ISSN: 2316-2341



REVISTA ENIAC PESQUISA

ENIAC Centro Universitário de Excelência

# *Previsão da sinistralidade em seguros de vida utilizando modelos de séries temporais*

## Claims forecast in life insurance using time series models

**Recebimento: 29/04/2023 - Aceite: 03/11/23 - Publicação:**
**Processo de Avaliação: Double Blind Review - DOI:** https://doi.org/10.22567/rep.v13i1.961

**Amanda Santos Pandolfi**
amandasantospandolfi@gmail.com
https://orcid.org/0009-0002-5252-8045
Universidade Federal de Minas Gerais (UFMG), Brasil

**Jussiane Nader Gonçalves**
jusianegoncalves@gmail.com
https://orcid.org/0009-0009-1285-3928
Universidade Federal de Minas Gerais (UFMG), Brasil

***RESUMO***

*A condição de não estaticidade dos seguros em geral se dá por se tratar de uma ciência que depende de fatores sociais e econômicos para se sustentar. O Seguro de Vida, em especial, é impactado por diversos agentes, desde políticas econômicas e culturais, até o desempenho das seguradoras quanto a oferta e demanda do seu produto e aspectos sociodemográficos que influenciam no comportamento do próprio segurado. O conhecimento sobre o risco identificado é fundamental para que a avaliação seja realista e precisa. Para garantir a solvência dessas empresas e assegurar os compromissos futuros, diversas premissas atuariais estão envolvidas no gerenciamento e precificação de riscos, como análise de frequência de sinistros, severidade, risco biométrico, análises de sensibilidade e subscrição, acompanhamento de sinistralidade, entre outras, cabendo ao atuário responsável adequar tais medidas a realidade da instituição de risco. Portanto, este trabalho tem como objetivo avaliar modelos de previsão para uma das premissas atuariais mencionadas, a saber, a sinistralidade, por meio de métodos de regressão e Box-Jenkins. Foram analisados os dados mensais disponíveis no Sistema de Estatísticas (SES) da Superintendência de Seguros Privados (SUSEP), entre os anos de 2011 e 2021, das oito seguradoras com maior volume de prêmio. A capacidade de previsão dos modelos foi avaliada comparando os valores previstos com os*

ISSN: 2316-2341

*dados reais do 1° semestre de 2022. O modelo de regressão foi o mais adequado para a previsão da sinistralidade, embora ambos os métodos tenham apresentado desempenho satisfatório.*

***Palavras-chave**: seguros de vida, sinistralidade, modelos de séries temporais.*

## ABSTRACT

The non-static nature of insurance, in general, is because it depends on many social and economic factors to sustain itself. Individual Life Insurance, in particular, is impacted by various factors that include economic and cultural policies, the performance of insurers in terms of supply and demand, and sociodemographic aspects that influence the behavior of the life insurance consumer. Analyzing each factor may facilitate the acceptance and pricing of insurance. Knowledge of the identified risk is essential for the evaluation to be realistic and accurate. To ensure the solvency of these companies and, thus, for them to fulfill their future commitments, the management and risk pricing involves several actuarial assumptions, such as analysis of claim frequency, severity, biometric risk, sensitivity and subscription analysis, and monitoring of claims experience. Consequently, it is up to the responsible actuary to manage such measures to the reality of the risk institution. Therefore, this paper aims to evaluate prediction models for one of the aforementioned actuarial assumptions, namely, loss ratio, through regression and Box-Jenkins models. This work analyzed monthly data available in the Statistics System (SES) of the Superintendence of Private Insurance (SUSEP), from the eight insurers with the highest premium volume, between 2011 and 2021. To perform model adequacy, we have compared the predicted values with the current data from the first half of 2022. Both regression and Box-Jenkins models showed satisfactory performance, but the regression model was more appropriate for predicting the loss ratio of the claims experience.

**Keywords:** life insurance, loss ratio, time series models.

ISSN: 2316-2341



## 1. INTRODUCTION

The study on insurance is non-static due to the fact that it is a Science that depends on many social and economic factors to sustain itself. According to the National Insurance School (ENS), Individual Life Insurance, especially, is impacted by several agents, such as economic and cultural policies within a macroeconomic conjecture, the performance of insurance companies regarding the supply and demand of their product, as well as sociodemographic aspects that influence the life insurance consumer's behaviour, for example, in a microeconomic scenario.

As stated by Huebner e Black (1976), a few factors such as age, sex, physical constitution, physical state, personal history, genetic inheritance, moral hazard, engaging in risky activities, profession, occupation, among others, should be analyzed by the insurance company for risk acceptance. Analyzing each factor may facilitate the acceptance and pricing of insurance. The knowledge of the identified risk is fundamental for a realistic and accurate evaluation. These are especially important because they are directly related to the premium volume needed by the risk institution in order to afford possible future claims sums from the subscriptions.

Life insurances generally last medium to long term, and their pricing and management entail several actuarial assumptions, such as analysis of claim frequency, severity, biometric risk, sensitivity and subscription analysis, and monitoring claims. It is an actuary's responsibility to adequate these assumptions to the insurance company's reality. In that way, the loss ratio evaluation is an important indicator in the field, given it represents the relation between the total retained claims amount and the premium volume income in a specific time window. Generally speaking, loss ratio indicates cost representativeness in relation to the premium income.

