Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Predictive accuracy of time series models applied to economic data: the European countries retail trade

S. Lima, A. M. Gonçalves & M. Costa

To cite this article: S. Lima, A. M. Gonçalves & M. Costa (25 Jul 2023): Predictive accuracy of time series models applied to economic data: the European countries retail trade, Journal of Applied Statistics, DOI: 10.1080/02664763.2023.2238249

To link to this article: https://doi.org/10.1080/02664763.2023.2238249

JOURNAL OF APPLIED STATISTICS
https://doi.org/10.1080/02664763.2023.2238249

Taylor & Francis
Taylor & Francis Group

Predictive accuracy of time series models applied to economic data: the European countries retail trade

S. Lima[a], A. M. Gonçalves [b] and M. Costa [c]

[a]MEtRICs Research Center, University of Minho, Guimaraes, Portugal; [b]DMAT – Department of Mathematics & CMAT – Center of Mathematics, University of Minho, Guimaraes, Portugal; [c]ESTGA – Águeda School of Technology and Management & CIDMA – Center for Research and Development in Mathematics and Applications, University of Aveiro, Aveiro, Portugal

ABSTRACT

Modeling and accurately forecasting trend and seasonal patterns of a time series is a crucial activity in economics. The main propose of this study is to evaluate and compare the performance of three traditional forecasting methods, namely the ARIMA models and their extensions, the classical decomposition time series associated with multiple linear regression models with correlated errors, and the Holt–Winters method. These methodologies are applied to retail time series from seven different European countries that present strong trend and seasonal fluctuations. In general, the results indicate that all the forecasting models somehow follow the seasonal pattern exhibited in the data. Based on mean squared error (MSE), root mean squared error (RMSE), mean absolute percentage error (MAPE), mean absolute scaled error (MASE) and U-Theil statistic, the results demonstrate the superiority of the ARIMA model over the other two forecasting approaches. Holt–Winters method also produces accurate forecasts, so it is considered a viable alternative to ARIMA. The performance of the forecasting methods in terms of coverage rates matches the results for accuracy measures.

ARTICLE HISTORY
Received 26 July 2022
Accepted 1 June 2023

KEYWORDS
Time series forecasting; retail trade forecasting; linear models; Holt–Winters; forecast accuracy

1. Introduction

Forecasting methods are a key tool in decision-making processes in many areas, such as Economics, Management, Finance, or Environment, and over the past several decades much effort has been devoted to the development and improvement of time series forecasting models.

In today's competitive global economy, accurate forecasting is crucial for profitable retail operations, since it supports most of the strategic planning decisions of any retail business, directly affecting revenue and competitive position [33]. Retail time series often exhibit strong trend and seasonal patterns. How to best model and forecast these patterns has been a long-standing issue in time series analysis.

There are several different approaches to deal with trend and seasonal time series, which can be divided into linear and nonlinear models. The available traditional statistical

CONTACT M. Costa marco@ua.pt

approaches include time series decomposition, exponential smoothing, time series regression, and autoregressive integrated moving average (ARIMA) models. These are linear models, in which predictions of future values are constrained functions of past observations. Because of their relative simplicity in terms of understanding and implementation, linear models have been widely used for time series forecasting [36]. An extensive review of the existing forecasting methods can be found in [15].

Although several comparative studies of forecasting models have been conducted in the literature, the findings are mixed as to the most suitable approach for accurate retail series forecasting.

In turn, [19,37] investigated the use of neural networks in forecasting aggregate retail sales, and both teams of researchers have concluded that the overall out-of-sample forecasting performance of neural networks does not outperform the traditional ARIMA models without appropriate data preprocessing. Aras *et al.* [5] also found that neural networks do not outperform the traditional forecasting techniques. In a global perspective, [22] evaluated eight widely used machine learning (nonlinear) models versus eight traditional statistical ones, including ARIMA and Holt–Winters models, all applied to a set of 3003 time series of M3. The authors found that traditional statistical methods are more accurate than nonlinear models, and that their computational requirements are considerably lower than those of machine learning methods. In fact, many observed time series exhibit non-linear characteristics, but nonlinear models do not necessarily produce better out-of-sample forecasts than linear models [11,23].

In the food retail segment, [32] compared ARIMA and Holt–Winters models for predicting demand data from a group of perishable dairy products, having concluded that Holt–Winters outperforms MAPE and Theil's U-statistic, has better adjustment, and captures the linear behavior of the series. Subsequently, [33] assessed the accuracy of demand forecasting between those two linear forecasting methods and two nonlinear forecasting models based on natural computing approaches. In general, the results showed that all the forecasting models somehow follow the seasonal pattern exhibited by the data, although nonlinear approaches performed better. Additionally, nonlinear methods also achieved more accurate results. The performance of the models was also tested by MAPE and Theil's U-statistic. Arunraj and Ahrens [6] developed a seasonal autoregressive integrated moving average with external variables (SARIMAX) model to forecast daily sales of a perishable food. The results showed that SARIMAX models yield better forecasts compared to seasonal naïve forecasting, traditional SARIMA, and multilayer perceptron neural network models.

Suhartono *et al.* [31] also proposed a SARIMAX model to forecast clothing monthly sales, which yielded better results compared to the SARIMA model. Ramos *et al.* [28] compared the forecasting performance of state space models and ARIMA models through a case study of retail sales of five categories of women's footwear. The results showed that the forecasting performance of these two models presented no significant difference via RMSE, MAE, and MAPE for both one-step and multi-step forecasts. Both models produced coverage probabilities that were close to the nominal rates.

Pillo *et al.* [25] introduced the support vector machine (SVM) in sales forecasting, establishing a comparison with some traditional statistical methods, namely ARIMA, simple exponential smoothing, and Holt–Winters models. Based on the mean squared error (MSE), the authors concluded that SVM provides better forecasts than the other applied

methodologies. Aye *et al.* [7] evaluated the performance of 26 forecasting models, including ARIMA and Holt–Winters, in forecasting aggregate seasonal retail sales. The authors underlined the difficulty in identifying a specific model as the best one for forecasting aggregate retail sales. Finally, [18] verified the accuracy of models in predicting revenue in the service sector based on 6 criteria to determine if the use of certain criteria could lead to the adoption of particular models.

In some cases, combining different models can increase the chance to capture different patterns in the data and thus improve forecasting performance. Several empirical studies have already suggested that combining several different models can often improve forecasting accuracy over the individual model [1,5,7,24]. Furthermore, the combined model is more robust regarding possible structure changes in the data [36].

An extended review of the research literature on the retail forecasting field is available in [16].

It follows from the foregoing that all commonly used forecasting models (linear or nonlinear) have their own characteristics, strengths and weaknesses [7]. In fact, different models capture different aspects of the series, and therefore none of them was identified as the universal model that fits every forecasting situation. The purpose of this work is to compare the forecasting performance of three traditional methods, namely the ARIMA models and their extensions, the classical decomposition time series associated with multiple linear regression models, and the exponential smoothing methods. These methods are selected due to their ability to model trend and seasonal fluctuations present in economic data, particularly retail sales data. The nonlinear approaches were not evaluated since the claims of their superiority were found to be exaggerated. According to [24], linear models should be the preferred ones if they are able to efficiently capture the underlying data-generating process. Moreover, linear models have the important practical advantage of easy interpretation and implementation in addition to their simplicity and low cost [33].

The remainder of the paper is organized as follows. The next section describes the time series used in the study. Section 3 describes the seasonal ARIMA model, Holt–Winters method, and multiple linear regression approach used for retail series forecast. The performance measures selected to evaluate and compare the accuracy of the forecasting models are also presented. The empirical results obtained in the research study are discussed in Section 4 based on a significant dataset of seven time series of European countries. The last section offers the concluding remarks.

2. Motivation for the analysis and data description

The data was collected from the Statistical Office of the European Union (Eurostat [30]), which provides official, harmonized statistics in the European Union and the euro area, offering a comparable, reliable and objective portrayal of European society and economy.

The trade sector is a key sector in the European economy. According to Eurostat, in 2015, this sector employed around 33 million people and represented 9.9% of the European Union's total gross value added. In terms of turnover, in 2016 the sector produced around 9.9 trillion euros, of which 57.8% corresponded to wholesale trade (−1% compared to 2015), 28.9% to retail trade (as in 2015), and the remaining 13.3% to automotive trade (+1%).

Retail trade is a dynamic and complex sector encompassing different types of companies whose structure reflects the cultural characteristics of the society in which it operates, bearing the impact of sociological, economic and technological evolution. Thus, from an economic point of view, it is paramount to study the retail trade sector to assess the performance of any national economy. In fact, in addition to representing almost 1/3 of the turnover of the commerce sector in 2015, retail trade held 58.4% of commercial companies and employed 8.7% of the EU's working population, which corresponds to approximately 18.8 million jobs.

In 2016, retail trade in Europe amounted approximately to 2.9 trillion euros. In terms of turnover, the most important retail markets in Europe were Germany (roughly 537.5 billion euros), the United Kingdom (roughly 480.3 billion euros), and France (roughly 440.9 billion euros). In addition to these countries, Italy and Spain also featured prominently, presenting values in the range of 300 and 200 billion euros, respectively. All these major markets, with the exception of Germany (−4.9%) and the United Kingdom (−7%), recorded an increase in retail turnover in 2016.

The economic situation caused by the 2008 crisis predictably impacted the trade trade sector, having influenced indicators such as turnover. In fact, in 2009 there was a reduction of more than 5% in European turnover, followed by an improvement of equal dimension in 2010. Thereafter, the turnover did not reach such low values again and presented a very positive evolution, having registered minimal decreases (−0.9%) only in 2013 and 2016.

