İSTANBUL UNIVERSITY PRESS

RESEARCH ARTICLE

A Comparative Assessment of Frequentist Forecasting Models: Evidence from the S&P 500 Pharmaceuticals Index

Christian MUNEZA[1] , Asad Ul İslam KHAN[1] , Waqar BADSHAH[2]

ABSTRACT

This paper compares three forecasting methods, the autoregressive integrated moving average (ARIMA), generalized autoregressive conditional heteroscedasticity (GARCH), and neural network autoregression (NNAR) methods, using the S&P 500 Pharmaceuticals Index. The objective is to identify the most accurate model based on the mean average forecasting error (MAFE). The results consistently show the NNAR model to outperform ARIMA and GARCH and to exhibit a significantly lower MAFE. The existing literature presents conflicting findings on forecasting model accuracy for stock indexes. While studies have explored various models, no universally applicable model exists. Therefore, a comparative analysis is crucial. The methodology includes data collection and cleaning, exploratory analysis, and model building. The daily closing prices of pharmaceutical stocks from the S&P 500 serve as the dataset. The exploratory analysis reveals an upward trend and increasing heteroscedasticity in the pharmaceuticals index, with the unit root tests confirming non-stationarity. To address this, the dataset has been transformed into stationary returns using logarithmic and differencing techniques. Model building involves splitting the dataset into training and test sets. The training set determines the best-fit models for each method. The models are then compared using MAFE on the test set, with the model possessing the lowest MAFE being considered the best. The findings provide insights into model accuracy for pharmaceutical industry indexes, aiding investor predictions, with the comparative analysis emphasizing tailored forecasting models for specific indexes and datasets.

Keywords: Forecasting Accuracy, Pharmaceutical Industry Indexes, S&P 500, ARIMA, GARCH, NNAR, Comparative Analysis

Introduction

For the last two decades and especially after the 2008 global financial crisis, financial markets have grown to represent a sizable share of national incomes (Ross, 2021). To keep up with the trend and make sense of vast amounts of available data, financial economists have devised methods ranging from algorithmic trading to forecasting methods in order to anticipate the market and maximize profits (Organisation for Economic Cooperation and Development (OECD, 2021). While these methods have grown in relevance as toolkits investors use to gauge evidence-backed predictors of markets, they are still far from perfect, with the case in point being their widespread association with the 2008 financial crisis due to how overreliance on them blinded many from seeing what was about to come. The past failures of financial forecasting methods and the need to find ways they can be more accurate are what fuel this academic exploration.

Despite the increase in academic research on financial forecasting models, most leading academic papers have focused on forecasting aggregate market indexes, such as the S&P 500 (Niaki & Hoseinzade, 2013), Dow Jones (Nasr Ben & Lux, 2016), Nasdaq Composite (Sunarya, 2019), Financial Times Stock Exchange (FTSE 100; Niu et al., 2020), National Stock Exchange of India (NIFTY 50; Mahajan et al., 2022), and the Shanghai Stock Exchange (SSE) Composite (Lin, 2018) indices, among others. This has contributed to the increased understanding of the behavior of the market as an aggregate, but the current scholarship remains insufficient at helping understand the behavior of particular industries. Though the gap is present in many industries, it is at its greatest in pharmaceutical and healthcare-related stock indexes (Harris, 2018).

After identifying the current gap in the literature as it relates to forecasting industry-specific indexes, the study aims to contribute to addressing the gap by testing forecasting accuracies on the pharmaceutical industry indexes of the S&P 500. This paper has significance for both academics and policies. Academic-wise, health stocks remain the least studied when compared to other stocks. Despite the huge data potential, the field as a whole remains the least explored, especially in terms of the use of data analytics (Harris, 2018). The significances policy-wise are also paramount. With the COVID-19 pandemic, climate change, increasing aging population, rise of noncommunicable diseases, research on longevity and better treatments, and the other host of global disease burdens caused by urbanization and lifestyle changes, the public sector will clearly face a need to invest more in pharmaceutical R&D and improved healthcare solutions (Simpkin et al., 2019). The public sector alone definitely won't meet the funding needed to address the emerging global disease burden, hence the clear need for increasing the role of the private sector in healthcare financing and funding (Simpkin et al., 2019). Both the individual and institutional investors that are currently investing in such

industries as oil and gas, infotech, automobiles and manufacturing, real estate, and others will need to familiarize themselves with and increase their share in healthcare investments. For this to happen, more study is needed to show investors that investment in healthcare is not only safe but also profitable.

