JEL: C22, C52, Q11

Tetiana Kmytiuk[1], Ginta Majore[2], Tetiana Bilyk[1]

[1]Kyiv National Economic University named after Vadym Hetman
[2]Vidzeme University of Applied Sciences
[1]Ukraine
[2]Latvia

TIME SERIES FORECASTING OF PRICE OF THE AGRICULTURAL PRODUCTS USING DATA SCIENCE

Purpose. *The purpose of our article is to research and forecast prices for agricultural products using the example of potato prices based on the most effective models using data science techniques.*

Methodology / approach. *Various forecasting models are explored, starting from baseline models like decomposition and exponential smoothing models to more advanced techniques such as ARIMA, SARIMA, as well as deep learning models including neural network. The data is split into training and testing sets, and models are validated using cross-validation techniques and optimised through hyperparameter tuning. Model performance is evaluated using metrics such as MAE, MSE, RMSE, and MAPE. The selected model is then used to generate future price forecasts, with uncertainty quantified through confidence intervals.*

Results. *The study successfully applied advanced data science techniques to forecast potato prices, leveraging a range of effective models. By analysing historical price data and using various forecasting methods, the research identified the most accurate models for predicting future price trends. The results demonstrate that the selected models can provide reliable forecasts. In particular, the results showed that the* $SARIMA\ (1,0,0)(0,1,1)[12]$ *model could achieve good forecast results when applied to real problems and, thus, can be effectively used for forecasting tasks especially considering seasonality. In addition, it should be noted that the* $ETS\ (M, A_d, A)$ *model has a higher prediction accuracy at the time intervals closest to the original data. The obtained results support using both models simultaneously for forecasting, which can compensate for the shortcomings of each of them. The models can be used separately, to more accurately predict the values for the required period, or a combination of them is also possible.*

Originality / scientific novelty. *The study's originality lies in development of methods for effectively accounting for seasonality in agricultural price data, such as using seasonal decomposition techniques or more advanced techniques that combine statistical and data science approaches. The novelty implies the implementation of real-time data processing and forecasting system allows for the timely prediction of price changes, enabling stakeholders to make more informed decisions.*

Practical value / implications. *Forecasting potato prices holds significant practical value for various stakeholders. For farmers, accurate forecasts enable informed decisions on the optimal times to plant, harvest, and sell their crops, thereby optimising their profits. In the supply chain, distributors and retailers can use these forecasts to manage inventory more effectively and plan contracts, reducing waste and avoiding shortages. Policymakers benefit from forecasts by anticipating market fluctuations and stabilising prices, which supports both consumers and producers. For consumers, stable pricing ensures better budgeting and helps avoid sudden price spikes, making essential foods more affordable. Overall, accurate price forecasting enhances market efficiency by reducing uncertainty and aiding investors in managing risk.*

Key words: *agriculture, price, seasonality, time series model, forecasting, neural network.*

1. INTRODUCTION

Price is a central element in the economic relations between producers and consumers in agriculture. Price reflects changes in the agricultural market and involves the economic interests of all participants in this sector. The competitiveness of products on the market and the final economic results for agricultural producers depend on the price level. The issue of price forecasting is one of the key issues today. With adequate price forecasting, a producer of agricultural products will be able to make better financial decisions, i.e. how to maximise their profits. Since the prices of agricultural products change over time, these changes can be recorded as a successive time series of data. A time series is a sequence of events usually observed at equally spaced time intervals. A characteristic feature of a time series is that the level of indicators in the next period largely depends on their level in the past. Mathematical models can be used to predict a time series. By constructing such mathematical models, a researcher can explain the behaviour of the time series and make a forecast for future periods. There are many models for forecasting, however, when choosing the appropriate model for forecasting the price of agricultural products, it is necessary to take into account the fact that a time series may have hidden characteristics such as lag of variables, autocorrelation, nonlinearity, nonstationarity, and seasonality. In such a case, it is impossible to use the traditional statistical and mathematical tools related to a time series, as it requires considerable experience and skills to select the appropriate type of model for this data set. To do this, it is important to conduct a full analysis of the effectiveness of the forecasting models and their suitability to describe a particular dynamic process. This problem is complex and requires significant mathematical calculations, which is impossible without the use of machine learning tools. Therefore, the question of finding the optimal forecasting model for a particular process, namely the pricing of agricultural products (potatoes) becomes relevant. A wide range of economic and mathematical modelling tools can be used to predict seasonal processes. In this study, we are considering and evaluating the following forecasting methods and models using time series that contain seasonality which are most suitable for the agricultural sector. The purpose of our article is to research and forecast prices for agricultural products using the example of potato prices based on the most effective models using data science techniques.

The structure of the paper is as follows: the introductory section focuses on the importance of the theme and justifies why this particular theme was chosen. Section 2 presents a literature review. Section 3 describes the methodological procedures of the building of famous time series methods and models, which can be used for phenomena and processes analysis in the agricultural sphere. Section 4 shows results, in particular, the construction of the econometric models and their forecasts, and statistical estimation based on Data Science tool and discusses the research findings, and Section 5 deals with the main conclusions of the paper.

2. LITERATURE REVIEW

A significant amount of research in the field of economic-mathematical modelling

is devoted to theoretical and practical methods of forecasting. The number of models and methods of forecasting is constantly increasing: researchers are developing new approaches and improving existing ones. Today, there are numerous types of time series forecasting models that are described in many scientific studies. According to Makridakis et al. (1998) a time series may be decomposed into four components: trend, cyclic, seasonality, and an irregular component. Mbuli et al. (2020) presents a comprehensive literature review on the application of decomposition methods of time series forecasting in his article. The precise procedure for estimating the trend, seasonal factor, and cycle terms is described by O'Connell et al. (1993). Vandeput (2021) presented the basic idea of simple exponential smoothing based on assumptions that the future will be more or less the same as the past. Among the existing forecasting methods that use exponential smoothing, the most well-known are Holt's linear method (Holt, 1957), Multiplicative Holt-Winters' method (Winters, 1960), and Brown's model (Brown, 1959). Hyndman and Khandakar (2008) discuss a family of 60 different exponential smoothing models and provide a new state-space approach to evaluate the likelihood function.

Wei (2006), and Kotu and Deshpande (2019), and Brockwell and Davis (1991) present the basic mathematical and statistical methods and models of time series analysis designed to identify their structure and forecast. A detailed discussion of the specific methods and models of their practical application for seasonal time series forecasting can be found in Gardner (2006) and Hyndman et al. (2008), Box and Jenkins (1976) proposed a quite successful variation of the Autoregressive Integrated Moving Average Model (ARIMA) model called the seasonal ARIMA (SARIMA) model. A hybrid model of ARIMA-PNN (probabilistic neural network) is presented in Khashei et al. (2012).

Recently, artificial neural networks (ANNs) are attracting more and more attention in the field of time series forecasting. A distinctive feature of ANN when applied to time series forecasting problems is its inherent ability to perform nonlinear modelling, without any assumptions about the statistical distribution, which is then followed by observations. The corresponding model is adaptively formed based on the given data. Over the past few years, a significant amount of research has been done on the use of neural networks for modelling and forecasting time series. In related literature Zhang et al. (1998), Hoptroff (1993) and Gately (1995) present different models of ANN forecasting. A neural network of time series prediction based on multilayer perceptron is described by Shiblee et al. (2008) and Rudenko et al. (2019). In Kmytiuk et al. (2021) the authors present popular recurrent neural networks. Conway et al. (1998) presented another widely used type of MLP that makes delayed time series predictions with neural networks based on a prediction of solar activity and demonstrates the tendency of neural networks to generate delayed predictions of specific features in the data.

The last stage of time series analysis can be forecasting its future (extrapolation) and determining the accuracy of this forecast based on the selected model. Nevertheless, despite such a wide variety of forecasting approaches, to date there is no

single one that shows equally high-quality results for a particular process. Another significant problem is choosing the best, in a sense, from the class of models. It often happens that several models produce suitable results. The ambiguity of the choice of a model can be observed both at the stage of selection of the deterministic component of the series, and when deciding the structure of a number of residues. Therefore, quite frequently, numerous forecasts are made using different models.

The prior empirical review allowed forming the following hypotheses, which are confirmed in the article:

Hypothesis H1: Historical Price Data Impact: The inclusion of extensive historical price data in time series models will enhance the precision of future price predictions.

Hypothesis H2: Seasonal Trends: Agricultural products often exhibit strong seasonal trends due to planting and harvest cycles, climatic conditions, and seasonal demand variations. Accurately modelling these trends is essential for precise price forecasting.

Hypothesis H3: Seasonal Decomposition: Techniques such as Seasonal Decomposition of Time Series (STL) can be used to separate the seasonal component from the overall trend and noise, providing clearer insights into the underlying patterns.

Hypothesis H4: Model Comparison: Advanced time series models like ARIMA/SARIMA and Neural Network will outperform traditional forecasting methods in terms of accuracy and reliability.

Hypothesis H5: Data Science Techniques: The application of data science techniques, such as machine learning and time series analysis, can significantly improve the accuracy of forecasting agricultural product prices, enabling better decision-making and planning for stakeholders in the agricultural sector.

3. METHODOLOGY

Let $Y_t, t_i \in T, (i = 1..n)$ be the set of observations of average prices for agricultural products (potatoes), which is obtained sequentially in time by measurements, and T is the set of samples of time at which observations were made. Observations are interpreted as the implementation of the stochastic process $\{Y_t: t \in T\}$ for time $T \subset R$. Moreover, the interval between neighboring values is the same, i.e. $\Delta t = t_{(i+1)} - t_i = const$. After ordering the set of values of the random variable y_i, which is observed at successive moments of time t_1, t_2 ... t_n, we obtain the sequence $y(t_1), y(t_2), \dots, y(t_n)$ which is called the time series and is denoted by $\{Y_{t_n}\}$.

There are three main purposes related to research into and the analysis of time series:

1. Description of the change of the studied feature over time and the identification of the properties of the studied series;
2. Determining the nature of the studied series;
3. Predicting the values of the studied feature for future moments of time.

Observational based forecasting is an important task for effective planning and

management of many economic processes. Many socio-economic indicators are characterised by the presence of cyclical intra-annual fluctuations, which are called seasonal. It is important that the forecasting model capture this pattern when a time series has a seasonal pattern.

Time series with an interval of less than a year (month, quarter), as a rule, contain seasonality. The seasonal component has a period m: ($m = 12$ for a number of monthly data; $m = 4$ for a number of quarterly data). In addition, it is known that m is a multiple of n (number of data) i.e., k is an integer. Obviously, if m is the number of months or quarters in a year, then k is the number of years represented in the time series $\{Y_{t_n}\}$.

Seasonal forecasting models are based on their non-seasonal counterparts, supplemented by means of displaying seasonal fluctuations. Seasonal models are able to reflect a relatively constant seasonal wave as well as dynamically change depending on the trend.

