Forecasting Models: An Application to Home Insurance

Luís Filipe Pires[1,3], A. Manuela Gonçalves[1,2(✉)], Luís Filipe Ferreira[3], and Luís Maranhão[3]

[1] Department of Mathematics, University of Minho, Guimarães, Portugal
mneves@math.uminho.pt
[2] Centre of Mathematics, University of Minho, Guimarães, Portugal
[3] International Insurance Group, Guimarães, Portugal

Abstract. Forecasting in time series is one of the main purposes for applying time series models. The choice of the forecasting model depends on data structure and the objectives of the study. This study presents a comparison of Box Jenkins SARIMA and Holt-Winters exponential smoothing approaches to time series forecasting to increase the likelihood of capturing different patterns in the data (in this specific case, home insurance data) and thus improve forecasting performance. These methods are chosen due to their ability to model seasonal fluctuations present in insurance data. The forecasting performance is demonstrated by a case study of home insurance monthly time series: total and frequency rate time series. In order to assess the predictive and forecasting performance of the two methodologies adopted, several evaluation measures are used, namely MSE, RMSE, MAPE, and Theil's U-statistics. A comparison is made and discussed, and the results obtained demonstrate the superiority of the SARIMA model over the other forecasting approach. Holt-Winters also produces accurate forecasts, so it is considered a viable alternative to SARIMA.

Keywords: Home insurance · Time series · Forecasting · SARIMA · Holt-winters

1 Introduction

A time series is a set of observations usually ordered in equally spaced intervals. Time series forecasting is an important area in which past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship. The model is then used to extrapolate the time series into the future. Forecasting methods are a key tool in decision-making processes in many areas, such as economics, insurance, management or environment. There are several approaches to modeling time series, but we decided to study and compare the accuracy of the Box Jenkins SARIMA and Holt-Winters exponential smoothing models for forecasting home insurance time series, because both models can increase the chance of capturing the proprieties and the dynamics of the

O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13377, pp. 514–529, 2022.
https://doi.org/10.1007/978-3-031-10536-4_34

data and thus improve forecast accuracy. Both methods have the ability to deal with time series seasonality. The time series analysis of both processes was carried out using the statistical software R programming language and specialized packages for modeling and forecasting.

The problem proposed by the International Insurance Group, Portugal, was to find the models that best fit and forecast (the focus will be on forecasting) the monthly time series of total home insurance, including storm damage and claims, contents damage, leaking or escape of water in the home, accidental damage, damage to fridge/freezer food in the home, building damage, theft or robbery, fire and others, in order to use them to forecast future values. For example, see the importance of this type of insurance in economic terms: at least 85% of homeowners in the U.S. have homeowners insurance, and policies cost on average $1,445 per year. While it's not a required form of coverage by the government, home insurance is typically required as a condition for applying for a mortgage and it is very valuable because of the protection it provides homeowners. Home insurance policies generally provide coverage for damage to a home's structure, damage to personal property and liability coverage in case the policy holder is considered at fault for property damage or bodily injury to another party. There is a growing number of accidents taking place, either caused by humans actions or by nature. Population growth at a national level certainly entails increasing numbers of accidents (recorded by insurance companies). Another major factor with tremendous weight in these records is climate change. Contrary to what one might assume at first glance, this phenomenon not only contributes to hotter summers, but also to colder winters, i.e., climate change increases the volatility of Earth' s temperature. This results in more intense and more frequent storms, which are ultimately reflected in insurance markets. In these markets, a claim is defined as an event that results in material loss to an insured individual.

In this study the claims data regard home insurance (total home insurance as defined above). Our data source are the records of the claims registered in the period from January 2015 to June 2021 on a monthly basis (data from Portugal). The main goal is to forecast these claims in monthly time series. Two time series are considered: the total home insurance (number of claims registered in the month) and a monthly frequency rate. This monthly frequency rate is defined as follows

$$\frac{\text{number of claims registered in the month}}{\text{number of people exposed to risk in the month}} \tag{1}$$

which varies between 0 and 1, where the number of people exposed to risk in a given month is equal to the number of portfolios open. It should be stressed that the main focus of this study is to establish accurate forecasting models to support managerial performance the decision-making process to improve the services provided to policyholders.

2 Methodologies

2.1 SARIMA Model

The Box Jenkins SARIMA$(p, d, q)(P, D, Q)_s$ is a short memory model and a very flexible model, given that it accounts for stochastic seasonality, and is one of the most versatile models for forecasting seasonal time series. Such seasonality is present when the seasonal pattern of a time series changes over time. The theory of SARIMA models has been developed by many researchers and its wide application results from the work by Box et al. [1], who developed a systematic and practical model-building method. Through an iterative three-step model-building process, model identification, parameter estimation and model diagnosis, the Box-Jenkins methodology has proven to be an effective practical time series modeling approach.

The SARIMA model has the following form

$$\Phi_p(B)N_P(B^s)(1-B)^d(1-B^s)^D Y_t = \Theta_q(B)H_Q(B^s)\epsilon_t, \tag{2}$$

where Y_t is the time series, with

$$\begin{aligned}
&\Phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^p,\\
&N_P(B^s) = 1 - \nu_1 B^s - \cdots - \nu_P P_s,\\
&\Theta_q(B) = 1 + \theta_1 B + \cdots + \theta_q B^q,\\
&H_Q(B^s) = 1 + \eta_1 B^s + \cdots + \eta_Q Q_s,
\end{aligned}$$

where s is the seasonal length, 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 model with the minimum AIC (Akaike's Information Criterion) value and the minimum BIC (Bayesian's Information Criterion) value is often the best model for forecasting [2]. We investigated the required transformations for variance stabilization and decided to apply logarithms to the time series under study.