The next sections of the study will be divided as follows: Literature Review, Methodology, Results and Discussion, and Conclusion and Recommendation. The Literature Review will provide a window into the works and findings of preliminary studies on the subject matter, while the Methodology section will detail the processes this study follows, such as data collection, data examination, and forecasting. The Results and Discussion section will present the outcome of each stage and provide commentary and interpretation, and the Conclusion and Recommendation section will provide the final takeaways from the study.

Literature Review

Sizable academic research is found to have compared the accuracy of forecasting methods on several indexes, usually with differing conclusions. Mahajan and Thakan's (2022) study "Modeling and Forecasting the Volatility of NIFTY 50 Using GARCH and RNN Models" sought to evaluate the forecasting accuracy of the generalized autoregressive conditional heteroscedasticity (GARCH) and neural network autoregression (NNAR) families of models. Motivated by the great volatilities within the Indian stock market such as the 2000s tech advancements that led to the boom of the Indian stock market and the crash that followed, they analyzed NIFTY 50 to identify the behavior of the Indian market's volatility and then evaluated the forecasting abilities of the above-mentioned models. They concluded that NIFTY 50 volatility is asymmetric and concluded the exponential GARCH, or EGARCH (1,1), and threshold autoregressive conditional heteroscedasticity, or TARCH (1,1), models to be the best at forecasting.

In a hybrid stock price index forecasting model based on variational mode decomposition (VMD) and long short-term memory (LSTM) network, Niu and Xu (2020) introduced a new hybrid model using VMD-LSTM to study the FTSE 100 Index. Their study is advantageous in that VMD decomposes the original complex series into a limited number of series with simpler fluctuation modes, thus overcoming the shortcomings of mode mixing found in the typically used empirical decomposition method (EDM). Another advantage is that LSTM filters out the critical previous information, thus making it better for financial time series forecasting than traditional recurrent neural networks. They concluded their VMD-LSTM hybrid model to be a better forecaster than single models.

Sunarya's (2019) study "Modelling and Forecasting Stock Market Volatility of Nasdaq Composite Index" evaluates the best models for both autoregressive integrated moving average

(ARIMA) and GARCH regarding the Nasdaq returns from March 1971-April 2019. They employed the standard data analytics methodology of data cleaning, manipulation, and model estimation. For ARIMA, they find ARIMA (8,0,6) model to be the best due to its lowest Akaike information criterion (AIC) value. They also determine the ARIMA-GARCH model combination and the best model for this emerges as ARIMA (8,0,6)-EGARCH (1,1). while modelling and forecasting the stock market volatility of SSE Composite Index using GARCH models, Lin (2018) examined the econometric features of the Shangai Stock Exchange (SSE) Composite Index and compared the forecasting ability of the GARCH family of models. Lin's results found SSE to have significant properties regarding time variance and clustering due to rapid information dissemination, fast capital flow, and undulating prices. Due to these phenomena, Lin's forecasting experiments concluded the EGARCH (1,1) model to outperform the GARCH (1,1) and TARCH (1,1) models.

Yadav and Sharma's (2018) study "Statistical Analysis and Forecasting Models for Stock Market" evaluated the accuracy of ARIMA, exponential smoothing, naive, seasonal naive, neural network, mean, and BoxCox transformation forecasting methods to predict the Bombay Stock Exchange (BSE) SENSEX opening, high, low, and closing prices from January 1997-January 2016. Their methodology utilized standard data analytical techniques for data cleaning and manipulation, as well as model estimation. They set the mean error as the accuracy criteria, and the exponential smoothing and neural network models emerged as the best ones. Islam and Nguyen's (2020) study "Comparison of Financial Models for Stock Price Prediction" evaluated the accuracy of ARIMA, artificial neural network (ANN), and stochastic process-geometric Brownian motion for forecasting the S&P 500 using daily adjusted closing prices from April 1, 2015-December 31, 2019. They set the standardized residuals as the accuracy criterion, with the ARIMA and stochastic process-geometric Brownian motion models emerging as the best ones for predicting short-term next-day prices. Their findings agree with those from Merh et al. (2010) on ARIMA predicting stock prices better than ANN but contradict those from Khashei and Bijari (2010), who had concluded ARIMA to be no better than ANN.