Therefore, the loss ratio analysis is a fundamental benchmark in order to manage the risk acceptance and evaluate the adequacy of insurance prices. Furthermore, it is a prudential indicator which helps assess institution risk levels and identify possible failings in the insurance subscription process. Based on this evaluation, the insurance company may take actions to minimize losses, such as proposing new pricing methods to adjust premium values. In this context, the goal of this paper is to evaluate statistical models to forecast loss ratio taxes. Especifically to carry out a study on Individual Life Insurance with Death Coverage, branch 1391, as defined by SUSEP.

This paper aims to predict the future loss ratio development through an investigation of its past behavior, as well as to evaluate stochastic methods which allow measuring the uncertainty of this variable. For that end, we have proposed using time series models.

A time series is made of values from a variable evaluated at distinct points in time, and its analysis is based on the assumption that there is an approximately constant causal system related to time, which has influenced past data and can continue to do so in the future (Chaves et al., 2014). This system generally creates patterns that can be detected in statistical analyses. According to Morettin and Toloi (2006), the main goals for analyzing a time series are: i) investigating the generating mechanism of the time series; ii) predicting future series values; iii) describing a time series behavior, such as seasonality and tendency; and iv) searching for relevant intervals in the data.

As explained by Gonçalves and Barreto-Souza (2020), the proposed model's adequacy evaluation is an important task, as it allows for the detection of deviations from the assumed response distribution, as well as detecting incorrect model specifications and inadequate adjustments. Additionally, Machado (2012) states that the best criterion to choose a forecast/prediction model is its predictive capacity, in other words, how close the predictions are to subsequently observed values.

Therefore, this study aims to: i) apply statistical forecast methods to data from insurance companies that work in the 1391 branch, available on SES (SUSEP's Statistics System - *Sistema de Estatísticas*, in Portuguese), in order to estimate loss ratios in the first semester of 2022; ii) assess the adjustment and applicability of the time series prediction methods by comparing them to real data available in the SUSEP database. The results of this analysis indicate that the regression and Box-Jenkins models have shown a satisfactory performance in predicting life insurance claims. However, the regression model has shown to be more adequate for this purpose.

## 2. THEORETICAL FRAME OF REFERENCE

### 2.1. Historical context

The concern in safekeeping and protecting our own life is a part of human instinct that has been present since the dawn of humanity. However, the first life insurances similar to what we know today were only structured amid the Industrial Revolution. They were motivated by

precarious life and work conditions, in which workers endured excessive work hours and had nothing to protect them. Disabled people at the time were simply dismissed and replaced, and similar cases illustrate the importance of insurance. Still in the 19th century, Otto Von Bismarck structured the first social insurance model, which covered illness, invalidity and workplace accidents, as a response to worker strikes and mobilizations at the time.

The first insurance registered in Brazil was the Marine kind. The Brazilian Commercial Code (Law Nº 556, from June 25th, 1850), forbade life insurance up until 1855, when it was authorized. As stated by Lima (2018), given the international market's interest on the life insurance rate in the country, Law Nº 294 was enacted in September 5th, discussing exclusively international life insurance companies. The law determined that these companies had technical reservations and resources invested in Brazil, to counter the risks taken here. The Evolution of the insurance field in the country sparked a need for the creation of an agency that oversaw these market operations, in order to ensure free competition and stability and respect for insured persons. As a result, the Superintendence of Private Insurance (*Superintendência de Seguros Privados*, SUSEP, in Portuguese) was created in 1966, and is responsible for the authorization, control and monitoring of the insurance markets, open supplementary pension plan, capitalization and reinsurance in Brazil. There is also a superior office that establishes general guidelines for the insurance field, the National Private Insurance Council (*Conselho Nacional de Seguros Privados*, CNSP, in Portuguese), whose president is also the country's Minister of Finance. CNSP is also responsible for deliberating on last instance pendings in the insurance field, regulates mandatory insurances and establishes the limits of insurance operations in the country.

## 2.2. Individual Life Insurance

In Brazil, this kind of insurance can be divided in survival coverage or risk hedge, also known as death coverage, and each type has its own specific laws. The first, governed by the CNSP Resolution Nº 348/2017, is structured under the capitalization financial regime and is defined as the coverage which allows payment of the insured capital, through the survival of the insured at the time of the hiring, or the purchase of immediate income, upon a one-time payment. Overall, there are several types of plans and modalities.

This study will investigate Life Insurance by Death Coverage, which is defined by the CNSP Resolution Nº 117/2004 as a life insurance coverage in which the claim event is not the

insured's survival to a pre-defined date. Unlike survival, this coverage admits Simple Distribution (RS), Capitalization (RC) and Coverage Capitals (RCC) financial regimes, and must be structured only in the defined benefit modality. Moreover, though there are collective and individual contracts, this study investigates only Individual Life Insurance.

Life insurance, funeral insurance, personal accidents insurance, educational insurance, travel insurance, loan insurance, daily insurance for hospital stays, loss of income insurance, and daily insurance for temporary incapacity are all examples of persons insurance with risk coverage. This paper will only analyze Life Insurance, identified by SUSEP as branch 1391. Depending on the hired plan, there are many insurance coverages that can be sold together or separate. According to the SUSEP Newsletter Nº 395/2009, which establishes the coding of insurance branches, the ones accounted for in the Individual Life Insurance Line are Death by Any Cause (*Morte por Qualquer Causa*, MQC, in Portuguese), Invalidity Caused by Disease and Invalidity by Any Cause (disease or accident), depending on each case.