In this study, time series for seven European countries are analyzed. These seven countries include Portugal (PT) and their main trading partners according to the Contemporary Portugal Database (PORDATA), namely Germany (DE), Spain (ES), France (FR), Italy (IT), the Netherlands (NL), and the United Kingdom (UK). Moreover, this group of countries has significant economic relations with each other, primarily because they share the European area and because of their geographical proximity. The variable analyzed is TOVT, which corresponds to indexes of total turnover in the context of retail trade. The objective of the turnover index is to show the development of the market for goods and services. Turnover comprises the totals invoiced by the observation unit during the reference period, and this corresponds to market sales of goods or services supplied to third parties. Turnover also includes all other charges (transport, packaging, etc.) passed on to the customer, even if these charges are listed separately in the invoice. Turnover excludes value added tax (VAT) and other similar deductible taxes directly linked to turnover, as well as all duties and taxes on the goods or services invoiced by the unit. Note that these indexes should be compared with the base year, in this case 2015, which corresponds to the index 100. The dataset used concerns the period from January 2000 to February 2018. Figure 1 presents the time series of the seven countries under analysis.

A seasonal variation is detected in an exploratory analysis of the observed values of TOVT in the time series these seven countries. The Christmas season has a marked influence on the retail trade. Every December, retail trade figures spike upwards and then contractions occur in January. Some of the countries under study present similar trends, namely France and the United Kingdom, showing rising trends over time. Portugal and Spain show decreasing trends between 2009 and 2014, thus corroborating the above-mentioned negative effect of the 2008 economic crisis.

Table 1 summarizes the descriptive statistics for the monthly measurements of the TOVT variable in the seven European countries under study in 2009–2014.

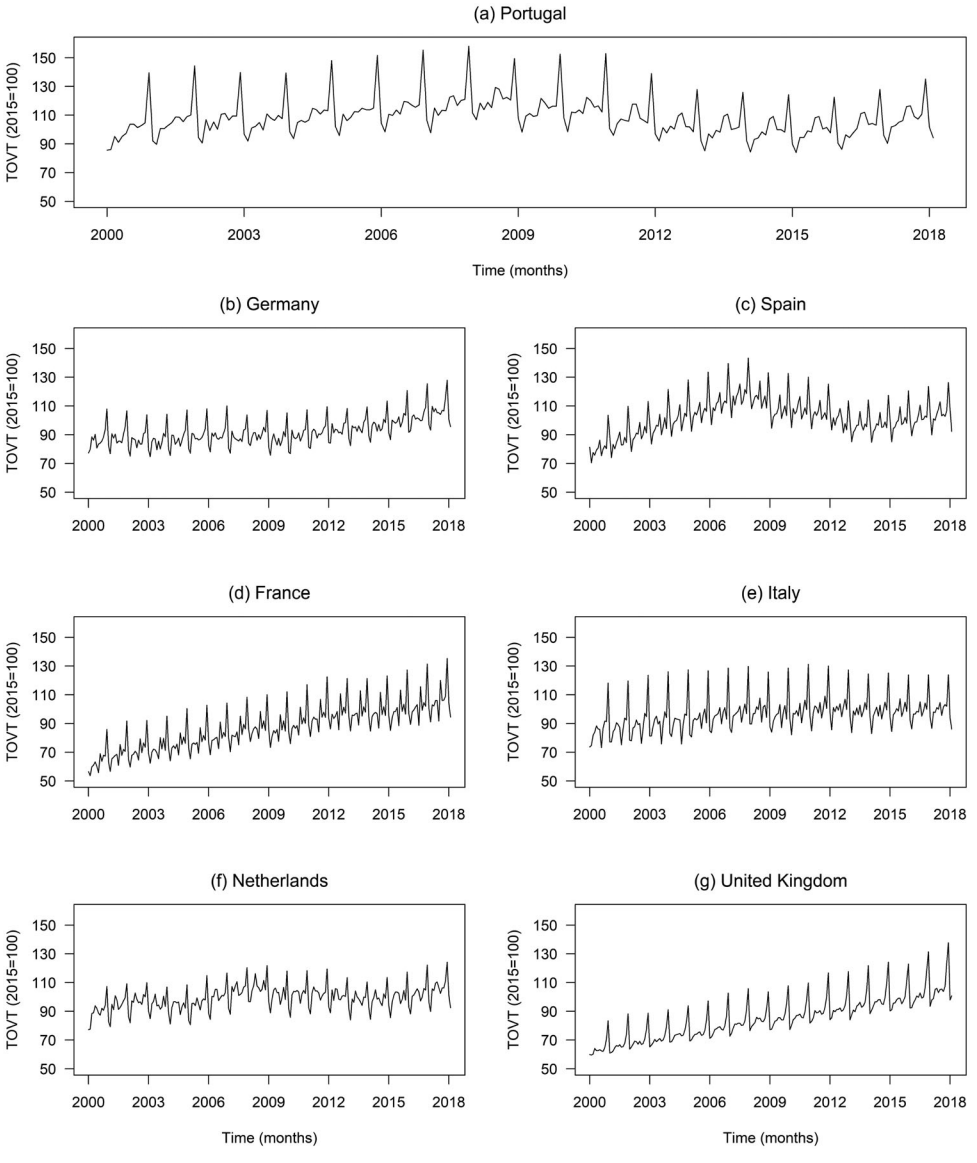

Figure 1. Index of total turnover in the context of retail trade for the seven countries under study between January 2000 and February 2018: (a) Portugal, (b) Germany, (c) Spain, (d) France, (e) Italy, (f) Netherlands, (g) United Kingdom. The base year is 2015 (TOVT = 100).

The Netherlands, Italy, and Germany present the lowest TOVT dispersion, with some constant periods during the period under observation, as graphical analysis suggests. Portugal and Spain present mean values slightly higher than 100, which means that the turnover is, on average, higher than that recorded in 2015. Regarding the countries with increasing trends, the mean of TOVT is widely lower than 100 due to the number of years registered before 2015, when the turnover was lower.

Table 1. Descriptive statistics of TOVT in seven European countries.

Country	Range	Mean	Sd	1st quartile	3rd quartile
Portugal	84.00–158.10	108.58	13.41	99.93	113.80
Germany	74.70–127.90	92.06	9.33	86.13	97.08
Spain	70.50–143.40	102.17	12.50	94.38	108.80
France	53.70–135.30	86.94	15.60	75.43	96.80
Italy	73.20–131.30	96.45	11.48	89.80	100.48
Netherlands	77.30–124.20	99.27	8.39	94.10	104.43
United Kingdom	59.60–137.70	85.49	14.93	73.70	95.30

3. Methods

3.1. Seasonal ARIMA models

The seasonal autoregressive integrated moving average (SARIMA) models, introduced by Box and Jenkins [10], are one of the most versatile linear models for forecasting seasonal time series, capable of representing both stationary and non-stationary data. These models are based on identifying the structure of the autocorrelations inherent to time data, describing the series as a linear combination of its own seasonal and nonseasonal lagged values and errors [5]. As most seasonal time series exhibit trends and/or seasonal variations, both seasonal and nonseasonal differencing are often used to stabilize the time series. In fact, in many important application areas, like engineering, economics, and environment, stationarity is really rare, and so the introduction of differencing to deal with nonstationarity was particularly important because it allowed applying research developed for stationary time series to nonstationary series [20].

There is a huge variety of SARIMA models. The general SARIMA model can be expressed as [37]

$$\Phi_p(B)N_P(B^s)(1-B)^d(1-B^s)^D Y_t = \Theta_q(B)H_Q(B^s)\epsilon_t, \tag{1}$$

with $\Phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$, $N_P(B^s) = 1 - \nu_1 B^s - \cdots - \nu_P B^{Ps}$, $\Theta_q(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$, $H_Q(B^s) = 1 + \eta_1 B^s + \cdots + \eta_Q B^{Qs}$, where s is the seasonal length (e.g. $s = 12$ for monthly data), B is the backshift operator defined by $B^k Y_t = Y_{t-k}$, $\Phi_p(B)$ and $\Theta_q(B)$ are the regular autoregressive and moving average polynomials of orders p and q, respectively, $N_P(B^s)$ and $H_Q(B^s)$ are the seasonal autoregressive and moving average polynomials of orders P and Q, respectively, and ϵ_t is a sequence of white noises with zero mean and constant variance σ^2. $(1-B)^d$ and $(1-B^s)^D$ are the nonseasonal and seasonal differencing operators, respectively. The roots of the polynomials $\Phi_p(B)$, $\Theta_q(B)$, $N_P(B^s)$, and $H_Q(B^s)$ should lie outside a unit circle to ensure causality and invertibility [28]. Model (1) is often referred to as the SARIMA$(p,d,q)(P,D,Q)_s$ model.

The selection of an appropriate SARIMA model is based on the Box-Jenkins methodology, a three-step iterative modeling approach consisting of model identification, parameter estimation, and diagnostic checking [20]. After identifying the model, the associated parameters are estimated and the residuals are obtained. It is important to check the residuals structure, since, for satisfactory models, the residuals should resemble independent and identically distributed (i.i.d.) or white noise process [35]. These three stages of the modeling process are typically repeated several times until an adequate model is selected.

In order to make a careful choice, *neighbor* models must be explored. The choice between two or more SARIMA models was based on the Akaike Information Criterion (AIC). The general formulation of the AIC can be expressed by $\text{AIC} = -2\log(L) + 2m$, where L is the maximum likelihood of the model and m is the number of estimated parameters [2]. The model that minimizes the AIC is considered to be the most appropriate one. However, if the differences between the models' AIC were residual (≤ 2), the most parsimonious model among all the representative ones – i.e. the one with the lowest number of parameters – should be selected as the final model for forecasting.