Once the model has been specified, its autoregressive, moving average, and seasonal parameters (SARIMA model) need to be estimated. The parameters of SARIMA models are usually estimated by maximizing the likelihood of the model.

2.2 Holt-Winters Model

The Holt-Winters method is an extension of the Holt method, and is applied whenever the data behavior is trendy and is seasonal. The seasonal type can be additive or multiplicative, depending on the oscillatory movement over the time period. In both versions, the forecasts will depend on the following three components of a seasonal time series: level, trend and seasonal coefficient. In addition, both are implemented in the Holt-Winters function of the *forecast* package in R. The additive version ought to be considered whenever the seasonal pattern of a series has constant amplitude over time [3]. In such case, the series can be written by $Y_t = T_t + S_t + \varepsilon_t$, where T_t represents the trend (the sum of the level and the slope of the series at time t), S_t is the seasonal component, and ε_t are error terms with mean 0 and constant variance. When a series displays a seasonal pattern characterized by amplitude that varies with the series level, the multiplicative version is a better choice. In such case, the series can be represented by $Y_t = T_t \times S_t + \varepsilon_t$. The multiplicative and additive Holt-Winters methods have the recursive equations presented in the Table 1. The exploratory analysis of all eight time series indicated the presence of a seasonal pattern.

Table 1. The recursive equations of the Holt-Winters methods.

Multiplicative H-W	$F_t = \alpha \frac{Y_t}{f_{t-s}} + (1-\alpha)(F_{t-1} + b_{t-1})$, $0 \leq \alpha \leq 1$
	$b_t = \beta(F_t - F_{t-1}) + (1-\beta)b_{t-1}$, $0 \leq \beta \leq 1$
	$f_t = \gamma \frac{Y_t}{F_t} + (1-\gamma)f_{t-s}$, $0 \leq \gamma \leq 1$
	$\hat{Y}_{t+k} = (F_t + kb_t)f_{t+k-ms}$, $m = 1, 1 \leq k \leq s, m = 2, s < k \leq 2s$, etc.
Additive H-W	$F_t = \alpha(Y_t - f_{t-s}) + (1-\alpha)(F_{t-1} + b_{t-1})$, $0 \leq \alpha \leq 1$
	$b_t = \beta(F_t - F_{t-1}) + (1-\beta)b_{t-1}$, $0 \leq \beta \leq 1$
	$f_t = \gamma(Y_t - F_t) + (1-\gamma)f_{t-s}$, $0 \leq \gamma \leq 1$
	$\hat{Y}_{t+k} = F_t + kb_t + f_{t+k-ms}$, $m = 1, 1 \leq k \leq s, m = 2, s < k \leq 2s$, etc.

The Bootstrap method introduced in [4] provides a way to estimate parameters, approximate a sampling distribution or derive confidence intervals when we have data but do not know the underlying distribution. If the population represented through a probability distribution and its parameters are unknown, the Bootstrap idea is to take (re-)samples $(y_1^*, y_2^*, \ldots, y_n^*)$, drawn with replacement from the original sample $(y_1, y_2, \ldots, y_n)$. Computing prediction intervals are an important part of the forecasting process and aim to indicate the likely uncertainty in point forecasts. The prediction intervals are usually based on the Mean Square Error (MSE), which denotes the variance of the one-step-ahead forecast errors [5]. Prediction intervals (whenever a normality assumption is verified) for both at one-step-ahead and at m-steps-ahead are given by the following expression:

$$\left[\hat{Y}_{k+m} - z_{1-\alpha/2}\sqrt{MSE_m}, \hat{Y}_{k+m} + z_{1-\alpha/2}\sqrt{MSE_m}\right]$$

where z is the quantile of probabilty $1 - \alpha/2$ of the standard Normal distribution and $MSE_m = \frac{1}{k-m}\sum_{t=m+1}^{k}(\epsilon_t^{(m)})^2$ denote the variance of the m-steps-ahead errors. The idea is to look at the bootstrap percentiles rather than the sampling distribution percentiles, and the confidence interval is based on the

Bootstrap distribution (i.e., on the percentiles). Different methods are available for the construction of Bootstrap confidence intervals: the percentile method, the percentile-t method, the bias-corrected method [4] and the accelerated bias-corrected method [4]. The Bootstrap percentile confidence interval is based on the quantiles of the Bootstrap estimates distribution and is obtained as follows: suppose F_k is the empirical cumulative distribution function $\{\hat{y}^b_{n+k}, b = 1, \ldots, B\}$, then the prediction interval is given by $\left[F_k^{-1}(\alpha/2), F_k^{-1}(1-\alpha/2)\right]$, where for an interval with 95% confidence and B replicates the limits of the intervals are the percentiles $F_k^{-1}(0.025)$ and $F_k^{-1}(0.975)$.