SES data shows that Persons Insurance has been growing expressively throughout the years in Brazil, and currently represents 16% of the revenue accumulated by segment in the country. Especially, Individual Life has shown an increase of 215% in the premium volume over 4,5 years (Jan/2017 to Jun/2022), which is around billions. Regarding claims, the absolute values are way smaller – in the millions, as expected, and the increase is not proportional to premiums. In the same 4,5 years, it has only increased around 58%.

From this data is possible to calculate the loss ratio index, observing the ratio between claims and premiums. This rate is essential for insurance companies because it allows them to analyze whether the degree of risk acceptance and pricing are adequate. According to data from SUSEP, it is possible to observe that the loss ratio does not oscillate much, except for the significant raise during the Covid-19 pandemic. Outside of that period, this index varies between 20% and 30%. At its maximum, in 2021, it hit approximately 44%, which is high, but in accordance with the scenario.

### 2.3. Related works

Vianna (2018) carried out a study which analyzes the loss ratio applied by private health insurance companies, with the purpose of determining the factors associated by this index and its possible impacts. Through her work, the author concluded that the loss ratio modifies according to changes in a few Heath Insurance Companies (HIC) factors, such as the number

of beneficiaries. Furthermore, she presented strategies to reduce accidents, for instance, preventive measures with respect to the financial-economic situation, by keeping inhouse resources that help the HIC honor their future commitments and the importance of financial availability, which helps the HIC honor short-term payments.

Still in the health insurance field, Guimarães and Alves (2009) developed a bankruptcy prediction model for HICs, in order to anticipate the financial capacity of these institutions to fulfil their contractual obligations towards clients and health service providers. For that end, the authors adjusted a logistic regression model from 17 financial indicators, taking 600 Brazilian health insurance companies into account. By comparing the proposed model to a more generic one, the authors concluded that the proposed model's performance was more precise.

A study by Carvalho and Gonçalves (2022) models the Supplementary Health Performance Index (IDSS) by evaluating the impacts of factors like region, preliminary intermediation notifications, modality and size of the HICs in the index prediction, which consists of quality of attention to health dimensions, guarantee of access, market sustainability and regulation processes' management. The authors validated and selected the proposed models through the evaluation of adequacy measures, arriving at the conclusion that the beta regression model has shown a better adjustment for the IDSS forecast.

In her descriptive analysis paper, Correa (2022) discusses the impacts of the Covid-19 pandemic on life insurance companies, in particular the increase of the loss ratio. Through this research the author concluded that insurance companies were capable of managing and absorbing the increase in this index satisfactorily. As a general result, this reflex can be observed in other operational ratios evaluated to assess the insurance companies' situation. On the other hand, on the subject of time series, Mori and Gonçalves (2016) have analyzed data from the Brazilian and American health care systems between 2000 and 2012, by using statistical models to assess time series. The time series models were adjusted to predict the number of people who own private health insurance in Brazil, while a regression model was adjusted to estimate new enrollments to the 14 health insurance plans evaluated in the United States. These models were developed to assist future decision-making, by understanding these variables' tendencies and guiding management in a more efficient manner.

In the context of the Brazilian insurance market there are few studies that apply time series models to evaluate ratios that allow for a financial-economic and actuarial analysis of the institutions at risk, especially for medium and long-term predictions. In that sense, this research aims to contribute as literature on this subject.

ISSN: 2316-2341



## 3. DATA AND METHODOLOGY

### 3.1. Data

The individual life insurance market is not uniformly distributed, resulting in a lack of proportionality in the premium volume between insurance companies that operate on this field. Among the 43 institutions that operate with this kind of insurance available at SES, the two largest concentrate approximately 68,24% of the premium amount. For this study, we have selected companies with a premium revenue greater or equal to 0,5 billion, given that this limit has selected the insurance companies that have over 90% of premium volume. Through this filtering process, 8 companies were selected. Their names were redactd in order to protect their image. These informations are available in the SUSEP "insurance market intelligence" panel.

Additionally, the variables selected to integrate the database were: Reference Date, Retained and Occurred Loss and Earned Premium. For the predictions we have used Competence and Loss Ratio, the last being calculated from the Earned Premium and the Losses values. SUSEP defines them as:

- Earned Premium: Issued Premium + Technical Provisions Variation.
- Amount of Claims: Notified Claims + Expenses + Accepted Coinsurance Amount – Ceded Coinsurance Amount – Savings and Reimbursements + Assistance Services + Technical Provisions Variation.

Furthermore, we have considered all the months which had complete data available in the SES. In that way, the series examined starts from January 2011, which represents 132 monthly observations or 11 full years, from 2011 to 2021. To evaluate the adequacy of the methods proposed we have collected informations from the 1st semestre of 2022, in order to compare the models' forecast.

### 3.2. Forecast Method by Regression

In a simple linear regression process it is assumed that any $f(\cdot)$ can be approximated in a straight line, such as

$$f(x) = \beta_0 + \beta_1 x + \varepsilon_i, \quad (1)$$

in which $\beta_0$ is intercepted, $\beta_1$ is the slope of the line, and $\varepsilon_i$ is the random error term. The goal of this model is to be a simplified approximation of the real relation between the variables of interest.