After model selection, point forecasts are readily calculated by replacing parameters with their estimates, and errors with the available residuals. Recursions can be used to obtain the forecasts for h steps ahead [20]. The forecast intervals for SARIMA models are based on the σ_h that denotes the standard deviation of the h-step ahead forecast errors. Assuming that the residuals are uncorrelated and normally distributed, the $(1-\alpha)100\%$ prediction interval for the h-step ahead forecast is

$$\left]\hat{Y}_{t+h|t} - z_{1-\alpha/2}\hat{\sigma}_h, \hat{Y}_{t+h|t} + z_{1-\alpha/2}\hat{\sigma}_h\right[, \tag{2}$$

where $\hat{y}_{t+h|t}$ is the h-step ahead point forecast, z is the appropriate quantile for the standard Normal distribution, $1-\alpha$ is the confidence level of the interval, and $\hat{\sigma}_h$ is the estimated standard deviation of the h-step ahead forecast errors. For a 95% prediction interval, z must be replaced with 1.960.

3.2. Exponential smoothing methods

Exponential smoothing refers to a set of methods that, in a versatile way, can be used to smooth and forecast a time series without the need to fit a parametric model. These methods belong to a class of local models that automatically adapt their parameters to the data during the estimation procedure, and therefore implicitly account for (slow) structural changes in level, trend, and seasonal patterns. The exponential smoothing forecasts are weighted combinations of past observations, with the weights decreasing exponentially as the observations come from further in the past – the smallest weights are associated with the oldest observations [17].

If the data have no trend or seasonal patterns, then simple exponential smoothing is appropriate. On the other hand, if the data exhibit a linear trend, then Holt's linear model should be used. The Holt–Winters (HW) method is an extension of the Holt's method, and is applied whenever data behavior is trendy and seasonal. Seasonality can be modeled in an additive or multiplicative way, depending on the oscillatory movement along the time period under observation. The additive version should be considered whenever the seasonal pattern of a series presents constant amplitude over time, while the multiplicative version is preferred when the amplitude of the seasonal pattern varies with the series level. In both versions, forecasts will depend on level, trend, and seasonal coefficient.

The Holt–Winters model is based on three smoothing equations: one for the level, one for trend, and one for seasonality. The recursive equations for both the multiplicative and the additive HW methods, with $h_s^+ = [(h-1) \bmod s] + 1$, are presented in Table 2, where Y_t is the observed data at time t, s is the length of seasonality (number of months in a

Table 2. Holt–Winters method recursive equations.

Additive HW method	Multiplicative HW method
Level: $l_t = \alpha(Y_t - s_{t-s}) + (1-\alpha)(l_{t-1} + b_{t-1})$	Level: $l_t = \alpha \frac{Y_t}{s_{t-s}} + (1-\alpha)(l_{t-1} + b_{t-1})$
Trend: $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$	Trend: $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$
Seasonal: $s_t = \gamma(Y_t - l_t) + (1-\gamma)s_{t-s}$	Seasonal: $s_t = \gamma \frac{Y_t}{l_t} + (1-\gamma)s_{t-s}$
Forecast: $\hat{Y}_{t+h\|t} = l_t + hb_t + s_{t-s+h_s^+}$	Forecast: $\hat{Y}_{t+h\|t} = (l_t + hb_t)s_{t-s+h_s^+}$

season), $h = 1, 2, \ldots$ is the forecast horizon, and $\theta = (\alpha, \beta, \gamma)^{\mathrm{T}}$ is the vector of smoothing parameters for level, trend and seasonality, respectively [17].

The initial states $l_1, b_1, s_{t-s}, \ldots, s_1$ and the smoothing parameters α, β, γ are estimated from the observed data using computer software. The smoothing parameters are constrained between 0 and 1, so that the equations can be interpreted as weighted averages. Simple exponential smoothing and Holt's method are derived from the above equations considering that the corresponding exponential parameters β and γ are to be set to zero [13].

The exponential smoothing methods are algorithms which ignore the error component and consequently only generate point forecasts. However, the error can be added to the model both in an additive or multiplicative way, which is extremely relevant for calculating the prediction intervals, [8]. In this study, additive errors were considered. After estimating both HW methods, the one with the lowest one-step mean squared error (in-sample) was selected.

Computing prediction intervals is an important part of the forecasting process, intended to indicate the likely uncertainty in point forecasts. However, the Holt–Winters methods do not provide good prediction intervals. In fact, a lot of different formulae have been proposed for obtaining prediction intervals, but several studies have shown that proposed intervals tend to be too narrow or unreasonably wide [9]. For this study, the forecast intervals are based on the mean squared error (MSE) that denotes the variance of the h-step ahead forecast errors [21]. The $(1-\alpha)100\%$ empirical prediction interval (if the normally assumption is verified) for the h-step ahead forecast when the time series presents a seasonal component with period s is

$$\left]\hat{Y}_{t+h|t} - z_{1-\alpha/2}\sqrt{\mathrm{MSE}_h}, \hat{Y}_{t+h|t} + z_{1-\alpha/2}\sqrt{\mathrm{MSE}_h}\right[, \tag{3}$$

where $\hat{Y}_{t+h|t}$ is the h-step ahead point forecast, $z_{1-\alpha/2}$ is the appropriate quantile for the standard Gaussian distribution, and $\mathrm{MSE}_h = \frac{1}{n-h-s+1}\sum_{t=h+s}^{n}[\epsilon_t^{(h)}]^2$ denotes the variance of the h-step ahead errors.

3.3. Multiple linear regression models with autocorrelated errors

Multiple linear regression can also be used to model time series with trend and seasonal patterns. The trend component is deterministic and can be modeled by polynomials of time t of some low orders. In the simplest case, the trend is modeled as a linear function of time, which corresponds to a linear regression model with a single explanatory variable

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t, \tag{4}$$

where β_0 is the level of the series when $t = 0$, β_1 is the amount of change in the time series associated with a one unit increase in time, t is the time variable (e.g. 1 to 90 for 90 equally spaced observations), and ϵ_t is the random error. However, some time series present a non-linear behavior that is not fully explained by the above model. In these cases, a polynomial trend from a higher order should be fitted to the data, that is

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p + \epsilon_t, \tag{5}$$

where t is the time variable, and p is the order of the polynomial that describes the trend component. For quadratic or cubic trends, p should be set to 2 or 3, respectively.

Economic time series are frequently influenced by real-world events (e.g. the economic crisis of 2008). In time series methodology, the impact of an event can be analyzed through segmented regression analysis, in which the time series is partitioned in specific points in time (change-points) – also known as interrupted time series. In the simplest case, there is only one event, and linear regressions are estimated for the two parts of the time series (the pre- and post-event segments), using two distinct parameters for each part: the level (intercept) and the trend (slope). A change in these parameters represents an effect of the event: a significant change in the level of a series indicates an immediate change, and a change in trend reflects a more gradual change in the outcome [34]. This model can be expressed as

$$Y_t = \beta_0 + \beta_1 t + \beta_2 \times Event_t + \beta_3 t \times Event_t + \epsilon_t, \tag{6}$$

where t is the time variable, β_0 and β_1 are the intercept and the slope for the pre-event trend, and β_2 and β_3 represent the post-event changes in the intercept and the slope. Therefore, the sum of the pre-event intercept (β_0) and its change (β_2) results in the post-event intercept, while the sum of β_1 with β_3 corresponds to the post-event slope. The dummy variable $event_t$ codes represent for whether or not each time point occurred before or after the event (0 for all points prior to the event; 1 for all points after, including the change-point). According to the time series under study, polynomial trends can also be included, as well as more change-points (events), which will increase the complexity of the model. When two or more events are considered, they are mutually exclusive, i.e. the first event does not extend its effect to the next event.

The seasonal component can be modeled either by seasonal dummy variables in a qualitative way or harmonic seasonal models that use trigonometric functions to describe the pattern of fluctuations seen across periods. In this study, the seasonal component was modeled by harmonic seasonal models. In fact, seasonal effects often vary in a smooth, continuous way, and instead of estimating a discrete intercept for each season, this approach can provide a more realistic model of seasonal change [14]. Also, using sine/cosine waves as independent variables is advantageous because, in most cases, several of these variables are not statistically significant, which allows for the adjustment of a more parsimonious model [4]. The harmonic seasonal model is formally expressed as [14]

$$Y_t = T_t + \sum_{i=1}^{s/2} (\alpha_i \cos(2\pi it/s) + \beta_i \sin(2\pi it/s)) + \epsilon_t, \tag{7}$$

where T_t is the trend model, α_i and β_i are the unknown parameters of interest, s is the number of seasons within the time period (e.g. 12 months for a yearly period), i is an index

ranging from 1 to $s/2$, and t is the time variable. When slopes vary over time, [4] suggest that a new term is added to the harmonic seasonal model. For these cases, the model can be written as

$$Y_t = T_t + \sum_{i=1}^{s/2}\left[\alpha_i \cos\left(\frac{2\pi it}{s}\right) + \beta_i \sin\left(\frac{2\pi it}{s}\right)\right] + \sum_{i=1}^{s/2}\left[\gamma_i t \cos\left(\frac{2\pi it}{s}\right) + \delta_i t \sin\left(\frac{2\pi it}{s}\right)\right] + \epsilon_t, \tag{8}$$

where T_t is the trend model, the first summation describes the seasonal variation, and the second summation describes the variation of slopes over the time period.