3.1. Seasonal Decomposition Method. Time series decomposition is a forecasting technique that divides or decomposes historical data into various components and uses them to create a forecast that is more accurate than a simple trend line. This time series schedule will provide a better understanding of the time series, and can be used to improve forecast accuracy. One of the most well-known methods of decomposing time into its components is classical seasonal decomposition (CSD). The basic idea of CSD is that there are four separate components, namely tendentious T_t, seasonal S_t, cyclic C_t and residual irregular (noise) ε_t, which make up the time series Y_t, which interact with each other (Makridakis et al. (1998). These components can combine in an additive or a multiplicative way (Mbuli et al., 2020). The general appearance of the additive model is as follows:

$$Y_t = T_t + S_t + \mathrm{C}_t + E_t, \quad t = 1,2,\ldots,n. \tag{1}$$

The general appearance of the multiplicative model looks like this:

$$Y_t = T_t \cdot S_t \cdot \mathrm{C}_t \cdot E_t, \quad t = 1,2,\ldots,n. \tag{2}$$

The Eqs. (1) and (2) assume that each level of the time series can be represented as the sum or multiplication of trend T, seasonal S, cyclic C and random E components.

The choice of one of the two models is based on the analysis of the structure of seasonal fluctuations. If the amplitude of oscillations is approximately constant, one can build an additive model of the time series, in which the values of the seasonal component are assumed constant for different cycles. If the amplitude of seasonal fluctuations increases or decreases, there will be a multiplicative model of the time series, which affects the levels of the series depending on the values of the seasonal component.

Therefore, the process of constructing additive and multiplicative models is reduced to the calculation of values of T, S and E for each level of the series, which involves the following steps (O'Connell et al., 1993).

Step 1. Alignment of the original series by the method of moving average.

1) summing up the levels of the series sequentially for each quarter (month) with

a shift of one point in time and aggregate the data;

2) dividing the amounts received by a number equal to the number of consolidated periods;

3) providing the obtained values in accordance with the actual moments of time, for which they find the average values of two consecutive moving averages - centered moving averages.

Step 2. Calculation of the values of the seasonal component S.

Step 3. Elimination of the seasonal component from the initial levels of the series and retention of aligned data $(T + E)$ in the additive or $(T \cdot E)$ in the multiplicative model.

Step 4. Analytical alignment of levels $(T + E)$ or $(T \cdot E)$ and calculation of T values using the obtained trend equation.

Step 5. Calculation of the values obtained by the model $(T + S)$ or $(T \cdot S)$.

Step 6. Calculation of absolute and/or relative errors.

If the resulting error values do not contain autocorrelation, they can replace the original levels of the series and then use the time series of errors E to analyse the relationships of the original level and other time series.

3.2. Exponential Smoothing Model. Exponential smoothing is one of the most successful classical forecasting methods. Exponential smoothing is a method of smoothing a time series which has a number of previous hours when forecasting, a computational procedure that includes the processing of all observations, taking into account the aging of information as it moves away from the forecast period (Vandeput, 2021). The principle of operation of this method is a parameter in that the previous values are taken into account with decreasing exponential weights. The simplest form of exponential smoothing is given by the formula (Holt, 1957):

$$y_{t+1} = \alpha y_t + (1 - \alpha)\, y_{t-1}, \qquad (3)$$

where y_{t+1} – the forecast value;

y_t – is the current value of the time series;

l_{t-1} – the value of the exponential means at the moment $(t - 1)$;

α – is the smoothing constant $0 < \alpha < 1$.

The simple exponential smoothing has one component included is the level, l_t, and is suitable for forecasting data without a clear trend or seasonal pattern. To take into account trends or seasonality, a combination of components is proposed: errors, trends and seasons. Each component can be combined in different ways, which are calculated by smoothing. These three terms (Error, Trend and Season) are formed ETS model. For example, errors can be combined additively or multiplicatively: $Error = \{\{A, M\}\}$; trend – additively, additive damped (A_d),) or excluded from the model (N (None)): $Trend = \{\{A, A_d, N\}\}$; seasonality – additively, or multiplicatively, or excluded from the model (N (None)): $Seasonal = \{\{A, M, N\}\}$. The combination of Error, Trend and Season can be done in three different ways, which provides many combinations of exponential smoothing models (Table 1), which are described in detail in (Hyndman et al., 2008 and 2018).

Table 1

Different combinations of ETS models

Models with none seasonality	Models with additive seasonality	Models with multiplicative seasonality
$ETS\ (A,N,N)$	$ETS\ (A,N,A)$	$ETS\ (A,N,M)$
$ETS\ (A,A,N)$	$ETS\ (A,A,A)$	$ETS\ (A,A,M)$
$ETS\ (A,A_d,N)$	$ETS\ (A,A_d,A)$	$ETS\ (A,A_d,M)$
$ETS\ (M,N,N)$	$ETS\ (M,N,A)$	$ETS\ (M,N,M)$
$ETS\ (M,A,N)$	$ETS\ (M,A,A)$	$ETS\ (M,A,M)$
$ETS\ (M,A_d,N)$	$ETS\ (M,A_d,A)$	$ETS\ (M,A_d,M)$

Source: authors' compilation based on (Hyndman et al., 2008 and 2018).

Among the existing forecasting methods that use exponential smoothing, the most well-known are Holt's linear method with additive errors ETS (A,A,N) (Holt, 1957), Multiplicative Holt-Winters' method with multiplicative errors ETS (*M*,*A*,*M*) (Winters, 1960), and Brown's model (Brown, 1959).

Exponential smoothing models, according to the combination of components, can be divided into the following types:

– linear models with additive error;

– linear models with multiplicative error;

– models with a linear trend, but a multiplicative error and a seasonal component;

– models with multiplicative error and a trend component or without seasonality, or with multiplicative seasonality;

– models that are difficult to evaluate, with a combination of additive and multiplicative elements.

Choosing the right model is based on recognising the key components of the time series (trend and seasonal) and the way they are included in the smoothing method (for example, additive, damping or multiplicative method).

To determine the best combination of ETS a model using several criteria such as Akaike's Information Criterion (AIC), Akaike's Information Criterion correction (AICc) and Bayesian Information Criterion (BIC) (Konishi and Kitagawa, 2008) can be used.

The advantages of this class of models are simplicity and efficiency. The advantages include the ability to account for the weights of the source information, the simplicity of computational operations, and the flexibility of describing the various dynamics of the processes. The method of exponential smoothing makes it possible to obtain an estimate of the trend parameters that characterise not the average level of the process, but the trend that has developed at the time of the last observation.

The main point of exponential smoothing is the correct choice of the smoothing parameter (smoothing constants) and the initial conditions, as the wrong choice can lead to negative consequences (Hyndman and Khandakar, 2008). In addition, one of the problems of applying this method relates to choosing the model that best recognises

the key components of the time series, such as seasonality, and the way in which it is included in the smoothing model.

3.3. Autoregressive Integrated Moving Average Model (ARIMA / Seasonal (SARIMA)). The Autoregressive Integrated Moving Average Model (ARIMA) is a time series forecasting technique. ARIMA is one of the most commonly used forecasting techniques in business today (Kotu and Deshpande, 2019). ARIMA refers to a model created by regressing the dependent variable only by its lag value and the current value and the lag value of the random error term in the process of converting non-stationary time series into stationary time series when taking into account several differences.

Box and Jenkins (1976) presented the equation of the ARIMA model (p, d, q):

$$Y'(t) = c + \Phi_1 Y'_{t-1} + \Phi_2 Y'_{t-2} + \cdots + \Phi_p Y'_{t-p} + \\ + \theta_1 \varepsilon_{(t-1)} + \theta_2 \varepsilon_{(t-2)} + \cdots + \theta_q \varepsilon_{(t-q)}, \quad (4)$$

where $Y'(t) -$ is the difference series;

$\Phi_p -$ is the coefficient of the first AR term;

$P -$ is the order of the AR term;

$\theta_q \ -$ is the coefficient of the first MA term;

$q -$ is the order of the MA term;

$\varepsilon_t -$ is the error.

The abbreviation for ARIMA is descriptive, capturing the key aspects of the model itself (Shumway and Stoffer, 2000). Briefly, they are:

AR: Autoregression: this is a process, in which the current value of the time series is regressed with its previous values, that is: y_{t-1}, y_{t-2} and so on. The order of lag is denoted as p.

I: Integration: The time series uses differencing to make it stationary. The order of the difference is denoted as d.

MA: Moving Average: In moving average models, the average current value of a stationary stochastic process is represented as a linear combination of the current and past error values $\varepsilon t, \varepsilon\ (t-1), \ldots, \varepsilon\ (t-p)$.

One of the main problems in modelling agriculture is the problem of seasonality. When analysing the oscillations in time series, along with the selection of random oscillations, there is a need to study periodic oscillations. As a rule, it is necessary to study seasonal fluctuations in order to exclude their influence on the general dynamics for the detection of pure (random) fluctuations. Seasonal fluctuations include all phenomena whose development shows a clear pattern of intra-annual changes, i.e. level fluctuations that are more or less steadily repeated from year to year.

A main disadvantage of the ARIMA model is that it does not take into account the seasonal factor. Since the average price of potatoes is seasonal, we cannot ignore it, and it must be checked.

The seasonal component of SARIMA models adds the following three components:

1. Seasonal Autoregressive SAR (P): This component captures the relationship

between the current value of the series and its past values, specifically at seasonal lags.

2. Seasonal Integrated I (D): Similar to the non-seasonal differencing, this component accounts for the differencing required removing seasonality from the series.

3. Seasonal Moving Average SMA (Q): This component models the dependency between the current value and the residual errors of the previous predictions at seasonal lags.

Seasonal effects can be overcome with a seasonal autoregressive moving average that was generalised to the ARIMA model to deal with seasonality by Box and Jenkins (1976).

$$\text{SARIMA} \quad \underbrace{(p,d,q)}_{\text{Non-seasonal part of the model}} \quad \underbrace{(P_s,D_s,Q_s)}_{\text{Seasonal part of of the model}}$$

A seasonal ARIMA model is formed by including additional seasonal terms in the ARIMA models. A SARIMA model with seasonal s periods is denoted by:

$$ARIMA\ (p,d,q)(P_s,D_s,Q_s)[s], \tag{5}$$

where p – autoregression order;

d – is the degree of differentiation;

q – is the number of moving averages;

s – refers to the number of periods in each season;

(P_s,D_s,Q_s) – represents (p, d, q) for the seasonal part of the time series:

P_s – is the order of seasonal autoregression, D_s – is the order of seasonal difference, Q_s – is the seasonal parameter of the moving average.