Linear regression, in a time series context, uses the least squares method to project a linear equation that best fits the presented data, consequently predicting the following months. $Y_t$ is assumed as an output or dependent time series, for $t = 1, \dots, n$. This Variable is being influenced by a collection of possible entries or independent series, such as $X_{t1}, X_{t2}, \dots, X_{tq}$, considered as fixed known inputs. Thus, this relation can be expressed by:

$$Y_t = \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_q X_{tq} + \varepsilon_t, \quad (2)$$

in which $\beta_1, \beta_2, \cdots, \beta_q$ are fixed unknown regression coefficients, $\varepsilon_t$ is a random error or noise process, and $\varepsilon_i \sim N(0, \sigma^2)$ (read as the errors that follow normal distribution with zero average and σ² variance). For further details on regression models in a time series context we suggest reading Wooldridge (2015).

### 3.3. Forecasting Method through Time Series Models

Shumway et al. (2011) describe a time series as a collection of observations {Y(t), t ∈ T} in which Y is the variable of interest that can either be discrete, continuous or multivariate, and T is the set of indexes. According to the authors, the obvious correlation shown by the sample of adjacent points in time can severely restrict the applicability of many conventional statistic methods, traditionally dependent on the assumption that these observations are independent and identically distributed. The systematic approach that answers the statistical questions raised by these time correlations are referred to as time series analyses.

The prediction of future values can be made through linear regression, as well as through more complex statistical methods, such as the autoregressive moving average models (Antunes et al., 2015). The purpose of these methods is to distinguish the pattern from any noise that may exist in the observations and later use that pattern to forecast future series values. Time series can be composed of the following factors:

- Tendency, which is the long-term series behavior;
- Seasonality, which are periodic fluctuations in the variable values;
- Cycles, which are those that provoke rise and fall oscillations in the series, softly and repetitively, throughout the tendency component;

- Error or irregular variation, which occurs due to an effect caused by one of the other factors. In the error components fluctuations show in a short period with an inexplicable deviation.

The models used to describe a time series are stochastic processes, controlled by probabilistic laws. Taking T as an arbitrary set, a stochastic process is a family $\{Y_t, t \in T\}$, in which for each $t \in T$, $Y_t$ is a random variable. A stochastic process is determined when its probability distribution functions are known. When they are not known, and one has only a sample of the process (the observed time series), the stationarity and ergodicity of the stochastic process are assumed. According to Morettin and Toloi (2006), these factors related to the variability of observations can be defined as:

- Stationarity: the statistics are not affected by variations in time. Therefore, a series is stationary if the two initial moments are constant through time, that is to say, $E(Y_t) = \mu$ and $Var(Y_t) = \sigma^2$.
- Ergodicity: if only one realization of the stochastic process is enough to obtain the entirety of its statistics. A process is ergodic if its time and sampling averages are the same when T tends to infinity.

Every ergodic process is also stationary, because the realization of a non-stationary process cannot contain all of the information needed to specify such process.

Diniz et al. (1998) state that a time series is stationary if the average $E(Y_t) = \mu$, which calculates the average data value; the variance $Var(Y_t) = \sigma^2$, which measures the degree of dispersion between a specific data and its subsequent, are constant over time. A unitary root test can verify that condition, given that the presence of one or more unitary roots indicates a non-stationary behavior in a historical series, or in other words, values tend to increase as time passes.

According to Granger and Newbold (1974), by adjusting a regression model, even if the parameters' significance is determined and the regression determination coefficient ($R^2$) is raised, it could still result in a spurious regression when the variable has a unitary root. That happens because the assumptions that the average and variance must be constant through time are violated, compromising results. Therefore, in a non-stationary series, the future forecast will not be efficient. In that sense, in order to verify the stationary condition, a unitary root test must be performed. For further details we recommend reading the aforementioned references.

ISSN: 2316-2341



### 3.3.1. Autocorrelation Functions

The autocovariance is the covariance between $Y_t$ and its value $Y_{t-k}$ separated by k time units, and is represented by:

$$\gamma_k = Cov[Y_t, Y_{t-k}] = E([Y_t - \mu][Y_{t-k} - \mu]), \quad (3)$$

for k = 0, $\pm 1$, $\pm 2$, ....

Autocorrelation, or Autocorrelation Function (ACF) is the standardized autocovariance, which determines the length and memory of a process. To clarify, it is the extension for which the value taken in t time depends on the one taken in t-k ou t+k time, represented by:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{Cov[Y_t, Y_{t-k}]}{\sqrt{Var(Y_t)Var(Y_{t-k})}}. \quad (4)$$

This idea of autocorrelation can be extended, that is, it is possible to calculate de correlation between two observations, eliminating the dependency on intermediary terms. Partial Autocorrelation Function (PACF) is expressed by:

$$Cov\left[Y_t, Y_{t-k} \middle| Y_{t-1}, \dots, Y_{t-(k+1)}\right]. \quad (5)$$

### 3.3.2. Autoregressive Integrated Moving Average Model – ARIMA

For Shumway et al. (2011), a classical regression, such as the one mentioned in 3.2, is at times not enough to explain every dynamic present in a time series. Otherwise, introducing correlation as a phenomenon that can be generated through lagged linear relations leads us to suggest autoregressive models (AR) and autoregressive moving average models (ARMA). Though complex, these adjust to seasonal and tendency factors, estimate adequate weight parameters, test the model and repeat the cycle when necessary.

The ARIMA model, proposed by Box and Jenkins (1994), is a general ARMA model applied to cases in which the data is non-stationary. The non-stationarity issue can be resolved through differentiations, which correspond to the integrated part (I) of the model. Both models can be used to forecast a series, given their past points. For these time intervals, Box-Jenkins recommend a minimal of 40 to 50 points. The data used in this study obliges to that requirement, given that the sample size is 132. The Box-Jenkins prediction is most useful when it is believed that subjacent factors cause the demand for products, services, income, and, in this case, loss ratio to behave in the same way in the past and the future.