However, time series often exhibit strong autocorrelation which often manifests in correlated residuals after a regression model has been fitted. This violates the standard assumption of independent (i.e. uncorrelated) errors. Thus, with correlated residuals, the standard deviations of the coefficients given by the linear model are not correct. This, of course, may lead to a wrong decision given by the t-test. Also, in general, the Ordinary Least Squares (OLS) estimators may lose their optimality properties if the residuals are not independent [4]. To overcome the autocorrelation problem, [4] propose a linear model of the form

$$Y_t = \beta_0 + \beta_1 X_t^1 + \beta_2 X_t^2 + \cdots + \beta_p X_t^p + \epsilon_t, \tag{9}$$

where the error component, ϵ_t, follows an autoregressive Gaussian stationary process of order k, AR(k), that is,

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \cdots + \phi_k \epsilon_{t-k} + a_t,$$

or, alternatively,

$$\Phi(B)\epsilon_t = a_t,$$

where a_t is a sequence of zero mean, uncorrelated normal variables, and $\Phi(B)$ is the autoregressive polynomial of order k.

When the residuals follow any autoregressive stationary process, Alpuim and El-Shaarawi [3] show that, under certain conditions of the design matrix $\mathbf{X}$, the OLS and Maximum Likelihood (ML) estimators are asymptotically equivalent and fully efficient. For this to be so, the set of p independent variables in time t, $\mathbf{X}_t^T = (X_t^1, X_t^2, \ldots, X_t^p)$, must verify a linear recursive relationship of the type

$$\mathbf{X}_t = \Psi \mathbf{X}_{t-1}, \tag{10}$$

where Ψ is a $p \times p$ matrix of constant coefficients. Most of the time varying regressors used in linear models verify this recursive relationship, as in cases of linear and polynomial trends, sin/cosine waves, dummy variables, etc. [4].

Also, for the linear model (9) with the condition (10), Alpuim and El-Shaarawi [3] show that the vector of OLS estimators is asymptotically normal with the variance/covariance matrix given by

$$\mathrm{Var}(\hat{\beta}) = \sigma_a^2 [\Phi(B)\mathbf{X}^T \Phi(B)\mathbf{X}]^{-1}, \tag{11}$$

with $\mathbf{X}$ representing the matrix that contains the whole set of the independent variables, and $\Phi(B)\mathbf{X}$ representing the matrix where each element is obtained by applying the operator

$\Phi(B)$ to the corresponding element of the matrix $\mathbf{X}$. More precisely, the generic element of the matrix $\Phi(B)\mathbf{X}$ is given by

$$X_t^{j*} = \Phi(B)X_t^j = X_t^j - \phi_1 X_{t-1}^j - \cdots - \phi_k X_{t-k}^j,$$

for $j = 1, \ldots, p$ and $t = k+1, \ldots, n$. In practice, the values for the autoregressive coefficients ($\phi_i, i = 1, \ldots, p$) and the variance of the white noise sequence, σ_a^2, are unknown. However, if n is large, we may replace them by consistent estimators, which allows obtaining an asymptotic test for the significance of each variable, based on the normal distribution (z-test) [4].

In harmonic seasonal models, a cosine curve with a certain period should be included or eliminated together with the corresponding sine with the same period, and vice versa, i.e. they should be included or eliminated in pairs. On one hand, this practice ensures that the OLS estimators are optimal and, on the other hand, that the formula for the variances (11) of the estimators can be applied. For the same reason, the time multiplied by the cosine with a certain period should be included or eliminated together with the time multiplied by the corresponding sine. Also, this variable should not be included without the cosine and sine with the same period [4].

After removing the non-significant explanatory variables from the full models, the selection between two or more final models is based on the adjusted coefficient of determination, R_a^2, corrected to take into account the autoregressive parameters (if applicable).

Just like the model needs corrections in the presence of correlated residuals, the prediction intervals should also be updated. The limits of a $(1-\alpha)100\%$ prediction interval for a general linear regression model are

$$\hat{Y}_i - t_{1-\frac{\alpha}{2};n-p-1}\sqrt{\hat{\sigma}^2\left(1 + \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\right)} \tag{12}$$

and

$$\hat{Y}_i + t_{1-\frac{\alpha}{2};n-p-1}\sqrt{\hat{\sigma}^2\left(1 + \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\right)}, \tag{13}$$

where $\hat{Y}_i$ is the point forecast, t is the appropriate quantile for the t distribution with $n-p-1$ degrees of freedom, and $\hat{\sigma}^2$ is the estimated variance of the errors. To obtain the correct prediction intervals, the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ should be replaced by $(\Phi(B)\mathbf{X}^T\Phi(B)\mathbf{X})^{-1}$.

3.4. Performance measures

Error metrics are the traditional way of quantifying the accuracy of a forecast [29]. To evaluate the predictive accuracy of the forecasting methods we applied the MSE (Mean Squared Error), the RMSE (Root Mean Squared Error), the MAPE (Mean Absolute Percentage Error), the MASE (Mean Absolute Scaled Error), and the Theil's U-statistic.

MSE is one of the most commonly used scale-dependent metrics. Based on squared errors, it is defined as

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2, \tag{14}$$

where Y_t represents the actual value, $\hat{Y}_t$ is the point forecast, and n is the sample size. Often, the RMSE $= \sqrt{\text{MSE}}$ is preferred to the MSE as it is on the same scale as the data.

Percentage errors have the advantage of being scale-independent, so they are frequently used to compare forecast performance between different data series. The most frequently used is MAPE

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100. \tag{15}$$

MASE uses scaled errors as an alternative to percentage errors when comparing forecast accuracy across series with different units. For non-seasonal time series, MASE is defined as

$$\text{MASE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{\frac{1}{n-1}\sum_{t=2}^{n}|Y_t - Y_{t-1}|}\right|. \tag{16}$$

For seasonal time series, MASE takes into account the seasonal period s

$$\text{MASE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{\frac{1}{n-s}\sum_{t=s+1}^{n}|Y_t - Y_{t-s}|}\right|. \tag{17}$$

Theil's U statistic allows a relative comparison of forecasting methods with naïve approaches and also squares the errors involved so that large errors are given much more weight than small errors. It is defined as

$$\text{U} - \text{Theil} = \sqrt{\frac{\sum_{t=1}^{n-1}\left(\frac{\hat{Y}_{t+1} - Y_{t+1}}{Y_t}\right)^2}{\sum_{t=1}^{n-1}\left(\frac{Y_{t+1} - Y_t}{Y_t}\right)^2}}. \tag{18}$$

Since there is no universally agreed-upon performance measure that can be applied to every forecasting situation, multiple criteria are therefore often needed to provide a comprehensive assessment of forecasting models [12].

4. Application to retail sales data

The data collected for the seven countries under study was divided into two sets, training data (in-sample data) and testing data (out-of-sample data) to test the accuracy of the three suggested forecasting models. The selected training period was from January 2000 to December 2016 (first 204 observations), and the test period was from January 2017 to February 2018 (last 14 observations).

The time series analysis was carried out using the statistical software R programming language and the packages *stats*, *forecast*, and *Metrics* [27]. Three linear models –SARIMA, Holt–Winters, and multiple linear regression – were built using the in-sample data. Each model's forecasting performance was evaluated in terms of the results obtained from the out-of-sample data that was excluded from the fitting and the model selection process. During the model selection, a 5% significance level was used.

A time series with heteroscedasticity often needs a logarithm transformation (generally, a Box-Cox transformation can be applied). In this work, such transformation was applied to the retail series before using SARIMA and multiple linear regression methodologies.

The models were fitted to the transformed data, and finally the forecasts were scaled back to their original units. However, the process of using a transformation, such as logarithm, and then applying an inverse transformation (as exponentiation) introduces a bias in the forecasts of the mean values. The $e^{\frac{1}{2}\sigma^2}$ correction factor can be used when the residual series of the fitted log-regression model is Gaussian white noise. In general, the distribution of the residuals from the log-regression is often negatively skewed, and in such a case a correction factor can be empirically determined using the mean of the anti-log of the residual series. In this approach, adjusted forecasts can be obtained from

$$e^{\hat{Y}_t} \sum_{t=1}^{n} e^{z_t/n},$$

where $\{\hat{Y}_t : t = 1, 2, \ldots, n\}$ is the predicted series given by the fitted log-regression model, and $\{z_t\}$ is the residual series from this fitted model [14]. In this study, the empirical correction factor was used because of its versatility and applicability to all cases.

To evaluate and compare the forecasting performance of different models, five error measures were selected – mean squared error (MSE), root mean squared error (RMSE), mean absolute percentage error (MAPE), mean absolute scaled error (MASE), and Theil's U-statistic. The coverage rates of the nominal 95% forecast intervals were also analyzed.

4.1. Model selection

As mentioned in Section 3, the selection of an appropriate SARIMA model follows the Box-Jenkins methodology, consisting of model identification, parameter estimation, and diagnostic checking. In the model identification stage, the autocorrelation function (ACF) and the partial autocorrelation function (PACF) are examined to help specify the model orders for both nonseasonal (p, d, q) and seasonal (P, D, Q) parts. Then, the model parameters were estimated iteratively via computer software, using either the method of maximum likelihood or conditional least squares. Diagnostic checking is applied to detect inadequacies in the fitted model and to suggest suitable modifications. In this stage, the significance of the model parameters is analyzed, and the residuals and their autocorrelations are inspected.

Selected models for each of the seven time series are presented in Table 3. Note that this model takes into account the transformed data – we applied a logarithm transformation to stabilize the variance.

The Holt–Winters method was applied in both additive and multiplicative seasonal approaches. In the German, Italian and Dutch cases, the additive model was the selected one, since it presented the lowest one-step MSE in-sample compared to the multiplicative model. In the other cases, the model selected was the multiplicative model (Table 4).

As regards the multiple linear regression, the model selection starts with trend component estimation. As previously mentioned, economic time series are frequently influenced by real-world events and so, when modeling the trend component, it may be necessary to partition the time series considering some change-points.

The general methodology of modeling using linear regression models with correlated errors is explained in detail by modeling the time series of Portugal. This methodology was adopted in the remaining six time series, whose results are presented above.