The general multiplicative ARIMA/SARIMA model is applied to the time series y_t and can be expressed as follows (Brockwell and Davis, 1991):

$$\Phi_P(L^s)\varphi_p(L)\Delta^d\Delta_s^D Y_t = \vartheta_q(L^s)\theta_q(L)\varepsilon_t, \tag{6}$$

where s is the seasonal length, for example s = 12 for monthly and s = 4 for quarterly data, L is the lag operator and ε_t is assumed to be a Gaussian white-noise process with mean zero and variance σ^2. The difference operator is Δ^d where d specifies the order of differencing and the seasonal difference operator is $\Delta_s^D$where D is the order of seasonal differencing. The difference operators are applied to transform the observed non-stationary time series Y_t to the stationary process Y_t' with the following equation (Brockwell and Davis, 1991):

$$Y_t' = (1-L)^d(1-L^s)^D Y_t, \tag{7}$$

where Φ_P and ϑ_q are the seasonal polynomials in the lag operator (or are called the seasonal AR and MA characteristics operators) and can be specified as follows:

$$\Phi_P(L^s) = 1 - \Phi_1 L^s - \Phi_2 L^{2s} \ -\ldots-\ \Phi_p L^{ps}, \tag{8}$$

$$\vartheta_q(L^s) = 1 - \vartheta_1 L^s - \vartheta_2 L^{2s} \ -\ldots-\ \vartheta_q L^{qs}. \tag{9}$$

Further $\varphi_p(L)$ and $\theta_q(L)$ are polynomials in the lag operator of P and Q respectively and the non-seasonal AR and MA characteristics operators:

$$\varphi_p(L) = 1 - \varphi_1 L - \varphi_2 L^2 - \ldots - \varphi_p L^p, \quad (10)$$

$$\theta_q(L) = 1 - \theta_1 L - \theta_2 L^2 - \ldots - \theta_q L^q. \quad (11)$$

Box and Jenkins (1976) propose a process for identifying, evaluating, and validating ARIMA models for a given set of time series data. This process is called the Box-Jenkins method and consists of the following three steps:

1. Identification. The time series is evaluated to determine if it is stationarity or non-stationarity, in the latter state the number of required cases needed to make it stationary is determined; seasonality in the dependent series is identified; the parameters of the ARIMA model for the data are determined; a decision as to which (if any) in the model should use the autoregressive or moving average component is made. Two diagnostic plots can be used to select the p and q parameters of ARIMA:

- Autocorrelation Function (ACF). The plot summarises the correlation of observations with the lag values. The x-axis shows the lag, the y-axis shows the correlation coefficient between -1 and 1 for negative and positive correlation.

- Partial Autocorrelation Function (PACF). The plot diagram summarises the correlations for observations with lag values that are not included in previous lag observations.

2. Estimation of parameters. Estimation involves the use of numerical methods to minimise loss or error. The most common methods use maximum likelihood estimation or non-linear least-squares estimation.

3. Diagnostic check of the model. This step is carried out by checking whether the estimation model corresponds to the specifications of a stationary one-dimensional process. In particular, the residuals must be independent of each other and constant in mean and change over time. To achieve this, making the plot the mean and variance of the residuals over time, and running a Ljung-Box test or plotting the autocorrelation, and partial autocorrelation of the residuals is useful for identifying errors in the specification (Ljung and Box, 1978; Long and Teetor, 2019). If the estimate is inadequate, it is necessary to return to the first step and try to build a better model.

3.4. Neural Network Autoregression Model (NNAR). An artificial neural network is a mathematical model based on the principle of organisation and functioning of biological neural networks – neural cellular networks of a living organism. From a mathematical point of view, it is a system of connected and interacting computing units, so-called artificial neurons. A neuron is a unit of information that receives information, calculates it, and passes it on. They are divided into three main types, as well as its corresponding layers. The neuron receives a set of signals. They can be either taken from the outside or transmitted by other neurons. The obtained values are multiplied by the weighting factor (Shiblee et al., 2008). A neuron forms a multilevel network with direct communication, where each level of nodes receives input from previous layers. The outputs of the nodes of one layer are the inputs for the next layer. The inputs to each node are combined using a weighted linear combination. The result is then changed using a nonlinear function before output:

$$z_j = \sum_{i=1}^{n} \omega_{i,j} y_i, \quad (12)$$

where ω – is the vector of synaptic weights (weighting factors of connection);

y – is the vector of input values.

Then the obtained value is passed to the activation function. Activation functions can take several forms. The type of activation function depends on the position of the neuron in the network. In most cases, the neurons of the input layer do not have the function of activation, because they perform the role of transmitting input data to the hidden layer (Hoptroff, 1993). A linear function is used for the source layer because a nonlinear activation function can distort the previous output. Logistic and hyperbolic functions are often used as hidden layer transfer functions. Formula (13) shows an example of the following function:

$$s(z) = \frac{1}{1 + e^{-x}}. \quad (13)$$

The value of this function is in the range of 0 to 1. The modified output is given to the next layer as an input. This process is useful for the network to be robust to outliers. The weights (ω) are randomly chosen and then learn from the observations by minimising the cost function.

In time series regression, lagged values of data are used as input by a neural network, which is called the Neural Network Autoregression (NNAR). The neural network consists of three layers: input, hidden, and output (Zhang et al., 1998). The input layer receives information, the hidden layer (or layers) processes it, and the output layer returns the results which are calculated based on the current input and the hidden state of the previous time step. The NNAR model has two components, p and k. p denotes the number of lagged values that are used as inputs, while k denotes the number of hidden nodes that are present. Output is denoted by $NNAR(p,k)$. For seasonal set of data, output is denoted by $NNAR(p, \mathrm{P}, k)_m$, where P is the number of seasonal lags, and m is the seasonal component. For a seasonal time series, it is better to include the last observation from the same term from the previous year and input layers can be present as $y_{t-m}, y_{t-2m}, \dots, y_{t-pm}$. An example of a neural network for seasonal time series is shown in Figure 1.

At the output layer, the forecast value of y_{t+1} is calculated by the formula:

$$y_{t+1} = b_0 + \sum_{i=1}^{k} w_i \, s(b_{0j} + \sum_{i=1}^{k} s(z)_k), \quad (14)$$

where b_0 – is the shift of the neuron of the output layer;

b_{0j} – shifts of neurons of the hidden layer;

w_i – the weight of the neurons, the outputs coming to the input of the neuron of the output layer;

y_{t+1} – the predicted value of the series for the $t+1$ time, k is the number of neurons on hidden layer;

$s(z)_k$ – the value of the activation function of neurons of the hidden layer.

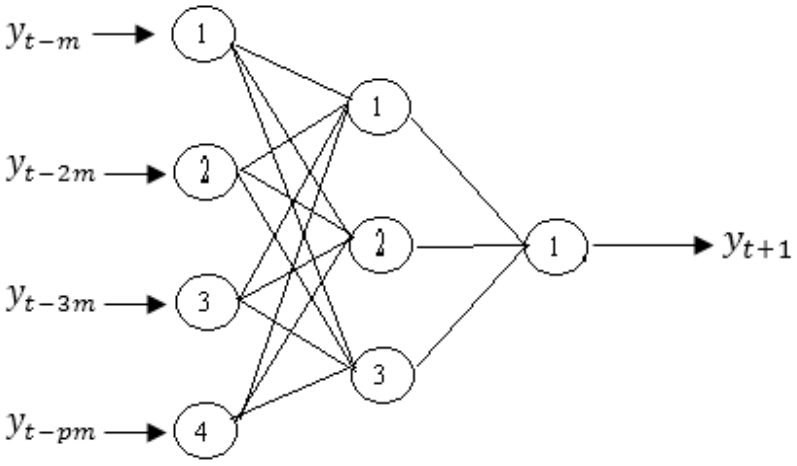

Figure 1. Neural network structure

Note. An example of a neural network for seasonal time series that has four input layers, three hidden layers, and one output layer.

Source: authors' elaboration based on (Zhan and Qi, 2005).

Another important point in building a model is the choice of learning speed for the neural network. While selecting the learning rate to train the neural network, we have to choose very carefully due to the following reasons:

1. If the learning rate is set too low, the training of the model will continue very slowly as we are making very small changes to the weights, as the step size determined by the equation of gradient descent is small. It will take many iterations before reaching the point of minimum loss.

2. If the learning rate is set too high it causes undesirable divergent behaviour to the loss function due to large changes in weights due to a larger value of step size. It may fail to converge (the model can give a good output) or even diverge (data is too chaotic for the network to train).

3. The allowable minimum learning speed of the neural network is 20 iterations

3.5. Model Estimation. Model quality assessment includes analysis of the accuracy, complexity and adequacy of the model. To assess the accuracy of the model and the complexity it needs to calculate the value of forecast errors, its forecasting accuracy is measured by the size of the forecast error – the difference between the forecast and the actual value of the studied variable.

It is extremely important to evaluate the forecast. To measure the accuracy of forecasts, the following characteristics are most often used: MSE, RMSE, MPE, MAE, MASE, MAPE, which must be minimised (Shcherbakov et al., 2013).

Mean Square Error (MSE):

$$MSE = \frac{1}{n}\sum_{t}(y_t - \hat{y}_t)^2. \quad (15)$$

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{MSE}. \quad (16)$$

Mean Percentage Error (MPE):

$$MPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{(y_t - \hat{y}_t)}{(y_t)}. \quad (17)$$

Mean Absolute Error (MAE):

$$MAE = \sum_{i=1}^{n}\frac{|y_t - \hat{y}_t|}{n}. \quad (18)$$

Mean Absolute Scaled Error (MASE):

$$MASE = \sum_{i=1}^{n}\frac{(y_t - \hat{y}_t)}{\frac{1}{T-m}(y_t - y_{t-m})}, \quad (19)$$

where m – seasonal period.

Mean Absolute Percentage Error (MAPE):

$$MAPE = \sum_{i=1}^{n}\frac{100|y_t - \hat{y}_t|}{n|y_t|}. \quad (20)$$

4. RESULTS

In this section, the results of analyses and predictions of time series of potato prices using the following methods and models: seasonal decomposition, exponential smoothing, SARIMA, neural network are presented. The construction of mathematical models was carried out in the software environment R and with the help of its open-source libraries.

4.1. Basic Data on the Time Series. Data for the average retail price of the potatoes that were used for analysis and forecasting were obtained from the official statistics portal of Latvia (Official statistics portal, 2021). The dataset contains monthly average retail price of the potatoes from Jan. 2005 to Dec. 2021 in national currency (euro per 1 kg).

The Figure 2 below shows a slight rising trend from 2005 to 2021 of data collection from the Table 2, and drawing conclusions about seasonality because every summer there is a peak, and every winter – a deflection. Seasonality of the average price on potatoes is due to the fact that some months of the year are more important in terms of activity or level. In our case it originates from climate and conventional seasons, which repeat from year to year. Moreover, one of the main reasons for the seasonality and peak of potato prices in the summer months can be explained by the beginning of the season and the shortage of supply, as well as uncertainty about the amount of harvest.

The blue line in the figure shows the average value of the price of potatoes for a given year and shows the growth trend from year to year. Taking 2005 as the base year, we can see that the average price of potatoes increased by 78 % in 2021.

