23.85	26.92	22.06	25.08
CHF/EUR	ELM	Before	10.82	10.83	10.07	10.54	10.86	12.40	12.24	13.04	12.41	13.95
		After	9.44	11.19	10.47	10.78	9.69	8.93	8.62	9.08	8.02	8.94
	MLP	Before	9.73	9.27	10.20	10.36	11.91	11.44	13.14	12.67	12.22	12.53
		After	15.01	17.21	16.12	17.21	18.14	16.89	17.35	16.59	17.37	16.47
	LSTM	Before	16.56	17.62	23.62	22.87	32.33	22.92	21.97	26.94	32.48	29.07
		After	26.09	29.52	26.92	19.06	38.55	28.89	21.54	22.15	15.40	29.21
GBP/EUR	ELM	Before	9.88	10.32	9.39	7.72	8.64	7.41	8.48	7.37	7.55	8.47
		After	8.53	8.81	8.51	8.98	9.91	10.23	9.91	10.71	10.24	7.91
	MLP	Before	14.51	14.34	16.52	19.01	19.15	19.15	19.49	18.07	19.94	19.48
		After	12.11	14.48	13.26	14.61	12.00	15.22	13.07	15.06	13.82	16.30
	LSTM	Before	15.94	20.58	25.65	25.95	27.17	27.83	29.99	30.75	31.15	32.43
		After	26.72	26.98	20.31	20.78	31.13	16.98	18.46	20.77	23.97	29.05

Source: author's work.

The numbers in the table indicate by how many percentage points the most practical combination of hyperparameters achieved higher accuracy than the least effective one.

When comparing the obtained results with descriptive statistics, the following conclusions can be drawn:

- the USD/EUR exchange rate (which fell after the pandemic outbreak and began to rise in about the middle of the studied period – see Figure 1) influenced the low accuracy for the ELM and MLP in the period after the outbreak of the pandemic; the vast majority achieved a level of accuracy below 50% (which they often exceeded before the pandemic began);

- as regards the pairs with minor exchange rate fluctuations, i.e. CHF/USD for the ELM and LSTM networks, the accuracy of forecasts after the COVID-19 outbreak in most cases was higher than before the pandemic began.

Table 11 shows that there were more instances when the difference between the highest and lowest relevance within a given network was higher after the COVID-19 outbreak than before it. The abovementioned situation occurred 50 times, and the reverse only 40 times, which means that the uncertainties related to the situation in the global economy resulted in more significant differences in the accuracy of forecasts generated by neural networks. If only the USD/EUR and CHF/EUR currency pairs were to be compared, in 40 out of 60 cases the difference was higher in the years 2020–2022 than before the outbreak of COVID-19. However, for the GBP/EUR pair, such a relationship cannot be indicated. As regards the ELM network forecasting of the GBP/EUR exchange rate, higher values were obtained in the period preceding the outbreak of COVID-19 rather than in the period after in individual forecast horizons in a 6 to 4 ratio. In contrast, these relationships were reversed for the MLP and LSTM networks, where the ratio was 1 to 9 and 3 to 7, respectively. A particular analogy may be observed here to Table 7, where the only currency pair achieving higher forecast accuracy in Q1 2020 than in Q4 2019 was the GBP/EUR pair.

5. Discussion and conclusions

The literature indicates a significant impact of the COVID-19 pandemic on various aspects of social and economic life. The research presented in this article aimed to examine the impact of the pandemic on the accuracy of forecasts generated by neural networks. This examination was carried out both on a broader scope (by comparing the accuracy of forecasts from the 10 quarters preceding 2020 and the 10 quarters following it) and in a narrower scope, by comparing the accuracy of forecasts generated in the quarters at the turn of the pandemic, i.e. the last quarter of 2019 and the first quarter of 2020.

Sciences dealing with economic and market phenomena seek forecasting methods resistant to various fluctuations. The comparison of the accuracy of forecasts on a broader scope indicates slight differences between them before and after the outbreak of the pandemic. The tables with average relevance shown in this article prove the above. For some neural networks, a decrease in accuracy was noticeable after the COVID-19 outbreak, while for others an increase in accuracy was achieved. This may suggest that these results were influenced not only by the COVID-19 pandemic. In conclusion, based only on a broader scope, it can be indicated that

neural networks are an appropriate tool for making forecasts in periods of uncertainty. The study showed no significant differences between the two studied periods. The neural networks achieved satisfactory levels of accuracy compared to the period before the outbreak of the COVID-19 pandemic.

However, the conclusions differ when only extreme periods are analysed, i.e. Q4 2019 and Q1 2020. Except for the ELM network forecasting CHF/EUR, where the accuracy was very low for all horizons (below 40% and for horizons from 8 to 10 days below 20%), the network's accuracy in Q1 2020 was not very low. An accuracy below 40% can be considered very low accuracy, while above 50% satisfactory. For most results, the accuracy oscillated around 40–50%. When comparing the accuracy with Q1 2020, it should be noted that the accuracy of 40–50% in many cases is the accuracy by about 5 to a dozen or so percentage points lower than the accuracy noted in Q4 2019.

In conclusion, the research results show that neural networks are a useful tool for forecasting exchange rates. However, like many other tools, neural networks have not been immune to the impact of COVID-19. Due to fluctuations in the markets, exchange rate quotations were more challenging to forecast, resulting in decreased accuracy of forecasts in the short term (less than 5 days). Nevertheless, this impact was reduced in the longer term (more than 5 days).

References

Abedin, M. Z., Moon, M. H., Hassan, M. K., & Hajek, P. (2021). Deep learning-based exchange rate prediction during the COVID-19 pandemic. *Annals of Operations Research*. Advance online publication. https://doi.org/10.1007/s10479-021-04420-6.