Table 3. Models' selection for the seven countries under study (with the logarithm transformation).

Portugal	model: SARIMA$(2,1,0)(1,0,1)_{12}$			AIC $= -933.72$	$\hat{\sigma}_a = 0.0211$		
	Parameter	ϕ_1	ϕ_2		ν_1	η_1	
	Estimate	-0.6810	-0.3283		0.9967	-0.5448	
	s.d.	0.0676	0.0665		0.0017	0.0661	
Germany	model: SARIMA$(2,1,1)(0,1,1)_{12}$			AIC $= -903.65$	$\hat{\sigma}_a = 0.0216$		
	Parameter	ϕ_1	ϕ_2	θ_1		η_1	
	Estimate	-0.6687	-0.5167	-0.3182		-0.6963	
	s.d.	0.1006	0.0830	0.1164		0.0598	
Spain	model: SARIMA$(2,1,0)(0,1,1)_{12}$			AIC $= -919.23$	$\hat{\sigma}_a = 0.0210$		
	Parameter	ϕ_1	ϕ_2			η_1	
	Estimate	-0.7339	-0.3786			-0.6376	
	s.d.	0.0683	0.0676			0.0722	
France	model: SARIMA$(2,1,0)(0,1,1)_{12}$			AIC $= -979.38$	$\hat{\sigma}_a = 0.0180$		
	Parameter	ϕ_1	ϕ_2			η_1	
	Estimate	-0.7112	-0.4841			-0.5893	
	s.d.	0.0644	0.0651			0.0683	
Italy	model: SARIMA$(0,1,1)(1,0,1)_{12}$			AIC $= -915.76$	$\hat{\sigma}_a = 0.0225$		
	Parameter			θ_1	ν_1	η_1	
	Estimate			-0.7051	0.9894	-0.2578	
	s.d.			0.0591	0.0047	0.0794	
Netherlands	model: SARIMA$(2,1,0)(0,1,2)_{12}$			AIC $= -920.53$	$\hat{\sigma}_a = 0.0202$		
	Parameter	ϕ_1	ϕ_2			η_1	η_2
	Estimate	-0.8869	-0.6425			-0.5899	-0.2967
	s.d.	0.0585	0.0568			0.0902	0.0868
United Kingdom	model: SARIMA$(0,1,1)(0,1,1)_{12}$			AIC $= -1104.60$	$\hat{\sigma}_a = 0.0129$		
	Parameter			θ_1		η_1	
	Estimate			-0.5942		-0.7162	
	s.d.			0.0597		0.0628	

Based on the visual analysis of the *decompose* output in R applied to the Portuguese case (Figure 2), we identified two possible models for the trend component. The first half of the time series can be easily explained by a linear trend, with a change-point occurring between January 2007 and December 2009. The remainder of the time series presented a quadratic behavior, which can be represented by two different trend models: one with a quadratic trend, or one with two linear trends, considering the occurrence of an event between January 2013 and December 2015. The selection of the change-points is made by comparing the adjustment quality (maximizing R_a^2) of several models that consider all the possible change-points within the defined time interval. For the Portuguese case, we identified two possible events: one in January 2008 and the other in April 2014 (see Figure 2). Note that those events match the beginning and the end of the financial crisis in Portugal. In fact, the European economic crisis began in 2008, although it was felt more intensely the following year, and it began to be overcome in Portugal in 2014.

Thus, the model for the trend component can be represented by

$$T_t = \begin{cases} b_0 + b_1 t, & t < t_1 \\ (b_0 + b_2) + (b_1 + b_3)t + b_4 t^2, & t \geq t_1 \end{cases},$$

Table 4. Initialization of level, trend, seasonal and exponential smoothing parameters of the best HW model for the seven countries under study.

Country model	Portugal mult.	Germany add.	Spain mult.	France mult.	Italy add.	Netherlands add.	United Kingdom mult.
$\hat{\alpha}$	0.3469	0.2087	0.3641	0.3365	0.1798	0.2168	0.3093
$\hat{\beta}$	0.0610	0.0336	0.0749	0.0329	0.0157	0.2238	0.0000
$\hat{\gamma}$	0.6259	0.3024	0.4829	0.5433	0.8463	0.2854	0.4881
$\hat{l}_1$	100.3229	87.2678	80.9233	63.4961	86.4834	89.6482	64.2994
$\hat{b}_1$	0.5067	0.1001	0.4557	0.4436	0.1541	0.3861	0.2271
$\hat{s}_1$	0.8818	−5.5146	1.0563	0.9467	−10.7736	−9.6632	0.9261
$\hat{s}_2$	0.8554	−11.4562	0.8748	0.8419	−10.8153	−13.5924	0.9302
$\hat{s}_3$	0.9579	2.4354	0.9783	0.9663	−3.7694	1.6701	0.9419
$\hat{s}_4$	0.9521	−0.6896	0,9378	0.9777	−1.9278	−2.4174	0.9743
$\hat{s}_5$	0.9685	1.9646	0.9833	0.9908	2.3431	6.7868	0.9916
$\hat{s}_6$	0.9834	−4.2562	1.0165	0.9955	0.8431	3.6868	0.9793
$\hat{s}_7$	1.0252	−3.9729	1.0596	0.8704	−0.7403	−1.4590	0.9793
$\hat{s}_8$	1.0220	−3.3646	0.9205	1.0695	−13.6861	−3.2674	0.9652
$\hat{s}_9$	0.9946	−1.9313	0.9721	0.9818	−0.2986	1.1910	0.9581
$\hat{s}_{10}$	1.0021	0.8187	0.9949	1.0406	4.4014	−0.5715	0.9948
$\hat{s}_{11}$	1.0136	5.9562	0.9673	1.0232	3.9181	2.5410	1.0842
$\hat{s}_{12}$	1.3434	20.0104	1.2387	1.2955	30.5056	15.0951	1.2750
MSE	6.1686	4.3834	5.5033	2.5204	4.4141	6.5603	1.4084

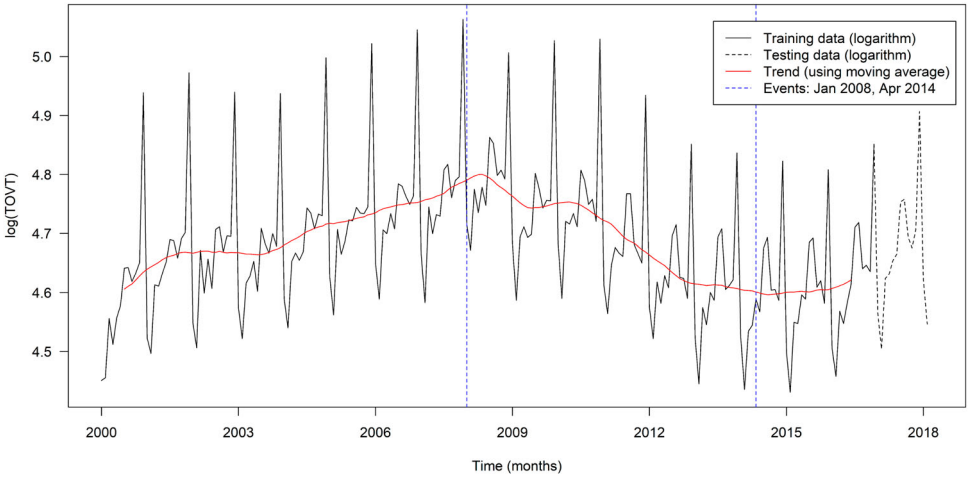

Figure 2. Trend component of the Portuguese time series (with logarithm transformation) using *decompose* in R. Identification of the two possible change-points: January 2008 and April 2014.

if we consider only one event (t_1), or

$$T_t = \begin{cases} b_0 + b_1 t, & t < t_1 \\ (b_0 + b_2) + (b_1 + b_3)t, & t_1 \leq t < t_2 \\ (b_0 + b_2 + b_4) + (b_1 + b_3 + b_5)t, & t \geq t_2 \end{cases},$$

if two events are identified, with t_1 and t_2 representing January 2008 and April 2014, respectively.

After this pre-selection of the trend model, we applied both full models (with trend and seasonal components, Equation (8)) to the Portuguese training data. However, this time

Table 5. Results for the adjustment of the final reduced model with two events to the Portuguese case (with logarithm transformation).

Model: Multiple regression, corrected with AR(3)			$R_a^2 = 0.9781$	$\hat{\sigma} = 0.0208$
Variables	Coeffs	Std. Dev.	*z*-Statistic	*p*-value
Intercept	4.607472	0.012937	356.1515	< 0.0001
Time	0.001739	0.000204	8.5454	< 0.0001
Event1	0.502848	0.036470	13.7878	< 0.0001
Event2	−0.325701	0.028562	−11.4031	< 0.0001
Cos12	0.042203	0.003478	12.1345	< 0.0001
Sin12	−0.053926	0.003572	−15.0984	< 0.0001
Cos6	0.056111	0.002390	23.4814	< 0.0001
Sin6	−0.029016	0.002373	−12.2271	< 0.0001
Cos4	0.072531	0.002585	28.0625	< 0.0001
Sin4	−0.028205	0.002618	−10.7746	< 0.0001
Cos3	0.063109	0.003598	17.5389	< 0.0001
Sin3	0.002070	0.003568	0.5800	0.5619
Cos2.4	0.061775	0.003285	18.8044	< 0.0001
Sin2.4	0.013402	0.003310	4.0487	0.0001
Cos2	0.015813	0.000976	16.1958	< 0.0001
*t*Event1	−0.004794	0.000355	−13.5107	< 0.0001
*t*Cos12	−0.000210	0.000029	−7.1494	< 0.0001
*t*Sin12	−0.000077	0.000030	−2.5452	0.0109
*t*Cos6	−0.000041	0.000020	−2.0221	0.0432
*t*Sin6	0.000158	0.000020	7.7950	< 0.0001
*t*Cos4	0.000009	0.000022	0.3979	0.6907
*t*Sin4	0.000047	0.000022	2.1139	0.0345
*t*Cos3	−0.000128	0.000030	−4.2705	< 0.0001
*t*Sin3	0.000031	0.000030	1.0489	0.2942
*t*Cos2.4	−0.000067	0.000028	−2.4136	0.0158
*t*Sin2.4	0.000042	0.000028	1.5249	0.1273
AR(3)	$\hat{\phi}_1 = 0.1579$	$\hat{\phi}_2 = 0.2704$	$\hat{\phi}_3 = 0.2863$	$\hat{\sigma}_a = 0.0162$

series exhibits strong autocorrelation, which manifests in correlated residuals after a regression model has been fitted. This could bring negative consequences for the model, since it could lead to wrong decision by the *t*-test. To overcome the autocorrelation problem, an autoregressive model of order 3, AR(3), is fitted to the residuals, following the methodology described in Section 3, to determine the new standard deviations of the coefficients, *z*-test statistics, and *p*-values.