The autoregressive part of models is structured by $Y_t$, which represents the current values of the series explained as a function of p past values. It is represented by AR(p) and demonstrated by:

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t, \quad (6)$$

In which $Y_t$ is stationary and the $\varphi_1, \varphi_2, \ldots, \varphi_p$ coefficients are constant ($\varphi_p \neq 0$). It is assumed that $\varepsilon_i \sim N(0, \sigma^2)$ are independent. Hence, this part of the model indicates that the variable is regressed to its own lagging values. On the other hand, assuming that white noises $e_t$ structure Moving Averages in a q order, MA(q) is expressed by:

$$Y_t = \varepsilon_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}, \quad (7)$$

in which $\theta_1, \theta_2, \ldots, \theta_q$ are parameters and $\theta_q \neq 0$. Again, we assume that $\varepsilon_i \sim N(0, \sigma^2)$, unless indicated otherwise. This consequently suggests that the regression error is actually a linear combination of the error terms, with values that occur simultaneously and on several moments in the past. It is important to point out that, unlike AR(p), the MA(q) is stationary for any $\theta$ parameter values.

Furthermore, the integrated part (I) indicates that the given values were replaced with the difference between its current and previous values and this differentiating process could have been carried out more than once, as explained above. In the results of this study, we will show that that is not the case, and that it was not necessary to differentiate the series because it is already stationary. Finally, the parts come together, forming the ARMA and ARIMA models, depending on the stationarity. Therefore, the model can be expressed as:

$$Y_t = \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \quad (8)$$

in which the q and p parameters represent autoregressive orders and Moving averages, respectively. In this paper, in order to calculate the forecast for future values with the model above, we have used the *arima* function from the *stats* pack in the R software.

## 4. RESULTS

### 4.1. Descriptive Analysis

The original data had two negative outliers and two loss ratio values higher than 1, actually being considerably higher in the third quarter. These rates correspond to the months Jan/11, Dec/11 May/16 and Dec/17. In that way, aiming for the most precise forecast possible with the available data, these outliers were adjusted. For outliers greater than 1, we used the average of the 10 highest loss ratio values lower than 1. Secondly, for the negative outliers we

have used an average of the 10 lowest loss ratio values greater than zero. An overview of the new data compared to the originals can be seen on Table 1.

**Table 1: Original and Adjusted Descriptive Data Statistics**

| Descriptive Statistics | Claims | | Loss ratio | |
|---|---|---|---|---|
| | Original | Adjusted | Original | Adjusted |
| Minimum | -4.046.445,00 | 2.335.441,00 | -0,0800 | 0,0800 |
| 1st Quarter | 9.465.237,00 | 9.465.237,00 | 0,2100 | 0,2100 |
| Median | 15.674.779,00 | 15.674.779,00 | 0,2600 | 0,2600 |
| Average | 20.821.577,00 | 20.900.212,00 | 0,2940 | 0,2852 |
| 3rd Quarter | 25.194.276,00 | 25.194.276,00 | 0,3225 | 0,3225 |
| Maximum | 95.432.888,00 | 95.432.888,00 | 1,4900 | 0,6400 |
| Standard Deviation | 18.232.735,00 | 18.134.014,00 | 0,1750 | 0,1197 |

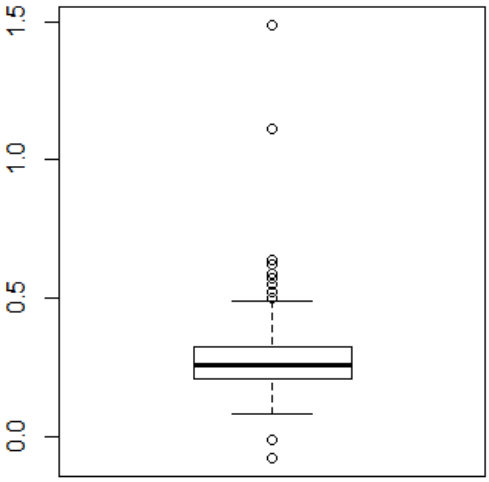
Source: Research data.

It is possible to observe that the minimum and maximum loss ratio values are closer to the average and median without outliers. It is also important to point out that even with the adjustment proposed for the values greater than 1, there are a few rates that are far too distant form the reality of the others, as analyzed in the third quarter values, that indicate that 75% of the data have shown a loss ratio of up to 0,32, in comparison to the maximum loss ratio of 0,64. The effectiveness of these adjustments can be better visualized through the following data loss ratio boxplots with and without outliers below.

Figures 1 and 2 indicate that the highest loss ratio limit is approximately 0,5. Though values over this limit are defined as outliers, they will be considered in the forecast because they show up frequently in the data set.

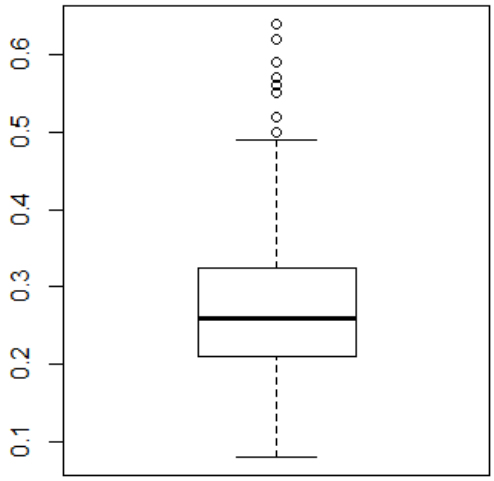
ISSN: 2316-2341

**Figure 1: Loss Ratio Boxplot (original)**




Source: Made by the authors.