Sharaff and Choudhary's (2018) study "Comparative Analysis of Various Stock Prediction Techniques" evaluated the accuracy of ARIMA, ANN, Holt-Winters, and NNARs for forecasting the S&P Bombay Stock Exchange using monthly closing prices from 2007-2012. Their methodology also involved the standard data analytics process of visualization, stationarizing, finding optimal parameters for models, and making predictions. They set the mean absolute percentage error (MAPE) as the accuracy criterion, with ANN emerging as the best model. Niaki and Hoseinzade's study (2013) "Forecasting S&P 500 Index Using Artificial Neural Networks and Design of Experiments" compared the predictive ability of ANN to traditional logit models. Their study included 27 financial and economic variables that tend to

influence the S&P500 movements and compared ANN and logit in terms of how they respond to these variables. They concluded ANN to be better at integrating influential variables and forecasting the index compared to the traditional logit model. Overall, the literature suggests that no decisive one-size-fits-all model exists that can be applied to stock predictions and that specific comparative scrutiny should be applied to each different index and dataset.

Methodology

The paper uses the ARIMA, GARCH, and NNAR models, and its methodology consists of four steps: data collection, data cleaning, exploratory analysis, and model building and forecasting in order to arrive at the best model. This section lays out the mathematical notions for the models, as well as the commentary on each of the methodological steps.

ARIMA Model

The ARIMA model is based on the following equation:

$$y_t = c + \emptyset 1 y_t - 1 + \emptyset 2 y_t - 2 \ldots + \emptyset p y_t - p + \theta 1 \varepsilon t - 1 + \theta 2 \varepsilon t - 2 \ldots \theta q \varepsilon t - q + \mathrm{e}_t \quad \textit{Eq. 1}$$

where y_t is the variable explained at time t; c is the constant; $\emptyset(i=1,2..p)$ and $\theta(j=1,2,\ldots q)$ are the model parameters; p and q are integers with p representing the autoregressive (AR) part and q representing the moving average (MA) part; and e_t is the error term.

GARCH Model

The GARCH model is derived from the following equation:

$$\sigma t = \sqrt{\omega + \sum_{i=1}^{p} \alpha^2 t - i + \sum_{i=1}^{q} \beta \sigma^2 t - i} \qquad \textit{Eq. 2}$$

where σt is the conditional standard deviation and its past values $\sigma t 1,2,\ldots$ are fed back into the process. $\sum_{i=1}^{p} \alpha^2 t - i$ represents the AR part of the model, and $\sum_{i=1}^{q} \beta \sigma^2 t - i$ represents the conditional heteroscedastic part of the model.

NNAR Model

The NNAR model comes from Eq. 3 as follows:

$$y_t = w_0 + \sum_{j=1}^{q} w_j . g \left(w_{0,j} + \sum_{i=1}^{p} w_{ij} . y_{t-1} \right) + \mathrm{e}_t \qquad \textit{Eq. 3}$$

where w_j $(j = 0,1,2,3..., q)$ and w_{ij} $(i = 0, 1, 2..., p; j = 0, 1, 2, ..., q)$ are the connection weights or model parameters, p is the number of input nodes, and q is the number of hidden nodes.

Methodology Steps

i.   Data Collection. The paper collected daily Pharmaceutical S&P500 closing prices from January 4, 2010-December 31, 2019 from Market Watch.

ii.   Data Cleaning. This process involves transforming the dataset into a time-series format readable by the program R by checking for the presence or absence of missing values and labeling columns appropriately.

iii. Exploratory Data Analysis. This process involves first plotting the dataset to see elements such as trends, seasonality, heteroscedasticity, and stationarity, then running stationarity tests, and finally applying logarithm and differencing techniques to transform the closing prices into stationary returns that will be used to build the forecasting models. The resultant dataset will be called "pharmareturns" and will be used in the following modeling stages based on the following formula:

$z_i = y_t - y_{t-1}$ *Eq. 4*

where z_i represents the returns, y_t represents the closing price at time t, and y_{t-1} represents the closing price at time t - 1.

iv. Model Building and Forecasting: This step involves the process of determining the best fit model for each of the ARIMA, GARCH, and NNAR models. The dataset on the returns (adjusted closing prices) will be divided into a training set constituting 70% of the dataset and a test set constituting the remaining 30%. The best fit model for each forecasting method will be found using the training set. Then the best fit model will be applied to the forecast, with its results compared to the test set. The model with the lowest mean average forecasting error (*MAFE*) will emerge as the best forecasting method.