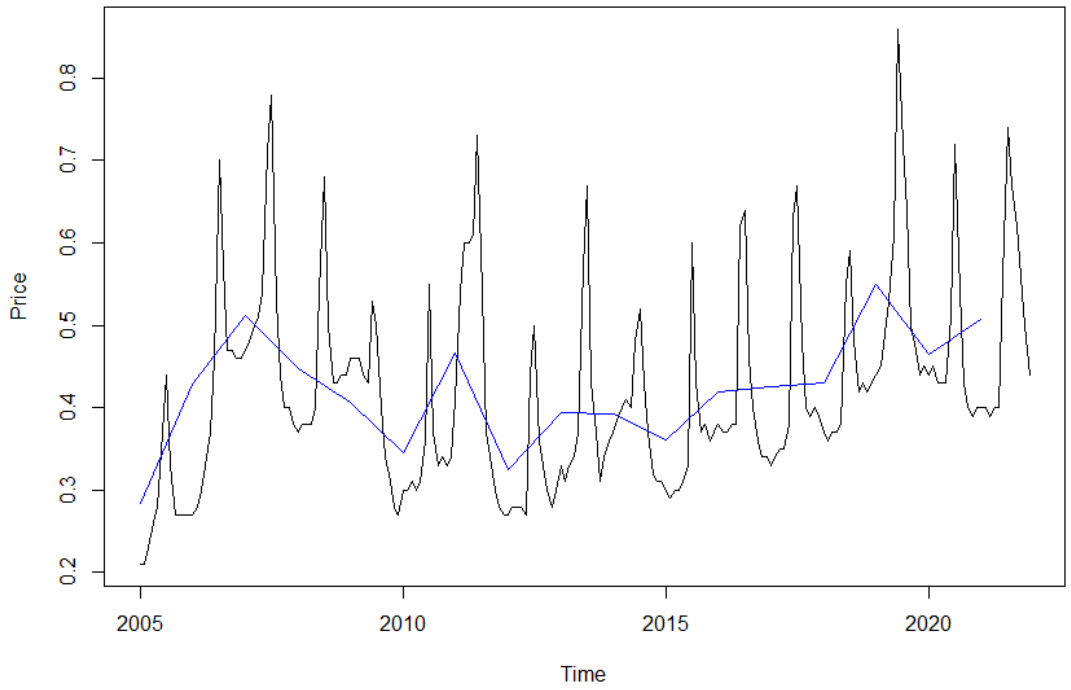

Figure 2. Monthly average price of potatoes dataset plot from Jan. 2005 to Dec. 2021 in Latvia

Source: authors' elaboration.

The dynamics of change in the average price of potatoes is shown in the box plot below (Figure 3). From this graph it is easy to see that the price has seasonal fluctuations.

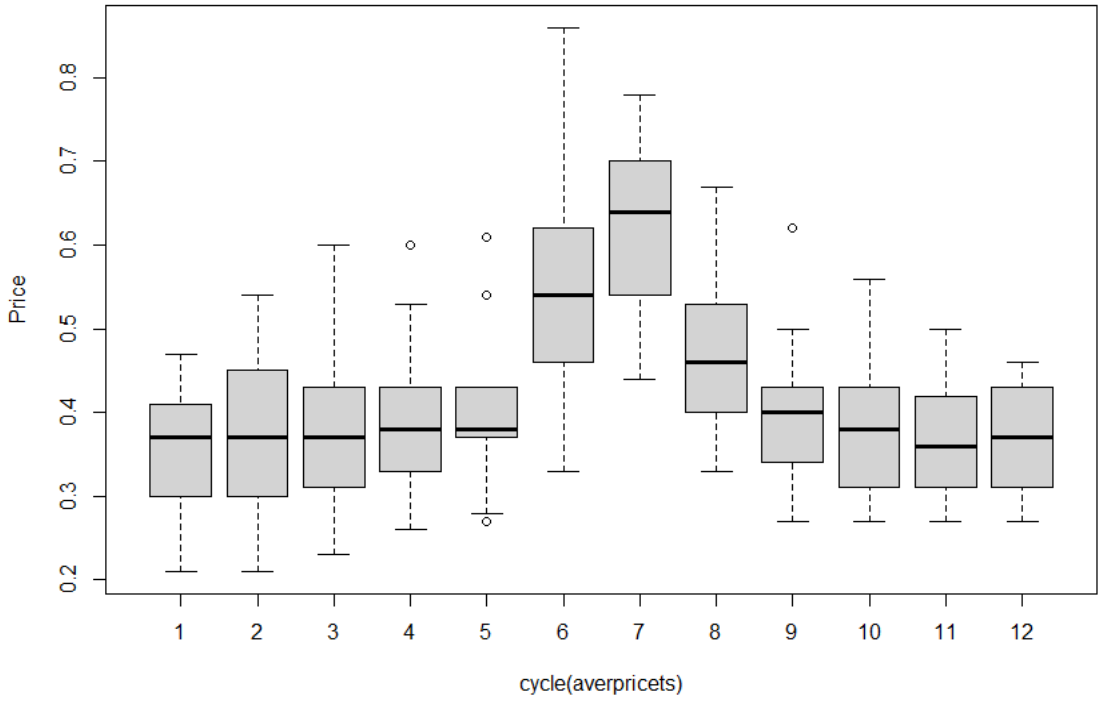

Figure 3. Box-Whisker plots of monthly average price of potatoes from Jan. 2005 to Dec. 2021 in Latvia

Source: authors' elaboration.

When a time series has a seasonal pattern, it is important that the forecasting

model captures this pattern. Analysis of seasonal time series allows us to make assumptions about the following values of the variable in the future. The accuracy of this prediction depends on the nature of the series and the method of analysis used. To verify the accuracy of the constructed models, it was decided to divide the original sample into two parts, namely – training and test.

According to the training sample for each model, we determined the optimal values of the coefficients. Then each model was tested on a test sample. We chose the best model for which the criteria based on equations (15)-(20) in the test sample are minimal and based on it we made a forecast. The Figure 4 below shows the division of data into training and test sets.

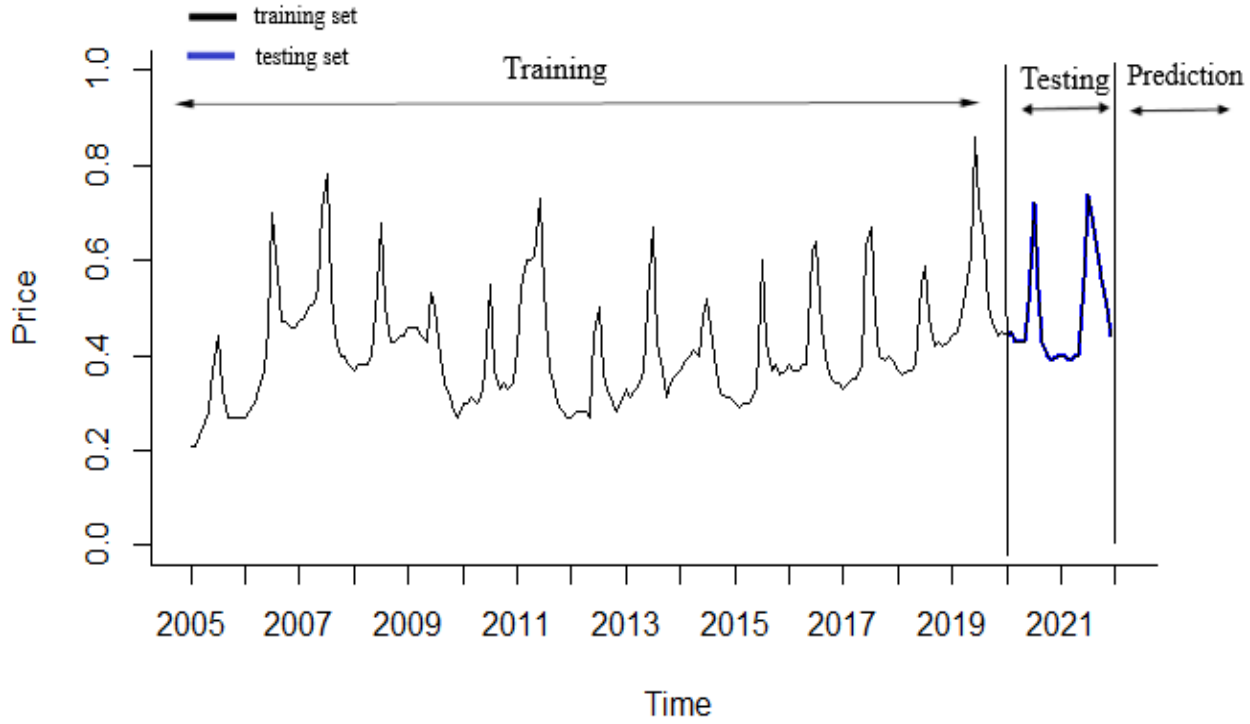

Figure 4. Graphical representation of the division of the time series into training and testing sets

Source: authors' elaboration.

We have 204 observations at monthly intervals (17 years), a good approach would be to keep the first 180 records (15 years) for training and the last 24 records (2 years) for testing. While such a division does not violate the condition of completeness of the training set.

4.2. Development of the Decomposition Method. Time series data can show a variety of patterns, so we need to divide the time series into components, each of which will represent the main category of the pattern. We did this using decomposition. We divide the data into its components: trend, seasonality, and white noise with this method.

As we can see from the Figure 5, decomposition gives results when the trend cycle tends to indicate excessively smooth rapid rises and falls of data, and the seasonal component is repeated from year to year. When the wave height increases, the amplitude of oscillations is not constant, that is why an additive model can represent

the average price of potatoes. The general model of the series can be represented as y = t+s+e, where t is a trend, s is seasonality, and e is an error. The decomposed series, and any assumptions about the additive combination, are comparable to the model constructed using the stl( ) function.