Abirami, S., & Chitra, P. (2020). Energy-efficient edge based real-time healthcare support system. *Advances in Computers*, *117*(1), 339–368. https://doi.org/10.1016/bs.adcom.2019.09.007.

Akser, M. (2020). Cinema, Life and Other Viruses: The Future of Filmmaking, Film Education and Film Studies in the Age of Covid-19 Pandemic. *CINEJ Cinema Journal*, *8*(2), 1–13. https://doi.org/10.5195/cinej.2020.351.

Ashraf, B. N. (2020). Stock markets' reaction to COVID-19: Cases or fatalities?. *Research in International Business and Finance*, *54*, 1–7. https://doi.org/10.1016/j.ribaf.2020.101249.

Balbo, N., Kashnitsky, I., Melegaro, A., Meslé, F., Mills, M. C., de Valk, H., & Vono de Vilhena, D. (2020). *Demography and the Coronavirus Pandemic* (Population and Policy Brief no. 25). https://www.population-europe.eu/research/policy-briefs/demography-and-coronavirus-pandemic.

Boserup, B., McKenney, M., & Elkbuli, A. (2020). Alarming trends in US domestic violence during the COVID-19 pandemic. *The American Journal of Emergency Medicine*, *38*(12), 2753–2755. https://doi.org/10.1016/j.ajem.2020.04.077.

Cullen, W., Gulati, G., & Kelly, B. D. (2020). Mental health in the COVID-19 pandemic. *QJM: An International Journal of Medicine*, *113*(5), 311–312. https://doi.org/10.1093/qjmed/hcaa110.

Daniel, S. J. (2020). Education and the COVID-19 pandemic. *Prospects*, *49*(1–2), 91–96. https://doi.org/10.1007/s11125-020-09464-3.

Das, A. K., Mishra, D., & Das, K. (2021). Currency Exchange Prediction for Financial Stock Market: An Extensive Survey. In P. K. Mallick, A. K. Bhoi, G. Marques & V. H. C. de Albuquerque (Eds.), *Cognitive Informatics and Soft Computing* (pp. 697–709). Springer. https://doi.org/10.1007/978-981-16-1056-1_54.

Gunay, S. (2021). Comparing COVID-19 with the GFC: A shockwave analysis of currency markets. *Research in International Business and Finance*, *56*, 1–13. https://doi.org/10.1016/j.ribaf.2020.101377.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, *53*(8), 5929–5955. https://doi.org/10.1007/s10462-020-09838-1.

Jastrzębski, M., Kabziński, J., Wasiak, G., & Zawiślak, R. (2015). Ocena efektywności techniki Extreme Learning Machine (ELM) do modelowania dwuwymiarowych nieliniowości w układach napędowych. *Prace Naukowe Instytutu Maszyn, Napędów i Pomiarów Elektrycznych Politechniki Wrocławskiej. Studia i Materiały*, *71*(35), 83–94.

Kartono, A., Febriyanti, M., Tri Wahyudi, S., & Irmansyah. (2020). Predicting foreign currency exchange rates using the numerical solution of the incompressible Navier-Stokes equations. *Physica A: Statistical Mechanics and its Applications*, *560*. https://doi.org/10.1016/j.physa.2020.125191.

Li, C., Su, Z.-W., Yaqoob, T., & Sajid, Y. (2022). COVID-19 and currency market: a comparative analysis of exchange rate movement in China and USA during pandemic. *Economic Research. Ekonomska Istraživanja*, *35*(1), 2477–2492. https://doi.org/10.1080/1331677X.2021.1959368.

Markova, M. (2019). *Foreign exchange rate forecasting by artificial neural networks*. Application of Mathematics in Technical and Natural Sciences: 11th International Conference for Promoting the Application of Mathematics in Technical and Natural Sciences, Albena.

Pfefferbaum, B., & North, C. S. (2020). Mental Health and the Covid-19 Pandemic. *The New England Journal of Medicine*, *383*(6), 510–512. https://doi.org/10.1056/NEJMp2008017.

Raifu, I. A., Kumeka, T. T., & Aminu, A. (2021). Reaction of stock market returns to COVID-19 pandemic and lockdown policy: evidence from Nigerian firms stock returns. *Future Business Journal*, *7*(1), 1–16. https://doi.org/10.1186/s43093-021-00080-x.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/H0042519.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0.

Smagulova, K., & James, A. P. (2019). A survey on LSTM memristive neural network architectures and applications. *The European Physical Journal Special Topics*, *228*(10), 2313–2324. https://doi.org/10.1140/epjst/e2019-900046-x.

Sun, S., Wei, Y., & Wang, S. (2018). *AdaBoost-LSTM ensemble learning for financial time series forecasting*. International Conference on Computational Science, Wuxi.

Usher, K., Durkin, J., & Bhullar, N. (2020). The COVID-19 pandemic and mental health impacts. *International journal of mental health nursing*, *29*(3), 315–318. https://doi.org/10.1111/inm.12726.

Wu, Y., & Gao, J. (2018). AdaBoost-based long short-term memory ensemble learning approach for financial time series forecasting. *Current Science*, *115*(1), 159–165. https://doi.org/10.18520/cs/v115/i1/159-165.

Zeren, F., & Hizarci, A. E. (2020). The impact of COVID-19 coronavirus on stock markets: evidence from selected countries. *Muhasebe ve Finans İncelemeleri Dergisi*, *3*(1), 78–84. https://doi.org/10.32951/mufider.706159.

Zhang, W., & Hamori, S. (2021). Crude oil market and stock markets during the COVID-19 pandemic: Evidence from the US, Japan, and Germany. *International Review of Financial Analysis*, *74*, 1–13. https://doi.org/10.1016/j.irfa.2021.101702.