2.3 Forecasting Models Evaluation

Let's denote the actual observation for time period t by Y_t and the estimated or forecasted value for the same period by $\hat{Y}_t$ and n is the total number of observations. The most commonly used forecast error measures are the mean squared error (MSE), the root mean squared error (RMSE), the mean absolute percentage error (MAPE), and Theil's U-statistics [2]. MSE, RMSE, and MAPE are defined by the following formulas, respectively:

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2 = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2, \text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}, \tag{3}$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100 \quad (\%)\,. \tag{4}$$

Theil's U-statistics 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

$$\text{U-Theil} = \sqrt{\frac{\sum_{t=1}^{n-1}\left(\frac{\hat{Y}_{t+1} - Y_{t+1}}{Y_t}\right)^2}{\sum_{t=1}^{n-1}\left(\frac{Y_{t+1} - Y_t}{Y_t}\right)^2}}. \tag{5}$$

Since there is no universally agreed-upon performance measure that can be applied to every forecasting situation, multiple criteria are often required to enable a comprehensive assessment of forecasting models [2]. When comparing forecasting methods, the method with the lowest MSE, RMSE, MAPE or Theil's U-statistics is the preferred one. Often, different accuracy measures will lead to different results as to the best forecast method.

3 Dataset

The study started with exhaustive comprehensive description of the data to understand their behavior over time but also globally. The first step in the analysis of any time series is the description of the historic series. It includes the graphical representation of the data. When a time series is plotted, common patterns are frequently found. These patterns might be explained by many possible cause-and-effect relationships. Common components are the trend, seasonal effect, cyclic changes and randomness. A more interesting and ambitious task is to forecast future values of a series based on its recorded past, and more specifically to calculate forecasting intervals. Therefore, identifying these components is important when selecting a forecasting model.

Table 2 presents descriptive statistics for the total monthly time series. As expected, standard deviation is higher and indicates a larger variability during the observed period. The mean and median are around 1466 and 1400 claims by month, respectively.

Table 2. Descriptive statistics of the total monthly time series.

Beginning	01/2015
End	06/2021
Dimension	78
Range	1013–2707
$Q_{0.25}$	1250
Median	1400
$Q_{0.75}$	1623
Mean	1466.13
Standard deviation	340.97
Variance	116258.6
Variation coefficient	0.23
Number of outliers	4

The years 2020 and 2021, in particular 2021, show a different behaviour from the previous years: 2021 presents a huge variability and higher values. This is due to COVID-19 pandemic context (Fig. 1). There are four outlier observations: January 2015 (1669 claims), March 2018 (2707 claims), January 2020 (2138 claims), and December 2020 (2131 claims).

Before implementing the modeling processes, it was decided to transform the data regarding the total home insurance (number of claims registered in the month) with values greater than 2000 claims per month (a threshold of 2000 claims). Thus, if the value observed (y_o) in a given month was greater than 2000 claims the following transformation is applied $y_o = \frac{y_{o-12}+y_{o+12}}{2}$ with regard to the annual seasonality ($s = 12$ months) inherent in the data (an increase at the end of each year, followed by a decrease at the beginning of the following year), which the graphic representations of the FAC and FACP indicate (see Fig. 2). Therefore, six values were transformed (Fig. 3).

Table 3 presents the values above the 2000 threshold and the transformed values.

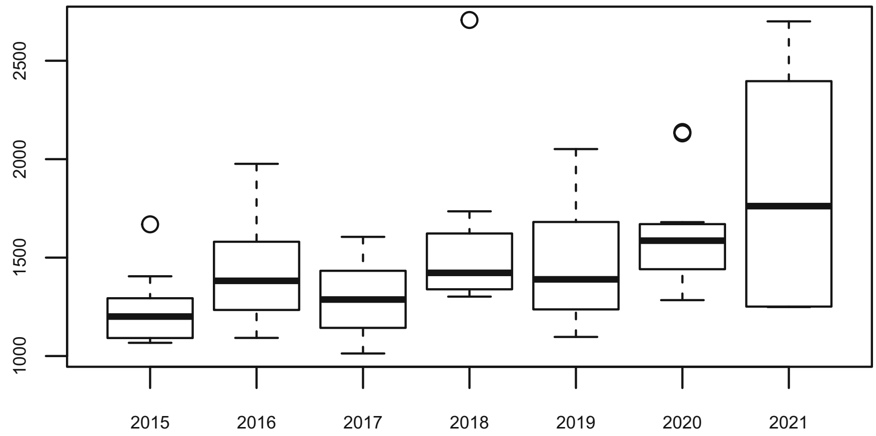

Fig. 1. Boxplots of the total time series from 2015 to 2021.

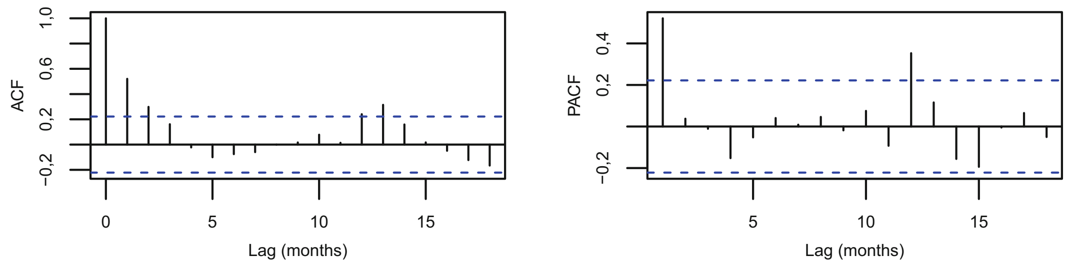

Fig. 2. ACF and PACF of the total monthly time series.