**Figure 2: Loss Ratio Boxplot (adjusted)**




Source: Made by the authors.

Through the adjusted data is possible to verify that loss ratio rates are mostly inferior to 0,35. The reduced proportion between claims expenses and the awarded premium have a direct impact on the financial solidity of these insurance companies, positively influencing the profit retention capacity. Moreover, this decrease helps proof the quality of the subscription and the effectiveness of the calculations for the risk pricing process. From this moment on, we have used the adjusted database for further analysis.

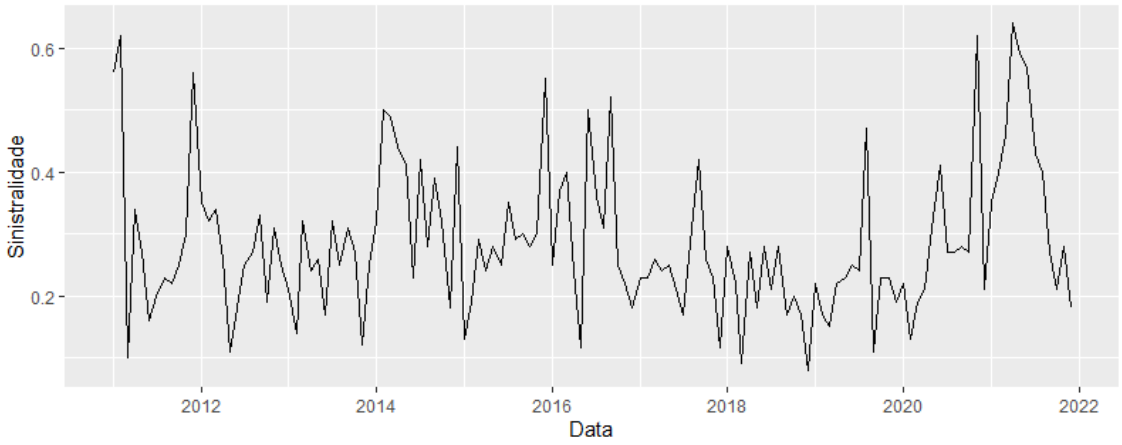
ISSN: 2316-2341





Figure 3 shows the complete series through time. From the graph it is possible to observe that the stationarity factor, essential for calculating the forecast by time series, will probably be met. This hypothesis will be proven through the unitary root tests.

**Figure 3: Loss Ration in the 2011 to 2021 period**




Source: Research data.

## 4.2. Normality and Unitary Root Tests

In order to test the normality of the date we have used the Shapiro-Wilk test. From the p-value of the test and considering a significance level of 5%, it is possible to conclude that the null hypothesis of the normality of the data is rejected. Considering this, the series data were transformed through the logarithm (the transformation tends to suppress bigger fluctuations that occur in parts of the series) and the test was administered for the transformed variable, indicating, at a 5% significance level, statistical evidence in favor of the normality hypothesis.

For the unitary root analysis, the augmented Dickey Fuller Test (ADF) was the first carried out. The null hypothesis of this test is that the series has a unitary root. If the statistic value, intrinsically calculated in the ADF test, is higher than the absolute value tabulated by Dickey-Fuller, the null hypothesis is accepted, and consequently, the series is considered non-stationary. We have used the *adf.test* of the *tseries* pack in the R software to carry out the test, wich showed statistical evidence in favor of the stationarity hypothesis.

Nusair (2003) proposes that crossing ADF and Kwiatkowski Philips Schmidt & Shin (KPSS) tests ensures a more precise conclusion on the stationarity of the series. Thus, in order to reduce the uncertainty of the ADF tests we have also carried out the KPSS test, by using the *PP.test* function, also from the *tseries* pack, which also pointed to statistic evidence in favor of the stationarity hypothesis.

ISSN: 2316-2341



### 4.3. Regression Models Results

In this section we have tested three different models. In the first, the Loss Ratio was adjusted by the time period t (which Jan/11 to Dec/21), plus t squared, corresponding to the tendency. The second model adds the seasonality factor into Model 1. Lastly, the third adds to Model 2 the autoregressive term of order 1 to the error terms $e_t$.

Next, we will present the comparison criteria, considering the normalcy of residues through the Shapiro-Wilk test and the Durbin-Watson statistic. Durbin-Watson (DW) statistic is an autocorrelation test of the first order of regression residues with values between 0 and 4. Values between 0 and 2 indicate a positive autocorrelation, while values between 2 and 4 indicate a negative autocorrelation. A value equal to 2 indicates that an autocorrelation was not detected in the sample. For the model selection in the case of a tie, we will consider the Akaike Information Criterion (AIC), with lower values being better.