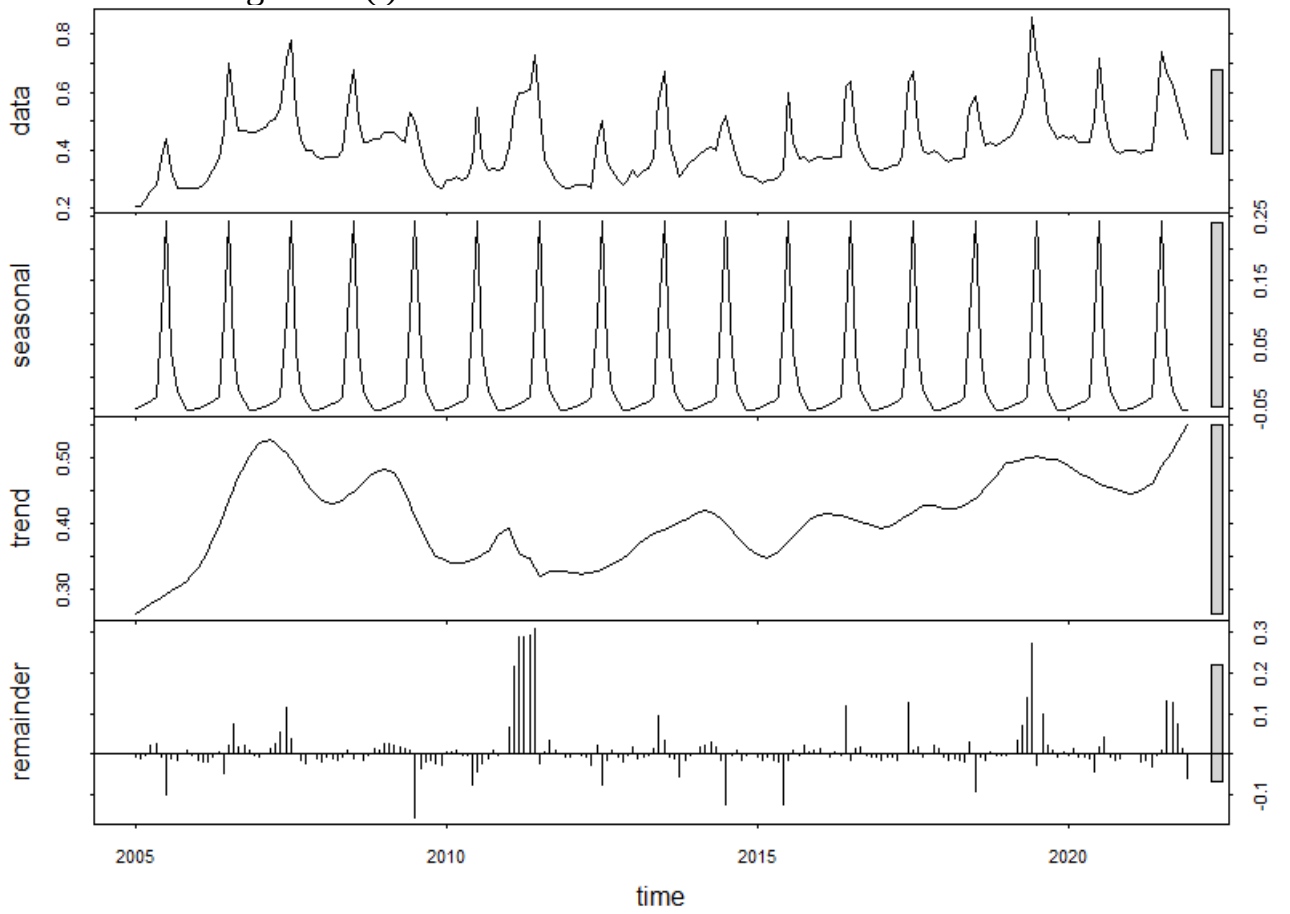

Figure 5. Decomposition of time series data

Note. The top panel depicts the actual data (training set) from Jan. 2005 to Dec. 2019 and the bottom three additional components, obtained as a result of classical decomposition: decomposed trend trait, the decomposed seasonal trait, and the decomposed random fluctuation trait.

Source: authors' elaboration.

Forecasts of stl ( ) use time series data, apply decomposition, and model the seasonally adjusted data using the model passed as model function or specified using the "naive" method. The forecasts assume that the coefficients in future periods will be obtained based on the value of the last observation.

In order to make the forecast, we decomposed time series for the following components: trend T_t, seasonal component S_t and the residuals E_t. The general model of the series was been represented as additive model $Y_t = T_t + S_t + E_t$

The resulting forecasts of the original data are shown in the Figure 6.

Forecasts of the average price of potatoes were built based on a naïve forecast of the seasonally adjusted data and a seasonal naïve forecast of the seasonal component, after an STL decomposition of the training set data.

According to the graph, we can say that the time series of the studied data remains the trend, and we can assume the existence of seasonal fluctuations.

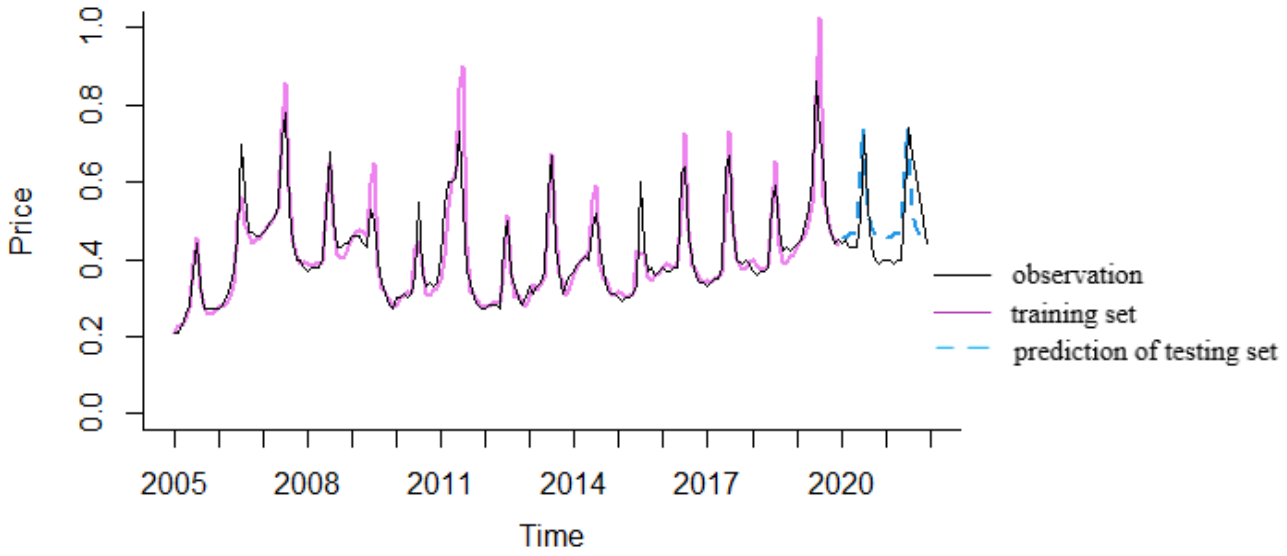

Figure 6. Forecasting the average prices of the potatoes in Latvia using decomposition method

Source: authors' elaboration.

4.3. Development of the ETS Model. The exponential smoothing model can be applied to time series data to produce smoothed data for presentation and to make forecasts. If the data has no trend or seasonal patterns, then simple exponential smoothing is appropriate or if the data demonstrate a linear trend, then Holt's linear method is appropriate. Nevertheless, in this case, the data is seasonal and these methods cannot solve the problem well. It is necessary to build the best model that can explain seasonality and predict future values. Because there are many combinations of exponential smoothing models, the construction and analysis of all of them will take a long time.

For this reason, this study's models were selected using the ets function of the forecast package *R*. This function automatically determines the best model of different combinations of ETS (the best model is elected according to its AIC) based on the exponential moving average filter.

We obtained the best exponential smoothing model as $ETS\ (M, A_d, M)$ described by the following components:

- Multiplicative error;
- Additive damped trend;
- Additive seasonality.

The ETS $(\mathrm{M}, \mathrm{A_d}, \mathrm{M})$ representation for this model may be written as:

forecast equation $$y_t = (l_{t-1} + \varphi b_{t-1} + s_{t-m})(1 + \varepsilon_t), \quad (21)$$

smoothing equation $$l_t = l_{t-1} + \varphi b_{t-1} + \alpha(l_{t-1} + \varphi b_{t-1} + s_{t-m}), \quad (22)$$

trend equation $$b_t = \varphi b_{t-1} + \beta(l_{t-1} + \varphi b_{t-1} + s_{t-m})\varepsilon_t, \quad (23)$$

seasonal equation $$s_t = s_{t-m} + \gamma(l_{t-1} + \varphi b_{t-1} + s_{t-m})\varepsilon_t. \quad (24)$$

The coefficient γ (gamma) always appears in the model that characterises seasonality. This is the smoothing factor for the seasonal component. We see that the

forecast equation consists of level, trend and seasonal component.

Table 2 shows the information coefficients and criteria of the best exponential smoothing model as $ETS\ (M, A_d, M)$ which were obtained by using the ets function in the programming language R.

Table 2

Results of the ets function

ETS (M, A_d, M)					
Smoothing parameters:					
α	β	γ	φ		
0.8132	0.0217	1e-04	0.8003		
Initial states:					
l			b		
0.2220			0.0244		
s_0	s_{-1}	s_{-2}	s_{-3}	s_{-4}	s_{-5}
0.8406	0.8486	0.881	0.9387	1.1309	1.5574
s_{-6}	s_{-7}	s_{-8}	s_{-9}	s_{-10}	s_{-11}
1.2977	0.9464	0.9169	0.9024	0.8805	0.859
Sigma: 0.0965					
AIC		$AICc$		BIC	
-221.9		-217.65		-164.42	

Source: authors' study.

The results of exponential smoothing in the Table 2 show the smoothing parameters $(\alpha, \beta, \gamma, \varphi)$ and initial states $x_0 = (l_0, b_0, s_0, s_{t-1}-, \dots, s_{t-11})$. We can see that $ETS\ (M, A_d, M)$ model has relatively low values of several criteria such as AIC of -221.9, $AICc$ of -217.65 and BIC of -164.42. All these criteria indicate that we fitted the better and the simpler model.

The Figure 7 visualises the model on the training set and prediction values on the test set. We see that our forecast projects a seasonal estimate into the future. Moreover, the forecast indicators are very close to those observed.

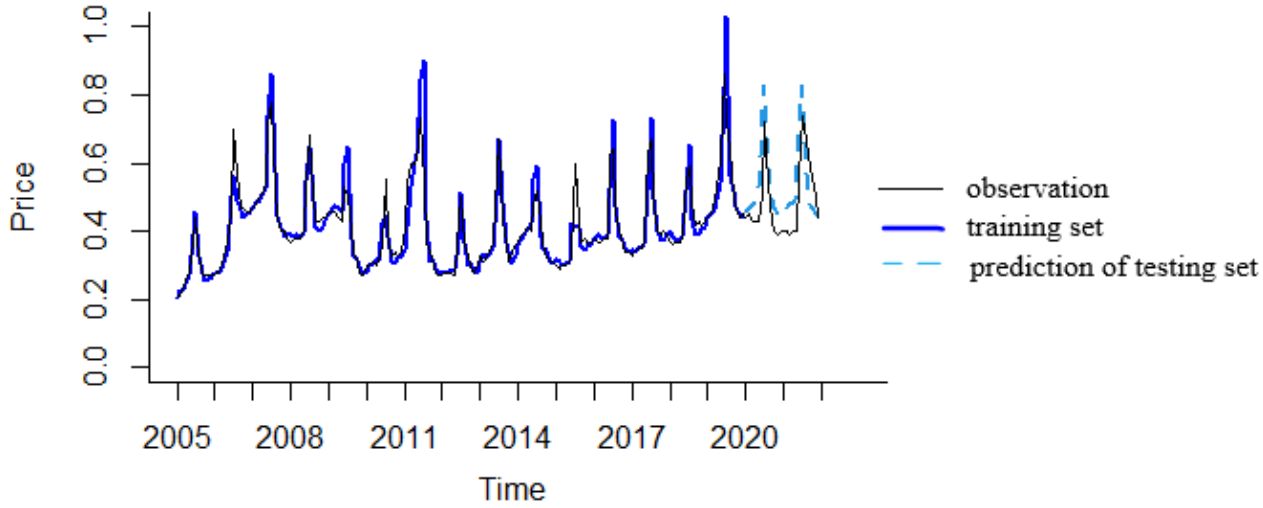

Figure 7. Forecasting the average prices of the potatoes in Latvia using the $ETS\ (M, A_d, A)$ model

Source: authors' elaboration.