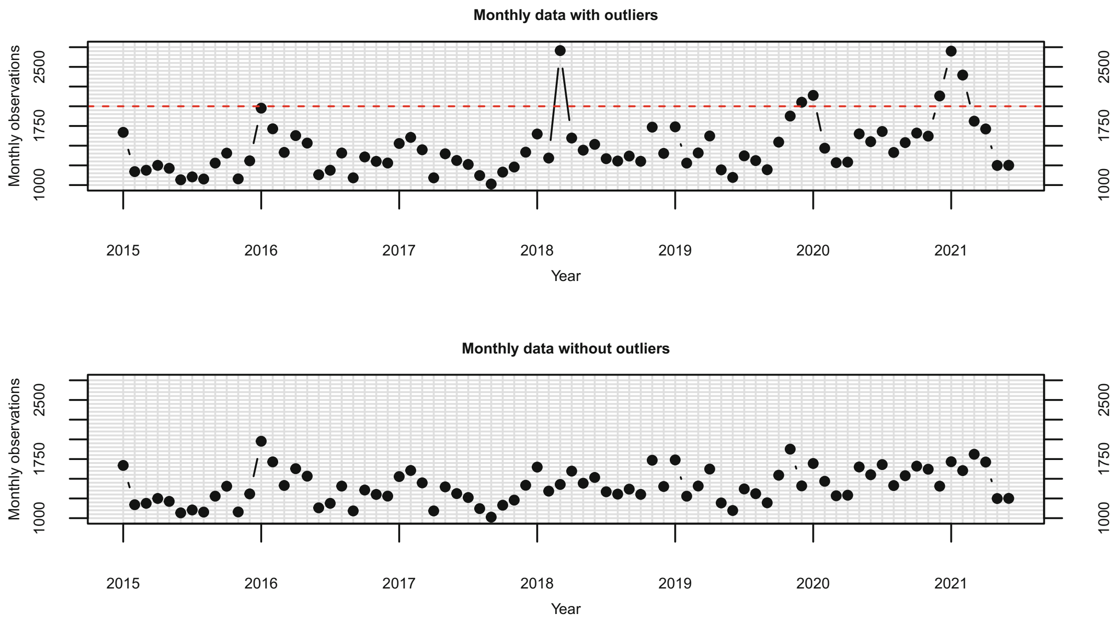

Fig. 3. Effect of the threshold smoothing in the total monthly time series.

Table 3. Values before and after transformation.

Date	Before	After
03/2018	2707	1428
12/2019	2051	1410
01/2020	2138	1694
12/2020	2131	1406
01/2021	2700	1716
02/2021	2396	1604

It should be noted that this transformation was not applied to the monthly frequency rate time series (a ratio ranging between 0 and 1). In the same context, the frequency rate time series also presents 12-month seasonality (Fig. 4).

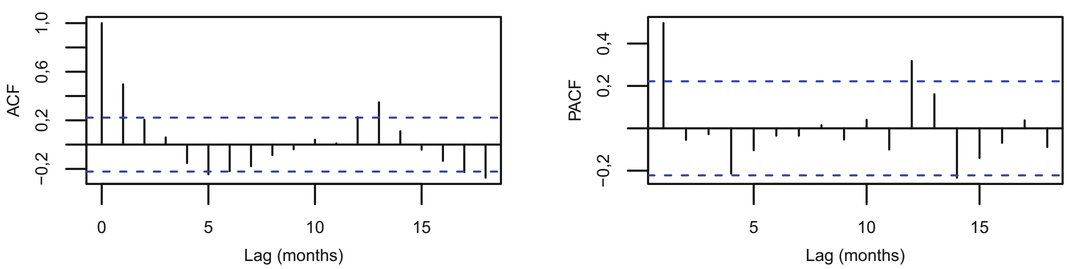

Fig. 4. ACF and PACF of the total monthly frequency rate time series.

Also, the graphical representation of the two time series clearly shows that time series exhibit seasonal behavior (Fig. 5 and Fig. 6), as expected due to the nature of the data. The monthly data exhibits a strong monthly seasonality (a period of 12 months).

4 Results

The results obtained from the application of the SARIMA and Holt-Winters methods are reported in this section. The methods considered in this study are applied to two data sets: training data (in-sample data) and testing data (out-of-sample data) in order to test the accuracy of the proposed forecasting models. This process is implemented regarding both time series: the total home insurance (number of claims registered in the month) and the monthly frequency rate. The selected training period was from January 2015 to December 2020 (first 72 observations/months) and was used to fit the models to the data, and the test period included the last 6 months, i.e., the period from January 2021 to June 2021 was used to forecast. This approach allows comparing the effectiveness of different methods of prediction.

Note that the final results that will be presented always refer to the number of claims registered per month, i.e., the rates time series are modeled and forecasted but the final results presented in this paper are transformed back to the initial values: the total home insurance (number of claims registered in the month).

4.1 SARIMA Model

The ADF (Augmented Dickey-Fuller) and KPSS (Kwiatkowski, Phillips, Schmidt, and Shin) tests (with a 5% significance level) were applied to the total home insurance time series to test series's stationarity and nonstationarity. A Box-Cox transformation was applied to the data, deciding on a $\lambda = -0.184$. Therefore, the SARIMA model will be fitted to this time series for the observations with Box-Cox transformation. The main task in SARIMA forecasting is to select an appropriate model order, i.e., the p, d, q, P, D, Q and s values ($s = 12$).

Tables 4 and 5 show the five models with lower AIC and BIC fitted to the training time series for both cases: the total home insurance (number of claims registered in the month) and a monthly frequency rate. All combinations of parameters p, d, q, P, D, and Q were tested.

Table 4. Adjustment of SARIMA models for the total monthly time series.