Results

Exploratory Data Analysis

The dataset used in the study contains the S&P 500 Pharmaceuticals Industry daily closing prices from January 4, 2010-December 31, 2019. The S&P 500 Pharmaceuticals Index was used for a number of reasons, such as its focus on industry leaders, relevance in the investment space, and benchmark comparison. Specifically, the 20 largest pharmaceutical companies tracked by the S&P 500 account for 78.7% of the global prescription market and are therefore a representative sample for studying global trends in pharmaceutical stock indexes (Mikulic,

2022). After the processes of cleaning and transforming the data into a time series format, the dataset was analyzed for trend, seasonality, heteroscedasticity, and stationarity.

Step 1: Plotting the Dataset

The data have been plotted to observe its long-run behavior (Figure 1). As can be seen from Figure 1, a clear upward trend is present, as well as heteroscedasticity that continuously increases over time.
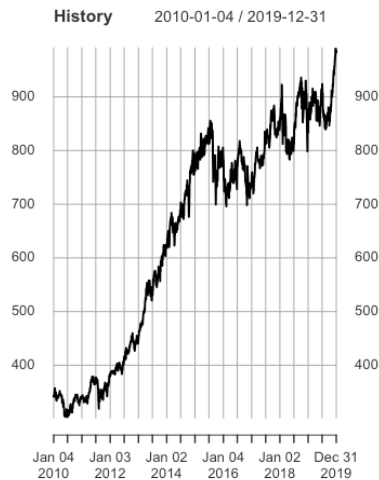
Figure 1: *S&P 500 Pharmaceutical Closing Prices from 2010 to 2019.*

Step 2: Unit Root Tests

Next, unit root tests were run to determine the stationarity of the data (see Table 1). A *p*-value of 0.5053 was obtained for the augmented Dickey-Fuller (ADF) test, indicating the null test is rejected, and the data are concluded to be nonstationary. For both the Kwiatkowski, Phillips, Schmidt, and Shin (KPSS) level stationarity and trend stationarity tests, *p*-values of 0.01 were obtained, indicating the null test for KPSS to again be rejected at a 5% significance and the dataset to be similarly concluded as nonstationary. After this analysis, the need is seen to exist for transforming the dataset into stationary data before applying the forecasting models to it.

Table 1. *Unit Root Test Results.*

TEST	Null Hypothesis	P-value
ADF Test	Unit Root	0.5053
KPSS Test	Stationary	0.01

Transforming the Dataset to Stationary Data

Due to the presence of heteroscedasticity and unit roots, logarithms will first be applied to the data to contain the heteroscedasticity and then it will be differentiated to remove the trend. The resulting dataset will represent the historical returns and is expected to exhibit white noise behavior. The resultant dataset will henceforth be referred to as "pharmareturns" and will be used for the rest of the modeling.

Figure 2: *S&P 500 Pharmaceutical Returns from 2010 to 2019.*

The graphs in Figure 2 compare the pretreatment and posttreatment datasets. While the pretreatment dataset exhibits unit roots, the posttreatment dataset exhibits white noise, which is in agreement with the financial literature on the historical nature of returns. The ADF and KPSS tests are rerun on the treated dataset to check for stationarity (see Table 2).

Table 2. *Unit Root Test Results on Returns.*

TEST	Null Hypothesis	*p*-value
ADF Test	Unit Root	0.01
KPSS Test	Stationary	0.1

The ADF test confirms the expected stationarity of the pharmareturns dataset at a 5% significance of p = 0.01. The KPSS tests also resulted in *p* = 0.1, thus the null hypothesis is no longer rejected and the pharmareturns dataset is concluded to indeed be stationary.