The non-significant variables were removed from both models (one by one, or two by two in the case of trigonometric variables), and the final reduced models were compared according to adjustment quality, i.e. maximizing the R_a^2 corrected for autocorrelation. The results for the selected reduced model, which has two change-points and three linear trends, are presented in Table 5.

The methodology illustrated with the data for Portugal was applied to the remaining time series. Some aspects of the modeling are presented in Table 6. With the exception of the time series for Italy, the AR(3) model was the model selected to model the time-correlation structure of the errors. Table 7 presents some results and estimates of some parameters of the regression models fitted to the seven countries under study. The values of the fitted coefficients of determination of the final regression models are high, indicating a good fit to the time series. Regarding the residuals of the regression models, they generally comply with the assumptions of normality, null mean and constant variance of the errors.

As an illustration of model fits and predictions, Figure 3 presents the empirical results of the model fitting, point forecasting and prediction intervals for the Portuguese case,

Table 6. Characterization of the regression models fitted to the logarithm of the retail trade series (with logarithm transformation).

Country	Trend	Events	Error model	Variables removed
Portugal	Three linear trends	Jan 2008, Apr 2014	AR(3)	t:event2; t:cos2
Germany	Two linear trends	Feb 2009	AR(3)	t:cos3; t:sin3; t:cos2.4; t:sin2.4; t:cos2
Spain	Three linear trends	Mar 2008, Sep 2012	AR(3)	event2; cos2; t:cos4; t:sen4; t:cos2
France	Two linear trends	Nov 2008	AR(3)	t:cos3; t:sin3; t:cos2
Italy	Four linear trends	Oct 2008, Apr 2012, Sep 2014	white noise	t:event1
Netherlands	One linear trend and Two quadratic trends	Jul 2002, Apr 2008	AR(3)	t:cos12; t:sin12; t:cos4; t:sen4; t:cos3; t:sin3; t:cos2.4; t:sin2.4; t:cos2
United Kingdom	Two linear trends	Oct 2008	AR(3)	t:cos2.4; t:sin2.4

Table 7. Some results and estimates of the regression models fitted to the time series of the seven European countries under study (R^2_{ac} – adjusted determination coefficient).

Country	Portugal	Germany	Spain	France	Italy	Netherlands	United Kingdom
Intercept	4.607472	4.454630	4.375235	4.151694	4.440673	4.471481	4.151591
time	0.001739	0.000379	0.004421	0.003168	0.001609	0.003656	0.002865
event1	0.502848	−0.207794	0.567825	0129318	−0.051117	0.276732	0.055154
event2	−0.325701	–	–	–	0.310524	0.605970	–
event3	–	–	–	–	0.184447	–	–
$\hat{\phi}_1$	0.1579	−0.0758	0.1412	0.2070	–	−0.1141	0.2546
$\hat{\phi}_2$	0.2704	0.1095	0.1658	0.1966	–	0.1422	0.1787
$\hat{\phi}_3$	0.2863	0.4837	0.2588	0.4422	–	0.6031	0.1732
$\hat{\sigma}_a$	0,0162	0.0174	0.2588	0.0148	0.0186	0.0171	0.0101
R^2_{ac}	0.9781	0.9594	0.9819	0.9920	0.9743	0.9530	0.9957

considering the three forecasting approaches under study. All the forecasting models are able to forecast trend movement and seasonal fluctuations exhibited by the data. According to the five evaluation measures, the most accurate model for explaining the behavior of the training sample is the multiple linear regression model. However, its predictive quality is not the best, since it is easily overcome by the two other forecasting approaches.

4.2. Accuracy measures of forecasting methods

Figure 4 illustrates the empirical results of three different forecasting models for the seven retail series under study. In general, all the forecasting models are capable of forecasting the trend movement and seasonal fluctuations exhibited by the data.

For each retail series, Tables 8 and 9 present the forecasting accuracy measures for in-sample and out-of-sample data, where the smaller values correspond to better forecasting accuracy. It can be observed from Table 8 that the multiple linear regression model outperforms the remaining methods in the training sample, for it is the one that better explains the behavior of TOVT for every country under study. However, a good fit (in-sample) does not necessarily translate into good out-of-sample forecasts [11]. In fact, in the out-of-sample period multiple linear regression was identified as the worst forecasting methodology (Table 9). Although the implemented models have taken into account some autocorrelation present in the data by including an autoregressive component, the regression models are still deterministic and may not be the most appropriate methodology

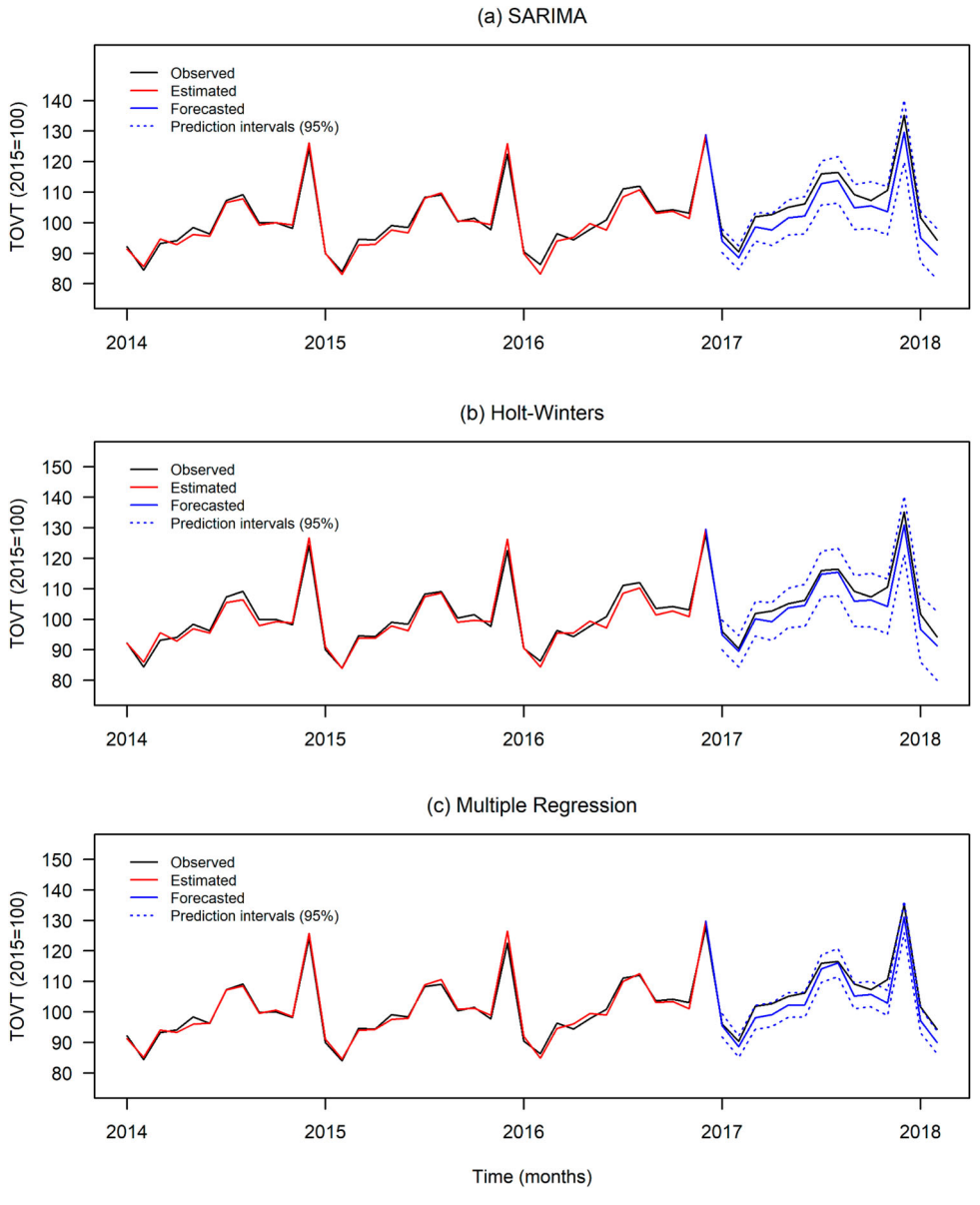

Figure 3. Model fitting and forecasting for the Portuguese case (from January 2014 to February 2018): (a) SARIMA; (b) Holt–Winters; (c) Multiple Regression.

for forecasting economic time series whose trend and seasonal components are constantly changing, [12].