4.4. Development of the SARIMA model. For seasonal phenomena there is a dense correlation between the levels of a series of dynamics corresponding to the same month. The seasonal data series in Figure 2 has an increase in summer and a decline in winter. We recommend first manually analysing the time series to verify our assumptions. We can explore seasonality with a correlogram. Construction of the autocorrelation function (ACF) and partial autocorrelation (PACF) plots is required.

The Figure 8 shows that the autocorrelation lag should be a multiple of 12, i.e. the January observation should be compared with the January observation, but from the previous year.

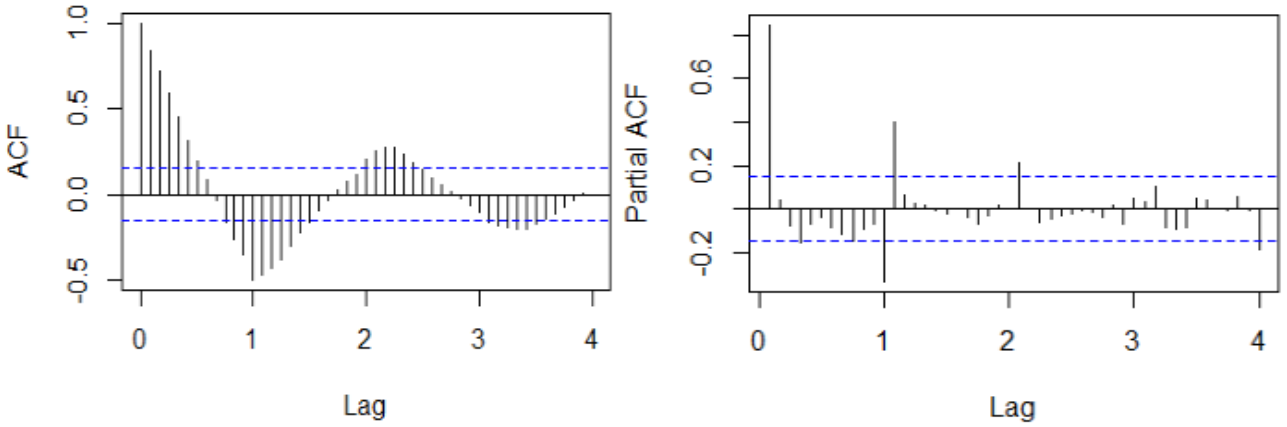

Figure 8. Correlograms of ACF and PACF of a time series with seasonal fluctuations

Source: authors' elaboration.

In the plots above of the seasonally differenced data, there are spikes in the PACF at lags 12, this may be suggestive of a seasonal AR(1) term. The ACF plot shows that the strongest positive correlation occurs at lag 1, which occurs after a period of negatively correlated lags. Moreover, the coefficient is significant (to override 95 % of the differences). This fact indicates a significant relationship between observations for one month, but for different years. In this regard, 12 is a suitable seasonal parameter for the model. For the PACF, we see that there is a strong correlation cut-off at lag 1. This means that the series follows the process AR (1) and the corresponding value for p = 1.

From the above, we see that there is a clear seasonal component in the time series. As also shown in the ACF chart, the ARIMA model will require an attached seasonal component.

We know that the order of the SARIMA model is denoted by the combination $(p, d, q)(P_s, D_s, Q_s)[s]$. Analysis of the correlogram and the seasonality plot does not give a 100 % guarantee of the correct choice of values of AR and MA, and the order of model combinations. This is often done manually, but the process is long and tedious. That's why the ARIMA model was fitted by using the $auto.arima$ ( ) functions of the forecast R package. Using the "$auto.arima$" function with seasonal influence in R package the best model was selected as $SARIMA$ $(1,0,0)(0,1,1)[12]$.

The methodology of the ARIMA model indicates that we should consider the

numerous models and choose a better model based on AIC, AICc and BIC criteria.

Table 3 below presents coefficients of model and information criteria such as the Akaike information criterion (AIC), the Akaike information criterion with a correction for small sample sizes (AICc), and the Bayesian information criterion (BIC).

Table 3

Results of the $auto.arima$ function

SARIMA (1,0,0)(0,1,1)[12]		
Coefficients:	$ar1$	$sma1$
	0.8681	-0.8864
Standard error	0.0402	0.0937
Sigma^2 = 0.002395 log likelihood = 259.9		
AIC	$AICc$	BIC
-513.80	-513.65	-504.43

Source: authors' study.

$SARIMA$ (1,0,0)(0,1,1)[12] was selected because it has the smallest values for AIC and BIC and is the less complex model (low p and q values).

Equation (6) describes the general expanded structure of the model, based on it, the structure of $SARIMA$ (1,0,0)(0,1,1)[12] where ($p = 1, d = 0, q = 0;\ P = 1, D = 1, Q = 1, S = 12$) can be written:

$$(1 - \varphi_1 L)(1 - L^{12})Y_t = (1 - \vartheta_1 L^{12})e_t. \quad (25)$$

The optimal SARIMA model includes a non-seasonal specification of $AR(1)$ order, no non-seasonal differencing, and MA, and then no the seasonal specification of seasonal AR, seasonal differencing that equal 1, seasonal MA(1), and the seasonal period that equals 12.

Using the estimated coefficients in Table 3, the SARIMA model can be written mathematically:

$$Y_t = 0.8681Y_{t-1} - 0.8864e_{t-12}. \quad (26)$$

The forecast result is given in the Figure 9.

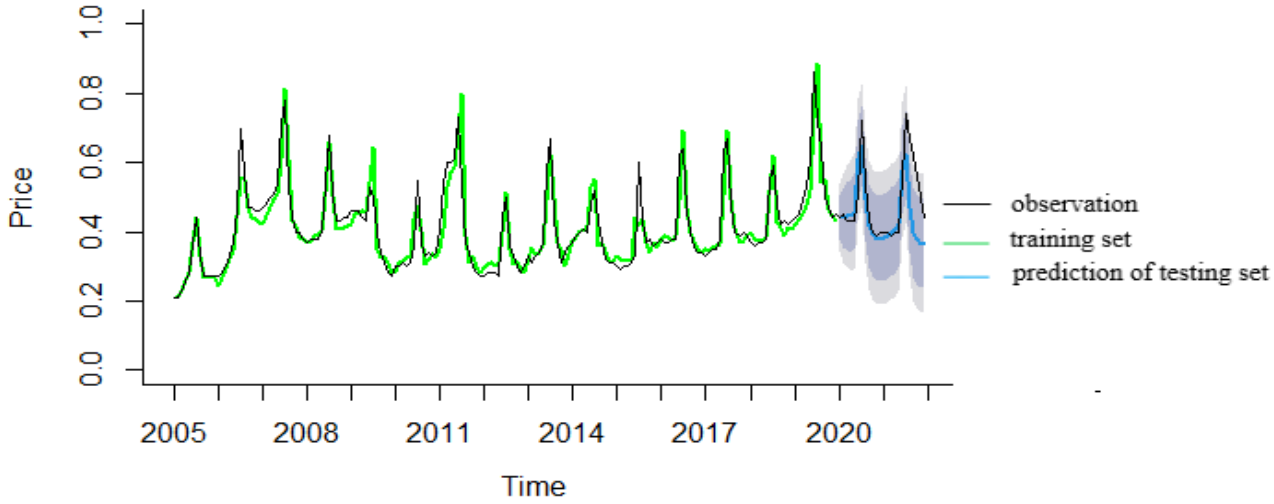

Figure 9. Forecasting the average prices of the potatoes in Latvia using the $SARIMA$ (1, 0, 0)(0, 1, 1)[12]

Source: authors' elaboration.

The Figure 9 provides information on forecasting results of potato average price based on *SARIMA* (1,0,0)(0,1,1)[12] model. We built the forecast with the testing set for the next two years. We obtained the results of forecast calculations, the values of the main criteria of forecast quality, as well as a graphical display of forecast values and 95 % confidence intervals. The data show an obvious seasonal pattern, with peaks observed in the summer quarter of each year.

We used the Ljung-Box test for diagnostics for the ARIMA model:

```
        Ljung-Box test

data:  Residuals from ARIMA(1,0,0)(0,1,1)[12]
Q* = 10.831, df = 22, p-value = 0.9771

Model df: 2.   Total lags used: 24
```

The results of the Ljung-Box test indicate that the model residues are statistically independent. The *p*-value for the Ljung-Box statistic was greater than 0.05, which means that the null hypothesis of independence for this residual series cannot be rejected.

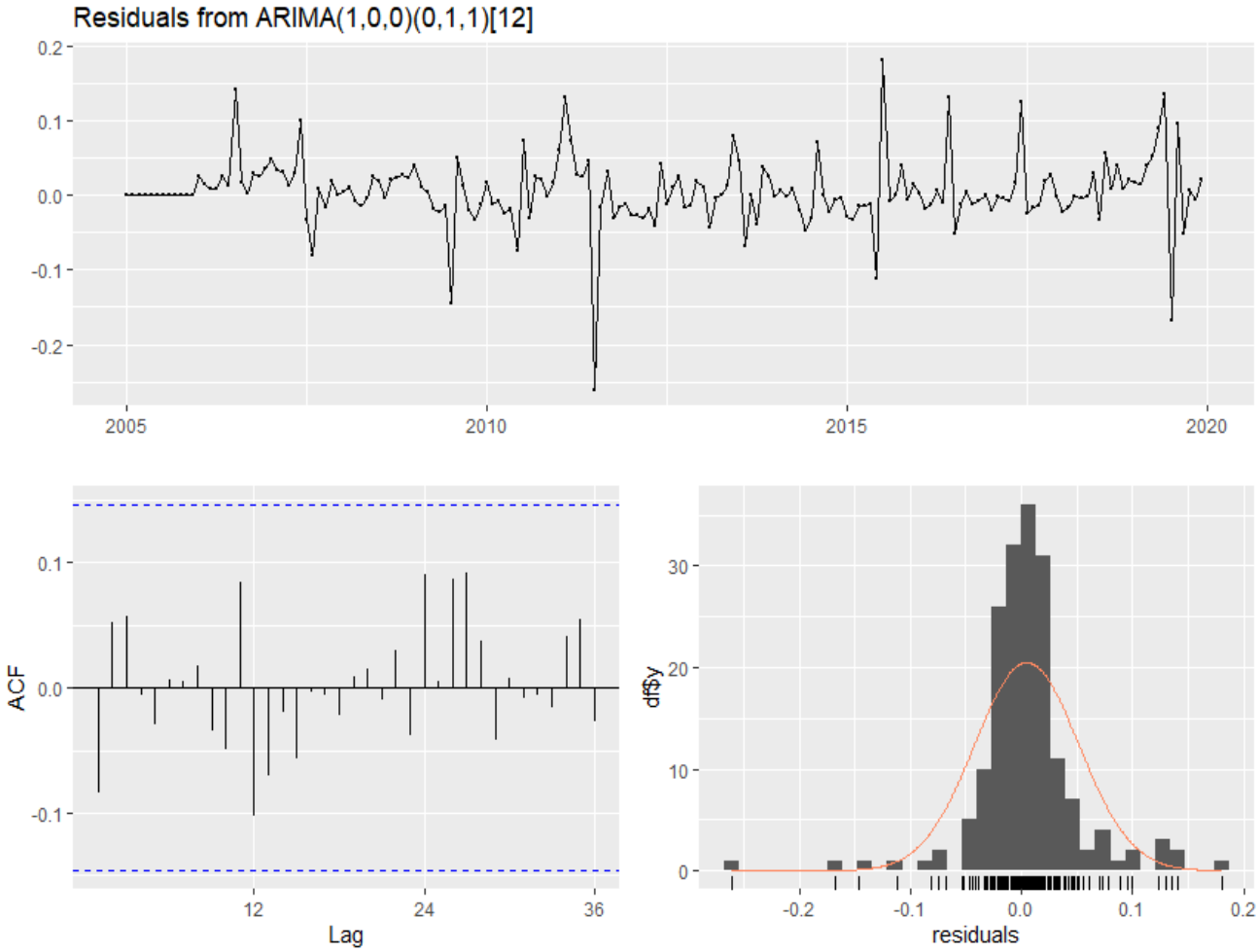

Figure 10. Residual plots for *SARIMA* (1, 0, 0)(0, 1, 1)[12]

Note. The top panel depicts a standardised residual plot the actual data from Jan. 2005 to Dec. 2019. The bottom ACF of the errors at various lags and Normal Plot of residuals.