Model Building and Forecasting

This section determines the best model for each method (i.e., the best ARIMA model, best GARCH model, best NNAR model). The pharmareturns dataset is first divided into a training set containing 70% of the data and a test set containing 30% of the data. Once the best model for each method has been determined, the best forecasting method will then be identified.

Determining the Best Models

ARIMA

Different p and q levels for the ARIMA(p, 0, q) model were experimented with in order to see which model has the least error based on the AIC value. This stage of the study experiments with p levels between 0-5 and q levels between 0-5. Table 3 lists the AIC values for some of the models.

Table 3. *ARIMA Models Estimation Results.*

ARIMA Model	Mean Specification	AIC Value
ARIMA (2, 0, 2)	non-zero mean	-11,182.78
ARIMA (0, 0, 0)	non-zero mean	-11,182.66
ARIMA (1, 0, 0)	non-zero mean	-11,182.94
ARIMA (0, 0, 1)	non-zero mean	-11,183.19
ARIMA (0, 0, 0	zero mean	-11,181.15
ARIMA (1, 0, 1)	non-zero mean	-11,183.89
ARIMA (2, 0, 1)*	non-zero mean*	-11,184.25*
ARIMA (2, 0, 0)	non-zero mean	-11,181.56
ARIMA (3, 0, 1)	non-zero mean	-11,181.05
* best model		

As Table 3 shows, the ARIMA (2, 0, 1) model has been identified as the best model due to having the lowest AIC value. The models were applied to the forecast and compared with the forecasts from the test set, with Table 4 showing ARIMA (2, 0, 1) to have been determined as the best forecasting ARIMA model with the lowest *MAFE* value of 0.0587138.

Table 4. *ARIMA Model Selection.*

Set	ME	RMSE	MAE	MFE	MAFE
Training set	2.016731×10^{-6}	0.007305894	0.06173646	0.0654321	0.6745051
Test Set	-8.167452×10^{-5}	0.009037684	0.006500222	0.05746563	0.0587138
ME = margin of error; RMSE = root mean square error; MAE = mean absolute error; MFE = maximum favorable excursion					

GARCH

Both the mean equation and the variance equation were found for the pharmareturns dataset. Table 5 shows the GARCH (1, 1) model with the mean model used being ARIMA (2, 0, 1). The model has a *MAFE* value of 0.000328.

Table 5. *GARCH Model Selection.*

GARCH Best Model	AIC	*MAFE*
GARCH Model (1, 1) Mean Model (2, 0, 1)	-6.5107	0.000328

NNAR

Similarly, different levels for *p* and *q* were experimented with for the NNAR model, with *p* representing the number of lagged values and *q* representing the number of hidden layers. Table 6 shows the NNAR (10, 6) model was obtained as the one with the lowest *MAFE*.

Table 6. *NNAR Model Selection.*

NNAR Best Model	*MAFE*
NNAR(10, 6)	8.147859×10^{-9}

The NNAR(10, 6) model was applied to the forecast and compared to the forecasts from the test set, and we obtain a mean average forecasting error of 8.147859e-09.

Best Forecasting Model Selection

This section determines the best forecasting model for the S&P 500 Pharmaceutical Index based on *MAFE*, with the results shown in Table 7.

Table 7. *Best Model Selection.*

Forecasting Model	*MAFE*	*RMSE*	*MAE*
ARIMA(2, 0, 1)	0.0587138	0.009037684	0.006500222
GARCH(2, 1) & (1, 1)	0.000328	0.025586	0.0185646
NNAR(10, 6)	8.147859×10^{-9}	7.951×10^{-5}*	6.546×10^{-5}*
*best model			

Table 7 shows the NNAR model to have conclusively emerged as the best model for the dataset, possessing a significantly lower *MAFE* than the other two models and also based on *RMSE* and *MAE*. The final conclusion is based on *MAFE* due to its wide acceptance in the academic literature as a measure of forecasting accuracy (Tofallis, 2017).

Conclusion and Recommendations

The aim of the study has been to compare some of the most commonly used forecasting models (i.e., ARIMA, GARCH, and NNAR) and to determine the best one regarding a dataset derived from the S&P 500 Pharmaceuticals Index. *MAFE* was used as the accuracy metric for determining the best forecasting model. The study involved rigorous processes for data cleaning, exploratory data analysis, model building, and best model selection, with the NNAR

model being determined as the best one due to it having the lowest *MAFE* value, which is widely used as a measure of forecasting accuracy. The study recommends NNAR be used when forecasting the S&P 500 Pharmaceuticals index as it forecasts are more reliable compared to the other models examined in this study.