Model	AIC	Model	BIC
$\text{SARIMA}(1,0,2)(0,1,1)_{12}$	862.62	$\text{SARIMA}(1,0,0)(0,1,1)_{12}$	870.01
$\text{SARIMA}(2,0,1)(0,1,1)_{12}$	863.26	$\text{SARIMA}(0,0,1)(0,1,1)_{12}$	870.04
$\text{SARIMA}(1,0,0)(0,1,1)_{12}$	863.73	$\text{SARIMA}(0,0,0)(0,1,1)_{12}$	871.92
$\text{SARIMA}(0,0,1)(0,1,1)_{12}$	863.76	$\text{SARIMA}(1,0,0)(0,1,2)_{12}$	872.54
$\text{SARIMA}(1,0,2)(0,1,2)_{12}$	863.82	$\text{SARIMA}(1,0,0)(1,1,1)_{12}$	872.85

Table 5. Adjustment of SARIMA models for the total frequency rate time series.

Model	AIC	Model	BIC
$\text{SARIMA}(2,0,2)(1,0,2)_{12}$	−487.91	$\text{SARIMA}(1,0,0)(0,0,0)_{12}$	−478.95
$\text{SARIMA}(2,0,1)(0,0,0)_{12}$	−487.51	$\text{SARIMA}(0,0,1)(0,0,0)_{12}$	−478.45
$\text{SARIMA}(2,0,1)(0,0,1)_{12}$	−486.79	$\text{SARIMA}(1,0,0)(0,0,1)_{12}$	−477.05
$\text{SARIMA}(2,0,2)(0,0,2)_{12}$	−486.27	$\text{SARIMA}(0,0,1)(0,0,1)_{12}$	−476.61
$\text{SARIMA}(1,0,0)(0,0,1)_{12}$	−486.16	$\text{SARIMA}(1,0,0)(1,0,0)_{12}$	−476.54

On the bases of the AICs e BICs criteria, it is preferred the $\text{SARIMA}(1,0,0)(0,1,1)_{12}$ model for total monthly time series, and the $\text{SARIMA}(1,0,0)(0,0,1)_{12}$ model for total frequency rate time series.

The estimation results of these two models can be consulted in more detail in Table 6.

Table 6. Characteristics of the SARIMA models.

Monthly time series	AIC = 863.73	BIC = 870.01	$\widehat{\sigma} \approx 0.0216$
SARIMA$(1,0,0)(0,1,1)_{12}$	Parameter	ϕ_1	η_1
	Estimate	0.3788	−0.6851
	Standard deviation	0.1238	0.2027
Frequency rate	AIC = −486.16	BIC = −477.05	$\widehat{\sigma} \approx 0.008$
SARIMA$(1,0,0)(0,0,1)_{12}$	Parameter	ϕ_1	η_1
	Estimate	0.3601	0.2101
	Standard deviation	0.1124	0.1300

Figures. 5 and 6 present the original values of the two time series, as well as the estimates in the modeling period (training period).

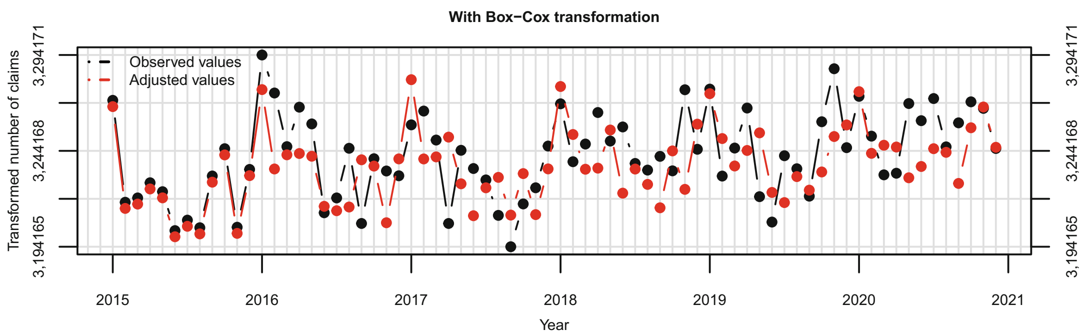

Fig. 5. Adjustment of the SARIMA model for the monthly total time series.

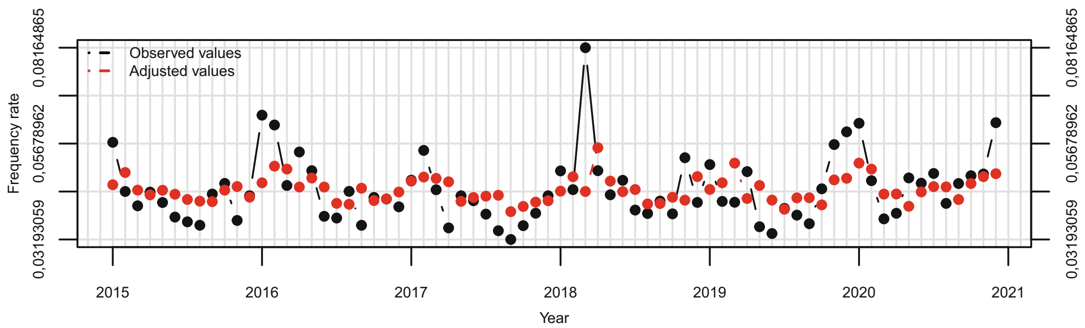

Fig. 6. Adjustment of the SARIMA model for the total frequency rate time series.

Once the forecasts' accuracy (punctual) is evaluated, it is essential to understand the effectiveness of the forecast intervals.