**Table 2: Criteria for the Regression Models' Selection**

| | AIC | Durbin-Watson Statistic | Durbin-Watson p-value | Shapiro-Wilk p-value |
|---|---|---|---|---|
| Model 1 | 147,90 | 1,5144 | 0,004 | 0,0394 |
| Model 2 | 165,47 | 1,5232 | 0,004 | 0,1330 |
| Model 3 | 160,47 | 2,0535 | 0,862 | 0,1246 |

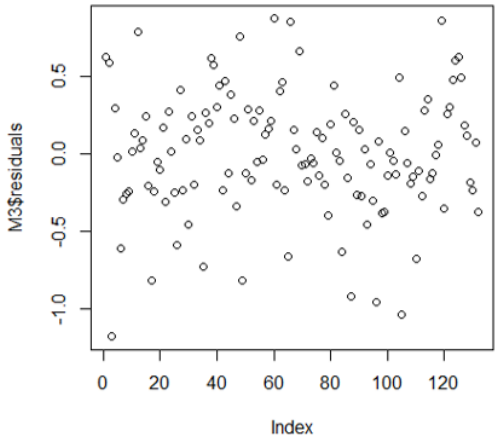
Source: Research Data.

Table 2 indicates that Model 1 was the only one that rejected the residue normalcy test (SW), in a significance level of 5%. Even though the Model 3 AIC is higher than the first model's, firstly we must evaluate the Durbin-Watson test, which indicates that Model 3 is the most adequate, given that the DW statistic is approximately 2, while the rest are between 0 and 2. For a more detailed analysis of the selected model, see the Model 3 residue graphs below.

From Figure 4 we can conclude that the residuals are well distributed and do not manifest a pattern, meaning a random cloud of points around zero. Figure 5 demonstrates that the residues show a slight asymmetry, indicating that an approximately normal distribution fits the data adjustment.

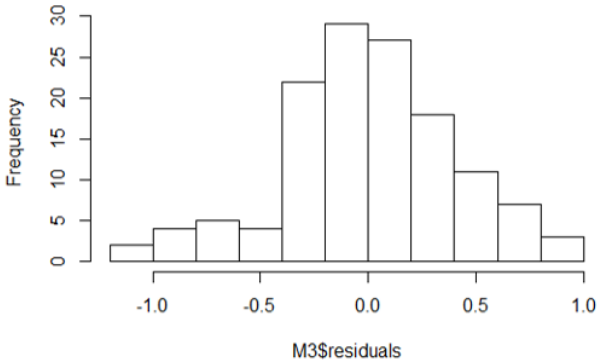
ISSN: 2316-2341

**Figure 4: Model 3 Residue Dispersion**




Source: Research Data.

**Figure 5: Model 3 Residue Histogram**




Source: Research Data.

Finally, from the described methodology and the chosen model, we arrive to the forecast for the 1st semester of 2022, as shown in Table 3. Real values were also included in order to compare the observed and predicted values.

**Table 3: Real versus Predicted Loss Ration – Regression Model**

| Month | Real | Prediction | Inferior IC | Superior IC |
|---|---|---|---|---|
| Jan/22 | 0,23 | 0,28 | 0,13 | 0,60 |
| Feb/22 | 0,29 | 0,30 | 0,13 | 0,66 |
| Mar/22 | 0,38 | 0,28 | 0,12 | 0,65 |
| Apr/22 | 0,34 | 0,33 | 0,14 | 0,81 |
| May/22 | 0,22 | 0,29 | 0,11 | 0,75 |
| Jun/22 | 0,29 | 0,31 | 0,11 | 0,89 |

Source: Research Data.

The forecast precision is notable, especially when compared to the real values observed for the period's loss ratio, given that all the confidence intervals contain observed values. Though there is a difference of approximately 30% in March and May, April and February show differences inferior to 3%.

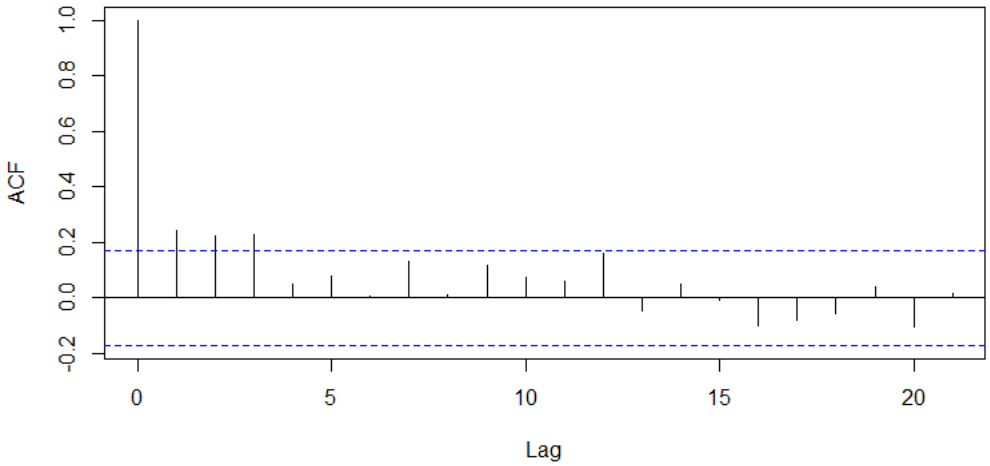
ISSN: 2316-2341



## 4.4. ARIMA Model Results

In a stationary process, the Autocorrelation Function (ACF) quickly drops to zero, indicating that the correlation regarding past time decreases exponentially. On the other hand, a slow decrease of the ACF suggests that the series in question is non-stationary and requires differentiation. Therefore, it is possible to observe 3 lags beyond the critical point in Figure 6.