Source: authors' elaboration.

We can make the following conclusions from Figure 11:

- standardised residuals do not show volatility clusters;
- autocorrelation function (ACF) does not show significant autocorrelation between residuals;
- the residuals are bell-shaped, which indicates that they are quite symmetrical;
- the p-value in the Ljung-Box test is large, which indicates that the residuals have no patterns, that is, all the information was extracted by the model and only noise remained.

4.5. Development of the NNAR Model. The R software allows the construction of neural networks with different initial values of weights and shifts of neurons of the hidden and source layers. As a forecast value it is necessary to calculate the average of the forecasts received by each network. The model of time series prediction using a neural network, implemented in the process of the study, is similar to the model of nonlinear autoregression of order p. The main difference is that the parameters of autoregression are not the estimates obtained by the method of least squares, and the weights of the neurons of the hidden layer and their displacements. The implemented neural network model usually has one hidden layer and p inputs. $Nnar(\,)$ functions of R were used and they defined the best neural network autoregression model as NNAR $(1,1,2)_{[12]}$. The number of neurons in the hidden layer k for each neural network was determined iteratively by testing different neuron schemas.

We used the "$nnetar$" function in R package for building *Neural Network Auto Regression Model.* We didn't specify components such as p, P, and k in the $nnetar()$ function, they are selected automatically. We used the Box-Cox power transformation to transform data into a "normal shape". In order to do this, we included a Box-Cox transformation with $lambda = 'auto'$ that helps to find the best value.

For our data, $nnetar()$ found a neural net as $NNAR\ (1,1,2)[12]$ where $p = 1, P = 1, k = 2$ and $m = 12$. It is mean that the model takes as inputs the average price at $t - 1$ and $t - 12$ and has one hidden layer with 2 neurons as shown in the Table 4 below.

Table 4

Results of the $nnetar$ function

NNAR (1,1,2)[12]
Average of 20 networks, each of which is a 2-2-1 network with 9 weights options were – linear output units
sigma^2 = 0.02825

Source: authors' study.

The result in the table shows training the neural network entails estimating 9 different parameters, 2 for each neuron in the hidden layer and output layer.

The Figure 11 shows visually the quality of the fitted NNAR model. The fitted values of the model follow perfectly all the components of the original data. The learning process of the NNAR model allows us to better understand the times series characteristics. All the components are well presented.

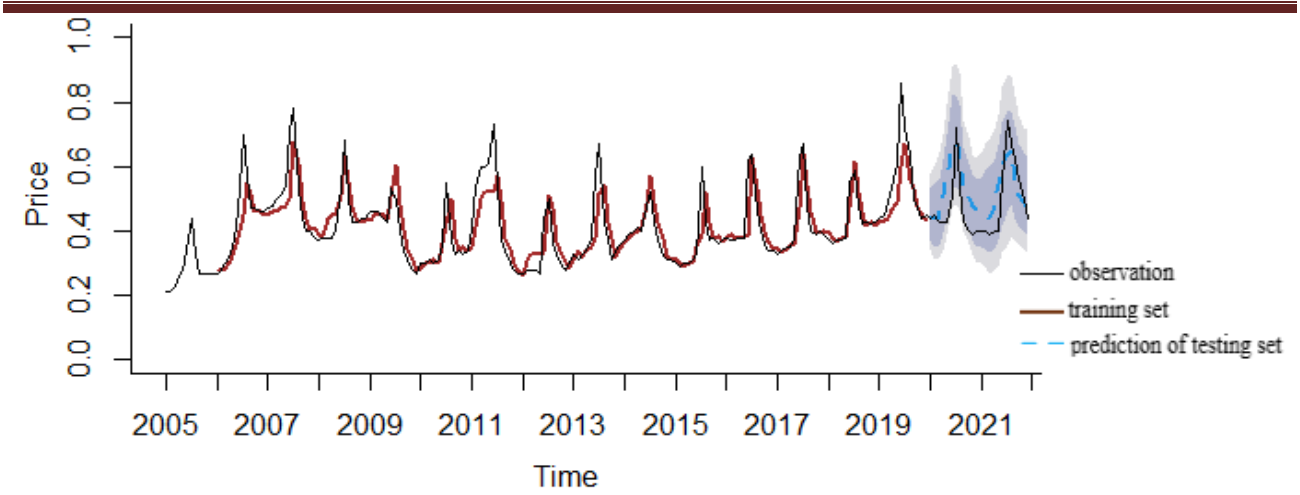

Figure 11. Forecasting the average prices of the potatoes in Latvia using the $NNAR\ (1,1,2)[12]$ model

Source: authors' study.

4.6. Measures of Accuracy. The predictive capabilities of the four methods can be compared by the measures ME, RMSE, MAE, MPE, MAPE, and MASE. For two different sets of time series, the results are shown in the Tables 5 and 6:

Table 5

Accuracy measures for the training set of average prices of time series on fitted models

Training set	ME	RMSE	MAE	MPE	MAPE	MASE
Decompositional	0.0014	0.0524	0.0299	0.3274	6.4675	0.3243
ETS	-0.0007	0.0530	0.0284	-0.1761	6.1072	0.3089
SARIMA	0.0047	0.0469	0.0288	0.3865	6.3370	0.3127
NNAR	0.0057	0.0664	0.0426	-0.6021	9.1729	0.4623

Source: authors' study.

Table 6

Accuracy measures for the testing set of average prices of time series on fitted models

Testing set	ME	RMSE	MAE	MPE	MAPE	MASE
Decompositional	-0.0182	0.0836	0.0529	-5.4881	11.2416	0.5738
ETS	-0.0464	0.0828	0.0739	-10.6612	15.3126	0.8021
SARIMA	0.0377	0.0661	0.0560	6.1889	10.1221	0.6077
NNAR	-0.0433	0.0817	0.0685	-10.9633	14.8585	0.7437

Source: authors' study.

In terms of the training set, all models fitted well, while the fitted time series by ETS and SARIMA models were especially approximate to the original data. As far as the testing set the SARIMA model was better than the other. It appeared that this model may be performed better in the validation set compared with other models.

As can be seen from the calculations, the RSME for the training set of all 4 models is below 10 %, which means that all the selected models very well describe the set of time series data. The mean absolute error, MAE, is the average (absolute) difference between the actual value and the forecast value is very low for four models. MAPE states that the error in all models was less than 10 % of the actual values.

Therefore, all models show high accuracy of forecasts and any forecast can be used. However, among all models, the best one can be identified as the one which has the lowest values compared to others on both sets: training and testing and it is model $SARIMA\ (1,0,0)(0,1,1)[12]$, it has the highest accuracy and the lowest error, which amounted to 10.1 % of the actual values.

Based on the best model, we obtained the following predicted values (Table 7).

Table 7

Future predictions of SARIMA (1,0,0) (0,1,1) [12] model

Month	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2022	0.372	0.241	0.502	0.172	0.572
Feb 2022	0.378	0.247	0.509	0.178	0.578
Mar 2022	0.392	0.261	0.523	0.192	0.592
Apr 2022	0.398	0.267	0.529	0.198	0.599
May 2022	0.417	0.286	0.548	0.216	0.618
Jun 2022	0.573	0.442	0.705	0.372	0.774
Jul 2022	0.618	0.487	0.750	0.417	0.819
Aug 2022	0.467	0.336	0.599	0.266	0.669
Sep 2022	0.396	0.264	0.527	0.194	0.597
Oct 2022	0.375	0.243	0.506	0.173	0.576
Nov 2022	0.363	0.232	0.495	0.162	0.565
Dec 2022	0.365	0.234	0.497	0.164	0.567

Source: authors' study.

We obtained a forecast that estimated the middle of the future range of possible values of data. Often, a forecast is accompanied by a prediction interval giving a range of values the variable could take with relatively high probability. Table 7 shows 80 % and 95 % intervals for the future average price of potatoes with lower confidence limit (Lo) and upper confidence limit (Hi).

5. DISCUSSION

Interpreting our results reveals several notable insights when compared to the findings of other researchers in the field. Our application of data science techniques, particularly machine learning models like ARIMA and neural network, showed a marked improvement in forecast accuracy for agricultural product prices. This aligns with the work of Sun et al. (2023) and Khashei et al. (2012), which also reported enhanced prediction precision using similar methodologies. However, our approach differed in the incorporation of testing on real data with a broader range of historical data. Also, using cross-validation techniques to assess the performance and robustness of various models.

In contrast, studies by Hyndman et al. (2008) and Mbuli et al. (2020) primarily focused on traditional time series models without the integration of machine learning techniques, resulting in less accurate forecasts. Our findings suggest that while traditional models can capture general trends, they may lack the ability to account for seasonality and complex non-linear relationships within the data. Additionally, the inclusion of feature programming in our models, particularly in handling seasonality

and trend decomposition, demonstrated significant benefits in improving forecast robustness, which was not extensively covered in previous studies.