Theoretically, the forecast intervals are calculated at 95% confidence level, which means that 95% of the intervals must include the real observation. That is, it is considered that the most effective interval forecasts are those whose effective coverage rate is closer to 95%. Note that the forecast intervals are obtained based on the testing series for each distinct series where (in this study) they contain only 6 observations (6 months) and, therefore, the analysis of coverage rates must be taken to consideration. Also the testing period is from January 2021 to June 2021, corresponding to half a year of COVID-19 pandemic, which had a tremendous impact on society as a whole, including on human behavior and consequently on the number of claims on home insurance which clearly behaved differently from previous periods. In the two time series under study, total monthly time series and total frequency rate, with coverage rates of 100% and 67%, respectively, are calculated. It should be noted that the 100% coverage rate is due to a greater amplitude of the forecasting interval. It is noteworthy that the model formulated for the time series corresponding to the total claims monthly presents better results (Figs. 7 and 8).

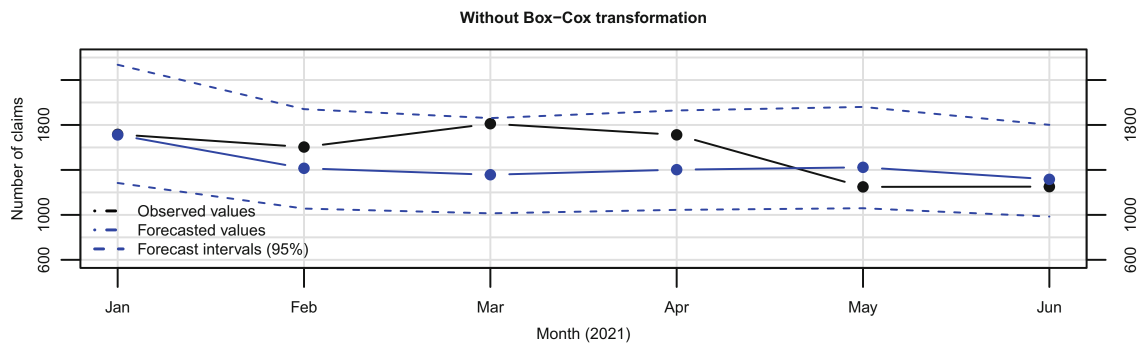

Fig. 7. Forecasts and forecast confidence intervals (95%) of the SARIMA model for the total monthly time series.

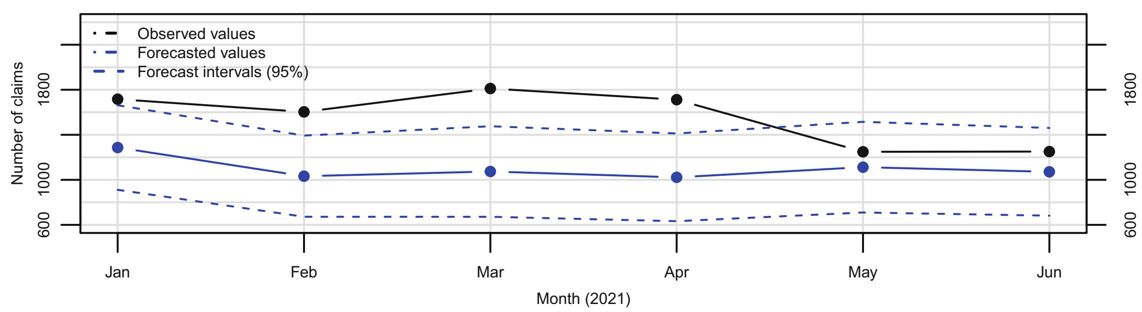

Fig. 8. Forecasts and forecast confidence intervals (95%) of the SARIMA model for the total frequency rate time series.

The models validation was assessed by means of the residuals analysis. The independency assumptions were assessed by estimating the autocorrelation and the partial autocorrelation functions of residuals and the assumption that the residuals are identically normally distributed were also verified (by performing the Kolmogorov-Smirnov test).

4.2 Holt-Winters Model

We applied the additive and multiplicative Holt-Winters to the first k observations (we considered the period from January 2015 to June 2015, the first semester), and we obtained the initial values for the smoothing parameters. We obtained the residuals and we calculated the MSE to compare the forecasting accuracy. The additive Holt-Winters models proved to have the best predictive performance. So, for all times series, we considered the models obtained by the additive Holt-Winters method. Tables 7 and 8 show the smoothing constants estimates of the two additive models for the training time series for both cases: the total home insurance (number of claims registered in the month) and a monthly frequency rate.

Table 7. Holt-Winters additive model parameters estimates for the monthly counts.

$\widehat{\alpha} \approx 0.1603$	$\widehat{\beta} \approx 0.0212$	$\widehat{\gamma} \approx 0.5974$	$\widehat{l}_1 \approx 1644.5000$	$\widehat{b}_1 \approx 8.4710$
$\widehat{s}_1 \approx 154.2854$	$\widehat{s}_2 \approx -112.3845$	$\widehat{s}_3 \approx -175.0366$	$\widehat{s}_4 \approx -87.0292$	$\widehat{s}_5 \approx -1.2906$
$\widehat{s}_6 \approx -110.3108$	$\widehat{s}_7 \approx -15.5516$	$\widehat{s}_8 \approx -200.5818$	$\widehat{s}_9 \approx -177.1819$	$\widehat{s}_{10} \approx -35.7976$
$\widehat{s}_{11} \approx 54.6681$	$\widehat{s}_{12} \approx -206.0083$			

Table 8. Estimates for the frequency rate Holt-Winters additive model's parameters.