In contrast, both SARIMA and Holt–Winters models performed well. SARIMA forecasted the series more accurately for Germany, Spain, France, Netherlands, and the United Kingdom than Holt–Winters and multiple linear regression models, regardless of the forecast error measure considered. For the remaining series (Portugal and Italy), the Holt–Winters method displays the best performance. Also, as stated by Kolkova [18], one

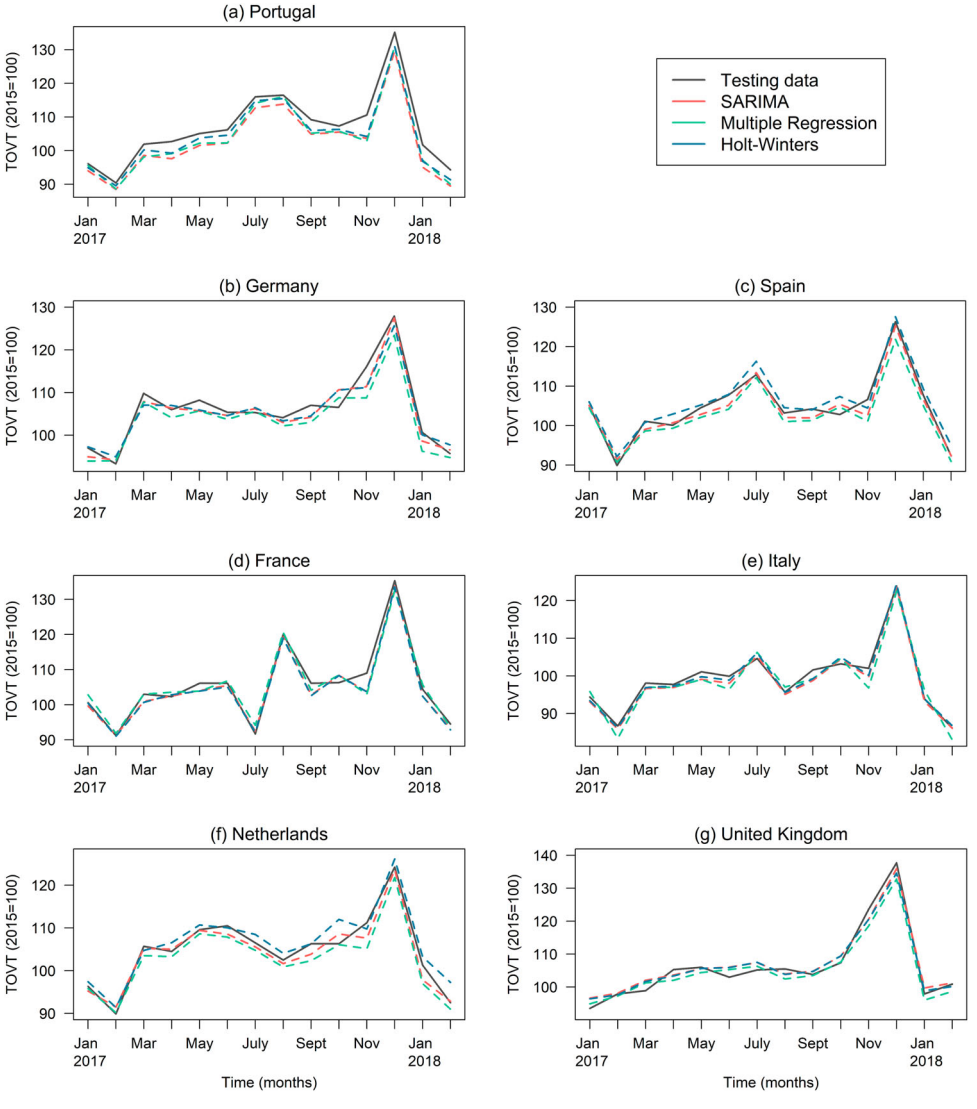

Figure 4. Out-of-sample fixed forecasting comparison for the retail series (between January 2017 and February 2018): (a) Portugal; (b) Germany; (c) Spain; (d) France; (e) Italy; (f) Netherlands; (g) United Kingdom.

accuracy measure could have been enough to select between different models, since the five criteria produced the same conclusions.

Thus, according to the order of accuracy shown in detail in Table 9, starting with the most accurate one, the forecasting models can be classified as follows: (1) SARIMA, (2) Holt–Winters, and (3) multiple linear regression. Similarly, [26] point out the Box-Jenkins as the best traditional methodology to forecast retail time series, with the multiple linear regression model producing the poorest forecasts.

Table 8. Forecast accuracy measures for the in-sample period from January 2000 to December 2016.

Country	Model	MSE	RMSE	MAPE (%)	MASE	U-Theil
Portugal	SARIMA	5.6295	2.3726	1.6541	0.5207	0.1653
	Holt–Winters	6.1686	2.4837	1.7152	0.5693	0.1730
	Regression	**3.3792**	**1.8382**	**1.2996**	**0.4108**	**0.1277**
Germany	SARIMA	3.5726	1.8901	1.6206	0.6070	0.2193
	Holt–Winters	4.3834	2.0937	1.8743	0.6940	0.2412
	Regression	**2.5055**	**1.5829**	**1.4000**	**0.5256**	**0.1803**
Spain	SARIMA	4.4262	2.1039	1.4716	0.3440	0.1877
	Holt–Winters	5.5033	2.3459	1.7175	0.4096	0.2088
	Regression	**2.6503**	**1.6280**	**1.2160**	**0.2838**	**0.1452**
France	SARIMA	2.1863	1.4786	1.3319	0.4068	0.1159
	Holt–Winters	2.5204	1.5876	1.4447	0.4665	0.1238
	Regression	**1.5557**	**1.2473**	**1.1979**	**0.3625**	**0.0967**
Italy	SARIMA	4.5915	2.1428	1.6587	0.7731	0.1541
	Holt–Winters	4.4141	2.1010	1.6224	0.7578	0.1506
	Regression	**2.5749**	**1.6047**	**1.3380**	**0.6133**	**0.1152**
Netherlands	SARIMA	3.8120	1.9524	1.5271	0.4997	0.2109
	Holt–Winters	6.5603	2.5613	2.0494	0.6911	0.2746
	Regression	**2.9233**	**1.7098**	**1.3610**	**0.4435**	**0.1849**
United Kingdom	SARIMA	1.2947	1.1379	0.9401	0.3272	0.1368
	Holt–Winters	1.4084	1.1868	1.0176	0.3550	0.1426
	Regression	**0.8291**	**0.9106**	**0.7858**	**0.2703**	**0.1101**

Table 9. Forecast accuracy measures for the out-of-sample period from January 2017 to February 2018.

Country	Model	MSE	RMSE	MAPE (%)	MASE	U-Theil
Portugal	SARIMA	18.5049	4.3017	3.7368	0.8401	0.3768
	Holt–Winters	**9.0174**	**3.0029**	**2.3260**	**0.5262**	**0.2604**
	Regression	13.7755	3.7115	3.0270	0.6782	0.3339
Germany	SARIMA	**4.9641**	**2.2280**	**1.7134**	**0.5977**	**0.2380**
	Holt–Winters	5.3537	2.3138	1.7875	0.6320	0.2586
	Regression	10.3858	3.2227	2.4758	0.8791	0.3333
Spain	SARIMA	**3.4618**	**1.8606**	**1.4417**	**0.6024**	**0.1905**
	Holt–Winters	4.6211	2.1497	1.6898	0.6957	0.2122
	Regression	7.3954	2.7195	2.1809	0.9292	0.2729
France	SARIMA	**4.4263**	**2.1039**	**1.5323**	**0.4602**	**0.1382**
	Holt–Winters	4.8154	2.1944	1.6128	0.4859	0.1458
	Regression	5.2163	2.2839	1.6972	0.5074	0.1471
Italy	SARIMA	2.0572	1.4343	1.2118	2.1898	0.1368
	Holt–Winters	**1.5008**	**1.2251**	**1.0207**	**1.8622**	**0.1176**
	Regression	6.2250	2.4950	2.2637	4.0110	0.2375
Netherlands	SARIMA	**3.2817**	**1.8115**	**1.3797**	**0.3795**	**0.1944**
	Holt–Winters	5.7452	2.3969	1.8355	0.4966	0.2683
	Regression	7.1199	2.6683	1.9896	0.5587	0.2934
United Kingdom	SARIMA	**4.2713**	**2.0667**	**1.7094**	**0.4876**	**0.1857**
	Holt–Winters	4.4513	2.1098	1.7335	0.5024	0.1912
	Regression	6.8310	2.6136	1.9652	0.5837	0.2471

4.3. Forecast interval coverage

The performance of the forecasting methodologies can also be evaluated by their ability to produce forecast intervals that provide coverages close to the nominal rates [28]. Table 10 shows the percentage of times that the nominal 95% forecast intervals contain the true observations, as well as the mean percentage for each forecasting method.

The results indicate that SARIMA and Holt–Winters models produce coverage probabilities that are close to the nominal rate for almost all of the retail series under study. The

Table 10. Forecast interval coverage for out-of-sample period from January 2016 to February 2018. Nominal coverage of 95%.

Country	SARIMA	Holt–Winters	Regression
Portugal	100.0	100.0	78.6
Germany	100.0	100.0	85.7
Spain	100.0	100.0	85.7
France	100.0	92.9	92.9
Italy	100.0	100.0	92.9
Netherlands	100.0	100.0	92.9
United Kingdom	85.7	85.7	71.4
Global	**98**.0	**96**.9	**85.7**

SARIMA models slightly overestimate the coverage probabilities of nominal forecast intervals in 6 of the 7 retail series, with a mean coverage of 98%. The Holt–Winters method also provides good coverage rates, underestimating them in 2 of the 7 retail series, and overestimating in the remaining ones. On the other hand, multiple linear regression has the less accurate forecast intervals. It underestimates the coverage probabilities of the nominal 95% forecast intervals in 100% of the retail series, with coverage probabilities ranging from 71.4% to 92.9%.