The findings from our study have significant practical implications for agricultural producers, distributors, policymakers, and other stakeholders within the agricultural sector. By using accurate price forecasts generated through advanced data science models, stakeholders can make more informed decisions, ultimately leading to improved outcomes:

1. Optimising Production Schedules: producers can leverage these forecasts to plan optimal planting, harvesting, and selling times. For instance, if forecasts predict an increase in prices in the near future, farmers might choose to delay selling their produce or invest in storage solutions to capitalise on higher market prices later, thereby increasing profitability;

2. Resource Allocation: With better foresight into price trends, producers can allocate resources like labor, fertilisers, and irrigation more efficiently. This allows for cost savings and reduces the risk of overproduction or underproduction, aligning supply more closely with expected demand;

3. Inventory Management: distributors and retailers can use price forecasts to manage their inventories more effectively. By anticipating periods of high demand or price fluctuations, they can adjust their procurement strategies accordingly, ensuring they have sufficient stock during peak times while avoiding overstocking during periods of low demand;

4. Supply Chain Optimisation: accurate forecasts can help in negotiating better contracts with suppliers and setting competitive prices for consumers. This ensures that supply chain operations are more synchronised with market conditions, reducing waste and increasing efficiency;

5. Market Regulation and Support Programs: policymakers can use these forecasts to anticipate market volatility and implement regulatory measures to stabilise prices;

6. Risk Management: investors and financial institutions involved in the agricultural sector can use these forecasts to assess risks more accurately;

7. Price Stability: improved price forecasts can lead to more stable consumer prices, reducing the impact of sudden price spikes on household budgets. This is particularly important for staple goods like potatoes, where price fluctuations can have a significant impact on food affordability for consumers.

In summary, the ability to accurately forecast potato prices using advanced data science models provides critical insights that can be used by various stakeholders to make strategic decisions, optimise operations, and ultimately contribute to a more stable and efficient agricultural market.

6. CONCLUSIONS

Time series forecasting of agricultural product prices using the example of potatoes with data science techniques offers significant potential for enhancing decision-making in the agricultural sector. This study highlights the importance of

capturing seasonality in forecasting models, especially when dealing with time series data that exhibit seasonal patterns. The article provides an overview of traditional forecasting methods and neural network models for modelling seasonal time series, with a specific focus on predicting the price of potatoes, since the chosen research object clearly characterises the seasonal component. Such time series forecasting models as a decomposition-based model, exponential smoothing model of ETS type: $ETS\ (M, A_d, A)$; autoregressive integrated moving average model of type: $SARIMA\ (1,0,0)(0,1,1)[12]$; and neural network autoregression model: $NNAR\ (1,1,2)_{[12]}$ were constructed and analysed. Predicted values that preserve seasonality were calculated for all models.

The performance of all models was tested based on the following characteristics: MSE, RMSE, MPE, MAE, MASE, MAPE. The results showed that the $SARIMA\ (1,0,0)(0,1,1)[12]$ model could achieve good forecast results when applied to real problems and, thus, can be effectively used for forecasting tasks. In addition, it should be noted that the $ETS\ (M, A_d, A)$ model has a higher prediction accuracy at the time intervals closest to the original data. The obtained results support using both models simultaneously, which can compensate for the shortcomings of each of them. The models can be used separately, to more accurately predict the values for the required time period, or a combination of them is also possible. The results of the study can be adapted and implemented at different levels of management. They provide a basis for developing long-term strategies and can be integrated into various aspects of business planning.

7. LIMITATIONS AND FUTURE RESEARCH

While our study demonstrates the potential of data science techniques in improving the accuracy of agricultural product price forecasting, several limitations need to be addressed. Firstly, our analysis relied on historical price data without consideration of external variables such as weather conditions, market demand, import/export regulations, and economic indicators etc. Second, the complexity of machine learning techniques such as regression analysis, decision trees and random forests were not used.

The limitations mentioned above indicate potential areas for future research that could improve our understanding of the needs of agricultural farmers. These limitations highlight the need for ongoing research and development to enhance the robustness, accessibility, and applicability of data science techniques in agricultural price forecasting.

An important area of further research is the combination of different forecasting methods, the use of time series filtering, and the application of deep learning techniques to predict time series. Integrating various forecasting methods can leverage the strengths of each approach, resulting in more robust and accurate predictions. Time series filtering techniques, such as moving averages and Kalman filters, can help smooth out noise and reveal underlying trends, improving model performance. Deep learning models, such as Long Short-Term Memory (LSTM) networks and

Convolutional Neural Networks (CNNs), offer advanced capabilities for capturing complex patterns and dependencies in time series data. These models can address the limitations of traditional machine learning by handling large volumes of data and learning intricate relationships.

Combining classical models like ARIMA with machine learning algorithms and deep learning frameworks can create hybrid models that improve forecast accuracy. Time series filtering methods can be incorporated into pre-processing steps to enhance data quality before feeding it into predictive models. Additionally, ensemble methods that aggregate predictions from multiple models can further reduce errors and increase robustness.

The use of deep learning in time series forecasting is particularly promising due to its ability to model non-linearity and long-term dependencies, which are common in agricultural price data. For instance, LSTM networks can capture seasonal effects and sudden changes in trends, making them well-suited for agricultural price forecasting. CNNs can be used to extract features from multivariate time series data, providing additional context and improving prediction accuracy.

Future research should focus on developing and testing these hybrid approaches in various agricultural contexts. Investigating the integration of external data sources, such as weather conditions and economic indicators, can further enhance model performance. Exploring different architectures and hyperparameters for deep learning models will also be crucial to optimise their effectiveness for specific forecasting tasks.

Acknowledgments: this research was partly supported by the European Commission, Research Executive Agency grant number 101079206, "Twinning in Environmental Data and Dynamical Systems Modelling for Latvia" (TED4LAT).

Conflicts of interest: the authors declare no conflict of interest.

REFERENCES

1. Bergmeir, C., & Benítez, J. M. (2012). Neural networks in R Using the Stuttgart neural network simulator: RSNNS. *Journal of Statistical Software*, 46(7), 1–26. https://doi.org/10.18637/jss.v046.i07.

2. Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: forecasting and control*. Holden-Day, San Francisco.

3. Brockwell, P., & Davis, R. (1991). *Time series: theory and methods* (2nd ed.). Springer, New York. https://doi.org/10.1007/978-1-4419-0320-4.

4. Brown, R. G. (1959). *Statistical forecasting for inventory control*. McGraw-Hill, New York.

5. Conway, A. J., Macpherson, K. P., & Brown, J. C. (1998). Delayed time series predictions with neural networks. *Neurocomputing*, 18(1–3), 81–89. https://doi.org/10.1016/S0925-2312(97)00070-2.

6. Gardner, E. S. (2006). Exponential smoothing: the state of the art – part II. *International Journal of Forecasting*, 22(4), 637–666. https://doi.org/10.1016/j.ijforecast.2006.03.005.

7. Gately, E. (1995). *Neural networks for financial forecasting*. New York, John

Wiley & Sons.

8. Holt, C. (1957). Forecasting trends and seasonal by exponentially weighted averages. *International Journal of Forecasting*, 20(1), 5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015.

9. Hoptroff, R. G. (1993). The principles and practice of time series forecasting and business modelling using neural nets. *Neural Computing & Applications*, 1, 59–66. https://doi.org/10.1007/BF01411375.

10. Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing the state space approach*. Springer, Germany. https://doi.org/10.1007/978-3-540-71918-2.

11. Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 27(3), 1–22. https://doi.org/10.18637/jss.v027.i03.

12. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). OTexts, Melbourne. Available at: https://otexts.com/fpp2.

13. Khashei, M., Bijari, M., & Ardali, G. (2012). Hybridization of autoregressive integrated moving average (ARIMA) with probabilistic neural networks (PNNs). *Computers & Industrial Engineering*. 63(1), 37–45. https://doi.org/10.1016/j.cie.2012.01.017.

14. Kmytiuk, T., & Majore, G. (2021). Time series forecasting of agricultural product prices using Elman and Jordan recurrent neural networks. *Neuro-Fuzzy Modeling Techniques in Economics*, 10, 67–85. http://doi.org/10.33111/nfmte.2021.067.

15. Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer, New York. https://doi.org/10.1007/978-0-387-71887-3.

16. Kotu, V., & Deshpande, B. (2019). Chapter 12 – Time Series Forecasting. In *Data Science* (2nd ed.), (pp. 395–445). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-814761-0.00 012-5.

17. Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297–303. https://doi.org/10.1093/biomet/65.2.297.

18. Long, J. D., & Teetor, P. (2019). *R cookbook: proven recipes for data analysis, statistics, and graphics* (2nd ed.). O'Reilly Media, USA. Available at: https://rc2e.com.

19. Makridakis, S. G., Wheelwright, S. C., & McGee, V. E. (1998). *Forecasting: Methods and Applications* (3d ed.). Wiley, New York.

20. Mbuli, N., Mathonsi, M., Seitshiro, M., & Pretorius, J. H. C. (2020). Decomposition forecasting methods: a review of applications in power systems, *Energy Reports*, 6(9), 298–306. https://doi.org/10.1016/j.egyr.2020.11.238.

21. Bowerman, B., & O'Connell, R. (1993). *Forecasting and time series: an applied approach*, 3rd ed. South-Western College Pub.

22. Official statistics of Latvia (2005–2021). *Average retail prices of selected commodity (euro per 1 kg, if other – specified) 2005M01 – 2021M12*. Available at: https://data.stat.gov.lv/pxweb/en/OSP_PUB/START__VEK__PC__PCC/PCC010m.

23. Rudenko, O., Bezsonov, O., & Romanyk, O. (2019). Neural network time series prediction based on multilayer perceptron. *Development Management*, 17(1), 23–34. https://doi.org/10.21511/dm.5(1).2019.03.

24. Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janosky, T. A., & Kamaev, V. A. (2013). A survey of forecast error measures. *World Applied Sciences Journal*, 24, 171–176. https://doi.org/10.5829/idosi.wasj.2013.24.itmies.80032.

25. Shumway, R. H., & Stoffer, D. S. (2011). *Time series analysis and its applications. With R examples*, 3 rd ed. Springer. https://doi.org/10.1007/978-1-4419-7865-3.

26. Shumway, R. H., & Stoffer, D. S. (2000). Time series regression and ARIMA models. In *Time series analysis and its applications. Springer texts in statistics*. Springer, New York. https://doi.org/10.1007/978-1-4757-3261-0_2.

27. Sun, F., Meng, X., Zhang, Y., Wang, Y., Jiang, H., & Liu, P. (2023). Agricultural product price forecasting methods: a review. *Agriculture*, 13(9), 1671. https://doi.org/10.3390/agriculture13091671.

28. Vandeput, N. (2021). *Data science for supply chain forecasting*, 2nd ed. Berlin, Boston, De Gruyter. Available at: https://www.researchgate.net/publication/350440225.

29. Wei, William W. S. (2006). *Time series analysis: univariate and multivariate methods*, 2nd ed. Pearson Addison Wesley. Available at: https://www.researchgate.net/publication/220693197.

30. Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3), 324–342. https://doi.org/10.1287/mnsc.6.3.324.

31. Zhang, P., Patuwo, E., & Hu, M. (1998). Forecasting with artificial neural networks: the state of the art. *International Journal of Forecasting*, 14(1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7.

Citation:

Стиль – ДСТУ:

Kmytiuk T., Majore G., Bilyk T. Time series forecasting of price of the agricultural products using data science. *Agricultural and Resource Economics*. 2024. Vol. 10. No. 3. Pp. 5–33. https://doi.org/10.51599/are.2024.10.03.01.

Style – APA:

Kmytiuk, T., Majore, G., & Bilyk, T. (2024). Time series forecasting of price of the agricultural products using data science. *Agricultural and Resource Economics*, 10(3), 5–33. https://doi.org/10.51599/are.2024.10.03.01.