$\widehat{\alpha} \approx 0.0839$	$\widehat{\beta} \approx 0.0128$	$\widehat{\gamma} \approx 0.3795$	$\widehat{l}_1 \approx 0.0526$	$\widehat{b}_1 \approx 0.0002$
$\widehat{s}_1 \approx 0.0081$	$\widehat{s}_2 \approx 0.0002$	$\widehat{s}_3 \approx -0.0012$	$\widehat{s}_4 \approx 0.0027$	$\widehat{s}_5 \approx -0.0034$
$\widehat{s}_6 \approx -0.006$	$\widehat{s}_7 \approx -0.0052$	$\widehat{s}_8 \approx -0.0091$	$\widehat{s}_9 \approx -0.0074$	$\widehat{s}_{10} \approx 0.0043$
$\widehat{s}_{11} \approx -0.0005$	$\widehat{s}_{12} \approx 0.0037$			

In Fig. 9 and Fig. 10 are represented the original values of the claims time series, the total home insurance (number of claims registered in the month) and a monthly frequency rate, respectively, and the estimates in the modeling period (training period).

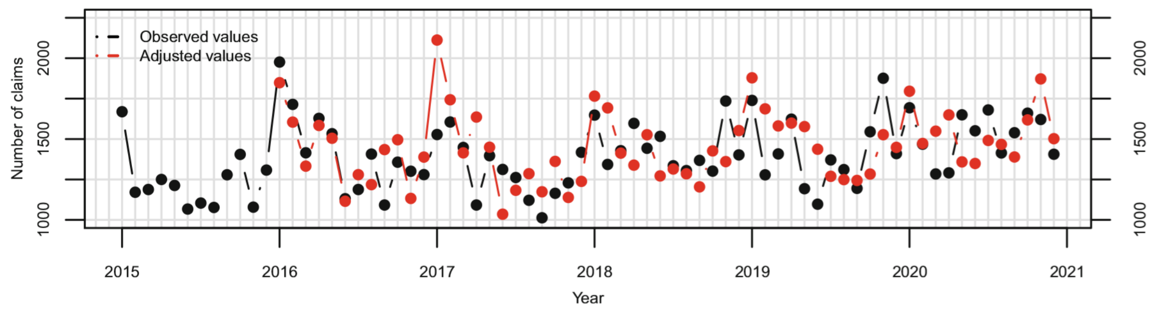

Fig. 9. Adjustment of the Holt-Winters model for the total monthly time series.

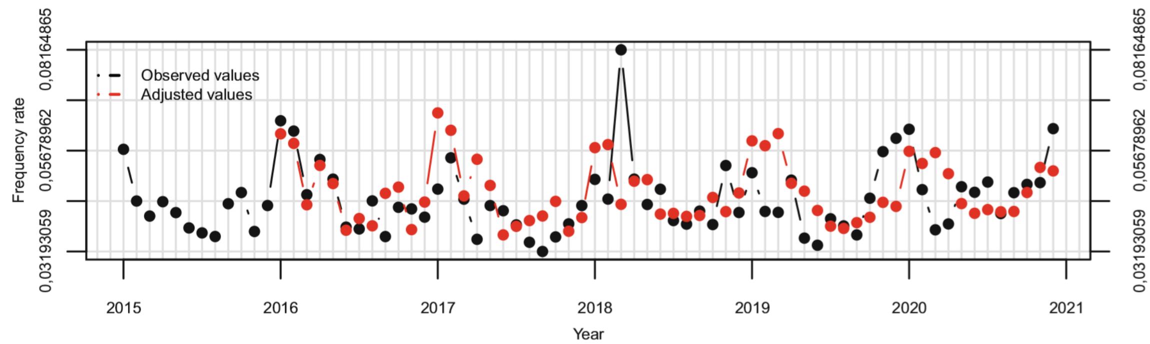

Fig. 10. Adjustment of the Holt-Winters model for the total frequency rate time series.

The Holt-Winters procedure associated to the Bootstrap resampling method was programmed in the R software by taking $B = 2000$ replicates in the residuals resampling process. Thus, the Bootstrap percentile confidence interval obtained is based on the quantiles of the Bootstrap estimates distribution. The forecasts from January 2021 to June 2021 for both time series processes were computed to assess the performance of the methodologies, namely by the forecast 95% confidence intervals range.

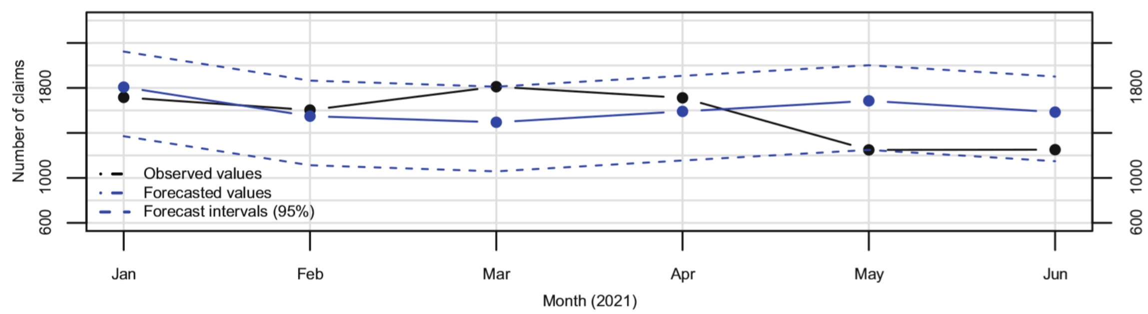

Fig. 11. Forecasts and forecast Bootstrap percentile confidence interval (95%) of the Holt-Winters model for the total monthly time series.