The study has been limited to normal forecasting methods and did not leverage machine learning tools such as supervised and unsupervised learning or more robust cross-validation techniques that use multiple training and testing datasets. As such, future studies can use machine learning tools and robust cross-validation techniques in order to obtain results with higher confidence. Moreover, future research should emphasize on exploring less-studied indexes, such as those from emerging and developing countries.

References

Harris J. (2018). *The opportunities and challenges of data analytics in health care.* Brookings. https://www.brookings. edu/research/the-opportunities-and-challenges-of-data-analytics-in-health-care/

Islam, M. R., & Nguyen, N. (2020). Comparison of Financial Models for Stock Price Prediction. *Journal of Risk and Financial Management, 13*(8), 181. https://doi.org/10.3390/JRFM13080181

Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with Applications*, *37*(1), 479–489. https://doi.org/10.1016/J.ESWA.2009.05.044

Lin, Z. (2018). Modelling and forecasting the stock market volatility of SSE Composite Index using GARCH models. *Future Generation Computer Systems*, *79*, 960–972. https://doi.org/10.1016/J.FUTURE.2017.08.033

Mahajan, V., Thakan, S., & Malik, A. (2022). Modeling and Forecasting the Volatility of NIFTY 50 Using GARCH and RNN Models. *Economies, 10*(5), 102. https://doi.org/10.3390/ECONOMIES10050102

Merh, N., Prakash Saxena, V., & Raj Pardasani, K. (2010). *A comparison between Hybrid Approaches of ANN and ARIMA for Indian Stock Trend Forecasting Computational Aspects of I-Function View project*. https://www. researchgate.net/publication/45602108

Mikulic, M. (2022). *Market share top pharma companies Rx drugs sales globally 2026 | Statista*. Statistica. https:// www.statista.com/statistics/309425/prescription-drugs-market-shares-by-top-companies-globally/

Niaki, S. T. A., & Hoseinzade, S. (2013). Forecasting S&P 500 index using artificial neural networks and design of experiments. *Journal of Industrial Engineering International*, *9*(1). https://doi.org/10.1186/2251-712X-9-1

Niu, H., Xu, K., & Wang, W. (2020). A hybrid stock price index forecasting model based on variational mode decomposition and LSTM network. *Applied Intelligence*, *50*(12), 4296–4309. https://doi.org/10.1007/S10489-020-01814-0

OECD. (2021). Sustainable and Resilient Finance OECD Business and Finance Outlook 2020 9HSTCQE*diefgj+ Artificial Intelligence, Machine Learning and Big Data in Finance Opportunities, Challenges and Implications for Policy Makers. *OECD Journal*.

Sean Ross. (2021). *What Percentage of the Global Economy Is the Financial Services Sector?* Investopedia. https://www.investopedia.com/ask/answers/030515/what-percentage-global-economy-comprised-financial-services-sector.asp

Sharaff, A., & Choudhary, M. (2018). Comparative Analysis of Various Stock Prediction Techniques. *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018*, 735–738. https://doi.org/10.1109/ICOEI.2018.8553825

Simpkin, V., Namubiru-Mwaura, E., Clarke, L., & Mossialos, E. (2019). Investing in health R&D: where we are, what limits us, and how to make progress in Africa. *BMJ Global Health*, *4*(2). https://doi.org/10.1136/BMJGH-2018-001047

Sunarya, I. W. (2019). Modelling and Forecasting Stock Market Volatility of Nasdaq Composite Index. *Economic and Accounting Journal*, *2*(3), 181–189. https://doi.org/10.32493/EAJ.V2I3.Y2019.P181-189

Tofallis, C. (2017). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, *66*(8), 1352–1362. https://doi.org/10.1057/JORS.2014.103

Yadav, S., & Sharma, K. P. (2018). Statistical Analysis and Forecasting Models for Stock Market. *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 117–121. https://doi.org/10.1109/ICSCCC.2018.8703324