However, it is important to notice that only the test period (14 observations) was used in the calculation of the coverage rates, and so one observation outside the correspondent interval is enough to drop the coverage rate to 92.9% (below the nominal 95%). Also, coverage rates can reach high values due to the wide amplitudes of the forecast intervals. Thus, the comparison between forecasting models should be done in a global way, with the coverage rate being just one more, and not the only one, performance indicator.

In this case, the performance of the forecasting methods in terms of coverage rates matches the results for accuracy measures, also leading to the following classification: (1) SARIMA, (2) Holt–Winters, and (3) multiple linear regression.

5. Concluding remarks

Many business and economic time series exhibit strong trend and seasonal variations, and retail time series are no exception. Since retailing is a widely competitive industry, accurate forecasts are extremely important to ensure the quality of the decision-making process, which greatly impacts effective management of retail business. This study compared the forecasting accuracy of three traditional linear forecasting models applied to retail time series of seven European countries from January 2000 to February 2018. Five accuracy measures were selected: MSE, RMSE, MAPE, MASE, and Theil's U-statistic. The performance of the forecasting methods was also evaluated in terms of coverage rates of the forecast intervals.

The five accuracy measures led to the same conclusions regarding the forecasting performance of the methodologies under study. Based on the empirical results, this study confirmed the previous works by [12,26], showing that multiple linear regression was not the most recommended approach to forecast retail time series. The SARIMA models provided superior point forecasts over the remaining methodologies, with the Holt–Winters model proving to be a viable alternative. In fact, these two methodologies have performed

well for forecasting retail time series, both in terms of accuracy forecasting and coverage rates.

For future work, it is possible to choose the traditional ARIMA models as a benchmark for forecasting retail time series and then evaluate the performance of new methods by comparison.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by the Portuguese FCT Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM and the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020. Susana Lima was financially supported by UMINHO/BI/145/2020.

ORCID

A. M. Gonçalves http://orcid.org/0000-0001-8491-6048
M. Costa http://orcid.org/0000-0001-7686-2430

References

[1] L. Aburto and R. Weber, *Improved supply chain management based on hybrid demand forecasts*, Appl. Soft Comput. 7 (2007), pp. 136–144.
[2] H. Akaike, *A new look at statistical model identification*, IEEE Trans. Autom. Control 19 (1974), pp. 716–723. doi:10.1109/TAC.1974.1100705.
[3] T. Alpuim and A. El-Shaarawi, *On the efficiency of regression analysis with AR(p) errors*, J. Appl. Stat. 35 (2008), pp. 717–737. doi:10.1080/02664760600679775.
[4] T. Alpuim and A. El-Shaarawi, *Modeling monthly temperature data in Lisbon and Prague*, Environmetrics 20 (2009), pp. 835–852. doi:10.1002/env.964.
[5] S. Aras, I.D. Kocakoç, and C. Polat, *Comparative study on retail sales forecasting between single and combination methods*, J. Bus. Econ. Manag. 18 (2017), pp. 803–832. doi:10.3846/16111699.2017.1367324.
[6] N.S. Arunraj and D. Ahrens, *A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting*, Int. J. Prod. Econ. 170 (2015), pp. 321–335. doi:10.1016/j.ijpe.2015.09.039.
[7] G.C. Aye, M. Balcinar, R. Gupta, and A. Majumdar, *Forecasting aggregate retail sales: The case of South Africa*, Int. J. Prod. Econ. 160 (2015), pp. 66–79. doi:10.1016/j.ijpe.2014.09.033.
[8] C. Bergmeir, R.J. Hyndman, and J.M. Benítez, *Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation*, Int. J. Forecast. 32 (2016), pp. 303–312. doi:10.1016/j.ijforecast.2015.07.002.
[9] J. Bermúdez, J. Segura, and E. Vercher, *Holt–Winters forecasting: An alter- native formulation applied to UK air passenger data*, J. Appl. Stat. 34 (2007), pp. 1075–1090. doi:10.1080/02664760701592125.
[10] G. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco, 1970.
[11] C. Chatfield, *Time-Series Forecasting*, Chapman and Hall/CRC, 2000.
[12] C.-W. Chu and G. Zhang, *A comparative study of linear and nonlinear models for aggregate retail sales forecasting*, Int. J. Prod. Econ. 86 (2003), pp. 217–231. doi:10.1016/S0925-5273(03)00068-9.

[13] C. Cordeiro and M. Neves, *Forecasting time series with Boot.EXPOS procedure*, REVSTAT Stat. J. 7 (2009), pp. 135–149.
[14] P. Cowpertwait and A. Metcalfe, *Introductory Time Series with R*, Springer, New York, 2009.
[15] J. De Gooijer and R. Hyndman, *25 years of time series forecasting*, Int. J. Forecast. 22 (2006), pp. 443–473.
[16] R. Fildes, S. Ma, and S. Kolassa, *Retail forecasting: Research and practice*, Int. J. Forecast. (2019). doi:10.1016/j.ijforecast.2019.06.004.
[17] R. Hyndman, A. Koehler, J. Ord, and R. Snyder, *Forecasting with Exponential Smoothing. The State Space Approach*, Springer, 2008.
[18] A. Kolkova, *The application of forecasting sales of services to increase business competitiveness*, J. Compet. 12 (2020), pp. 90–105. doi:10.7441/joc.2020.02.06.
[19] J. Kuvulmaz, S. Usanmaz, and S. Engin, *Time-series forecasting by means of linear and nonlinear models*, in *Advances in Artificial Intelligence*, Vol. 3789, Springer eds., MICAI 2005, 2005, pp. 504–513.
[20] G. Ljung, J. Ledolter, and B. Abraham, *George Box's contributions to time series analysis and forecasting*, Appl. Stoch. Models Bus. Ind. 30 (2014), pp. 25–35. doi:10.1002/asmb.2016.
[21] S. Lima, A.M. Gonçalves, and M. Costa, *Time series forecasting using Holt–Winters exponential smoothing: An application to economic data*, AIP Conf. Proc. 2186 (2019), Article ID 090003. doi:10.1063/1.5137999.
[22] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, *Statistical and machine learning forecasting methods: Concerns and ways forward*, PLoS ONE 13 (2018), Article ID e0194889. doi:10.1371/journal.pone.0194889.
[23] N. Meade, *Evidence for the selection of forecasting methods*, J. Forecast. 19 (2000), pp. 515–535.
[24] Y. Pan, *Predicting aggregate retail sales using hybrid ARIMA*, Proceedings of the 7th Global Business and Social Science Research Conference, 2013.
[25] G.D. Pillo, V. Latorre, S. Lucidi, and E. Procacci, *An application of support vector machines to sales forecasting under promotions*, 4OR-A Q. J. Oper. Res. 14 (2016), pp. 309–325. doi:10.1007/s10288-016-0316-0.
[26] M. Qi, I. Alon, and R. Sadowski, *Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods*, J. Retail. Consum. Serv. 8 (2001), pp. 147–156. doi:10.1016/S0969-6989(00)00011-4.
[27] R Core Team, *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. Available at https://www.R-project.org/.
[28] P. Ramos, N. Santos, and R. Rebelo, *Performance of state space and ARIMA models for consumer retail sales forecasting*, Robot. Comput. Integr. Manuf. 34 (2015), pp. 151–163. doi:10.1016/j.rcim.2014.12.015.
[29] B. Seaman, *Considerations of a retail forecasting practitioner*, Int. J. Forecast. 34 (2018), pp. 822–829. doi:10.1016/j.ijforecast.2018.03.001.
[30] Statistical Office of the European Union, Eurostat database, *Turnover and volume of sales in wholesale and retail trade – monthly data*, 2018. Available at https://ec.europa.eu/eurostat/data/database.
[31] S. Suhartono, M.H. Lee, and D. Prastyo, *Two levels ARIMAX and regression models for forecasting time series data with calendar variation effects*, AIP Conf. Proc. 1691 (2015), Article ID 050026. doi:10.1063/1.4937108.
[32] C. Veiga, C. Veiga, A. Catapan, U. Tortato, and W. Silva, *Demand forecasting in food retail: A comparison between the Holt–Winters and ARIMA models*, WSEAS Trans. Bus. Econ. 11 (2014), pp. 608–614.
[33] C. Veiga, C. Veiga, W. Puchalski, L. Coelho, and U. Tortato, *Demand forecasting based on natural computing approaches applied to the foodstuff retail segment*, J. Retail. Consum. Serv. 31 (2016), pp. 174–181. doi:10.1016/j.jretconser.2016.03.008.
[34] A. Wagner, S. Soumerai, F. Zhang, and D. Ross-Degnan, *Segmented regression analysis of interrupted time series studies in medication use research*, J. Clin. Pharm. Ther. 27 (2002), pp. 299–309. doi:10.1046/j.1365-2710.2002.00430.x.

[35] A. Zamani, H. Haghbin, M. Hashemi, and R.J. Hyndman, *Seasonal functional autoregressive models*, Journal of Time Series Analysis (2021). doi:10.1111/jtsa.12608.
[36] G. Zhang, *Time series forecasting using a hybrid ARIMA and neural network model*, Neurocomputing 50 (2003), pp. 159–175. doi:10.1016/S0925-2312(01)00702-0.
[37] G. Zhang and M. Qi, *Neural network forecasting for seasonal and trend time series*, Eur. J. Oper. Res. 160 (2005), pp. 501–514. doi:10.1016/j.ejor.2003.08.037.