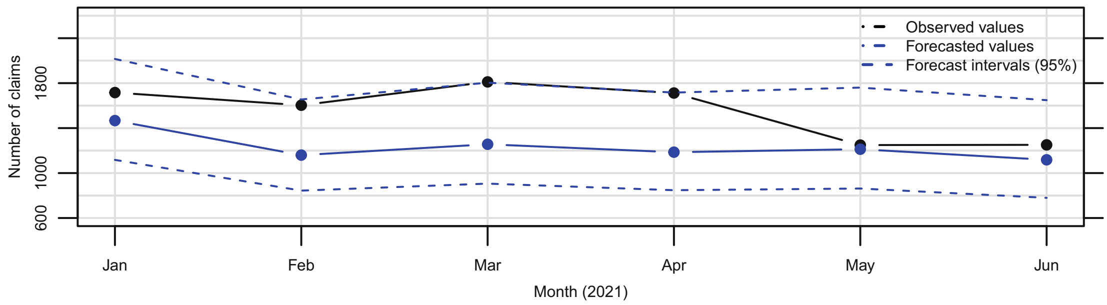

Fig. 12. Forecasts and forecast Bootstrap percentile confidence interval (95%) of the Holt-Winters model for the total frequency rate time series.

Figure 11 and Fig. 12 present the original values, the forecasts in the forecasting period (testing period) and the Bootstrap forecast confidence intervals for a 95% confidence level for the two Holt-Winters models. The coverage rate of empirical confidence of corrected forecasts is 83% and 100% (this happens because the amplitude of the forecast interval is larger), for the total monthly time series and total frequency rate time series, respectively, of the confidence intervals with a 95% confidence level (5 and 6 observations of the testing series belong to the confidence interval).

4.3 Models Performance

Tables 9 and 10 show the results of the accuracy measures calculated for the entire observation period (a total period of observation of 78 months), training and testing periods for the two methods applied to the time series under study. The performance comparisons of the competing models (SARIMA and Holt-Winters) were evaluated using MSE, RMSE, MAPE, and Theil's U- statistics. The results obtained showed that the SARIMA, which requires fewer parameters to be estimated, is (predominantly) more accurate than Holt-Winters and performs better for all period times (total, training and test periods). From the two models performed, we selected the most adequate model which has the lowest forecast error when comparing predicted data using a suitable test set: SARIMA. Therefore, the SARIMA models can more efficiently capture the dynamic behavior of the total monthly time series and total frequency rate time series compared to Holt-Winters.

Table 9. Evaluation metrics of the SARIMA models of both time series.

	Time series	MSE	RMSE	MAPE	U-Theil
Total	Monthly counts	189950.30	435.83	9.90	0.78
	Frequency rate	39239.89	198.09	12.11	0.83
Training set	Monthly counts	33028.16	181.74	9.69	0.80
	Frequency rate	37320.45	193.19	11.78	0.81
Testing set	Monthly counts	62092.11	249.18	12.40	1.18
	Frequency rate	62273.19	249.55	16.00	1.23

Table 10. Evaluation metrics of the Holt-Winter models of both time series.

	Time series	MSE	RMSE	MAPE	U-Theil
Total	Monthly counts	48410.05	220.02	12.68	0.98
	Frequency rate	331902.60	576.11	36.70	2.52
Training set	Monthly counts	46119.56	214.75	12.36	0.97
	Frequency rate	350710.00	592.21	38.41	2.60
Testing set	Monthly counts	71315.00	267.05	15.81	1.36
	Frequency rate	143828.40	379.25	19.53	1.71

5 Conclusions

The main objective of this study was to establish accurate forecasting models to enable the Insurance Company level to (monthly) forecast the number of housing claims with good accuracy to optimize costs at the managerial level. The SARIMA models provided superior point forecasts over the remaining methodology, with the Holt-Winters model proving to be a viable alternative. The lowest observed value was 249 claims (root mean squared error) per month for the forecasting with the total monthly time series modeling process (see Table 9). Contrary to expectations, claims modeling via monthly rate did not result in better estimates or better forecasts for the process of obtaining accurate claims numbers. Thus, future research should only consider the forecasting process of the time series via the original data (number of claims by month). In fact, for the total home insurance, both model processes in terms of accuracy forecasting and coverage rates have performed well regarding the number of claims registered by month.

Acknowledgements. A. Manuela Gonçalves was partially financed by Portuguese Funds through FCT (Fundação para a Ciência e a Tecnologia) within the Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM.

References

1. Box, G., Jenkins, G., Reinsel, G.: Time Series Analysis, 4th edn. Wiley, Hoboken (2008)
2. Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. Online Open-Acess Texbooks. http://otexts.com/fpp. Accessed 5 Oct 2020
3. Kalekar, P.S.: Time series forecasting using holt-winters exponential smoothing. Kanwal Rekhi School Inf. Technol. **4329008**, 1–13 (2004)
4. Efron, B., Tibshirani, R.: Bootstrap methods for standard errors, confidence intervals, and another measures of statistical accuracy. Stat. Sci. **1**(1), 54–77 (1986)
5. Cordeiro, C., Neves, M.M.: Séries Temporais e Modelos de Previsão. Aplicação da Metodologia Bootstrap. Actas do XI Congresso Anual da SPE, Lisboa, pp. 153–164 (2003)