**Figure 6: Loss Ratio Autocorrelation Function**

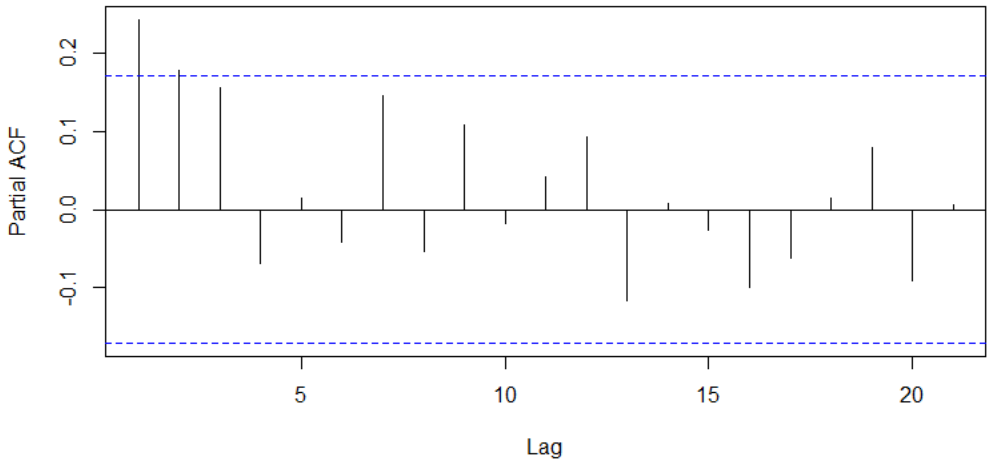



Source: Research Data.

In the autorregressive process, the Autocorrelation Funciton (ACF) shows an exponential decrease, or cushioned sinusoidal. In turn, the Partial Autocorrelation Function (PACF) shows significant peaks in lags 1 through p, followed by an abrupt drop to zero. Based on Figure 7, it is possible to identify 2 lags outside of the critical point.

ISSN: 2316-2341

**Figure 7: Loss Ratio Partial Autocorrelation Function**




Source: Research Data.

The combination of the AR(p) and MA(q) models originates the ARMA(p,q) model. The p parameter from the autoregressive part can be identified on the PACF graph. The identification is possible through the number of lags outside of the critical points in the graphs. Therefore, given the 3 points observed in the ACF and the 2 seen in the PACF, we can conclude that the model is an ARMA(2,3) or an ARIMA (2,0,3). However, to further minimize the number of model parameters, we have used the *coeftest* function from the *lmtest* pack in the R software. Through this function we could determine which coefficients are in fact significant, complementing the autocorrelation graphs' analysis. In the tests we have considered the coefficients with the highest p-value until we arrived at a final result, an ARIMA(1,0,1).

Complementing this inferential process, we also used the *auto.arima* function from the *forecast* pack. This function uses a variation of the algorithm developed by Hyndman anf Khandakar (2008), which combines unitary root tests to obtain an ARIMA model. Lastly, among all the options analyzed by the methods mentioned above, we selected the models with the lowest AICs, as shown in Table 4.

**Table 4: AIC Criterion for the Selection of ARIMA Models**

| Modelo | AIC |
|---|---|
| ARIMA(1,0,1)(0,0,1)[12] | 134,31 |
| ARIMA(1,0,1)(1,0,0)[12] | 134,64 |
| ARIMA(1,0,1) | 136,86 |
| ARIMA(2,0,2) | 137,40 |

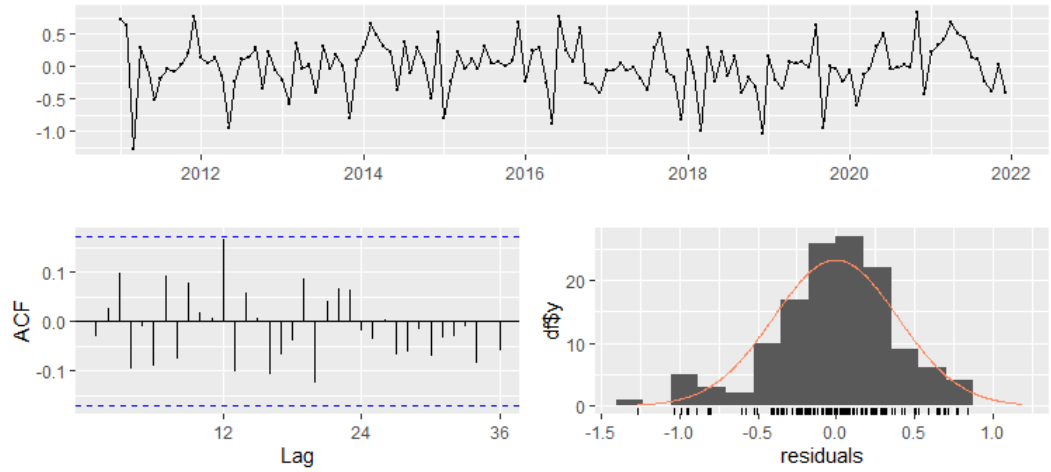
Source: Research Data.

Therefore, Table 4 indicates that the best model for loss ratio is the ARIMA(1,0,1)(0,0,1)[12]. This result validates the previously performed analysis to estimate the model, adding a term that refers to the development into the seasonal part. The 12 number between brackets is the number of periods, which goes along with the data, that are annual observations divided between the months January through December.

### 4.4.1 Diagnostic Measures

Next, we will consider for diagnostic the ARIMA(1,0,1) and ARIMA(1,0,1)(0,0,1)[12] models, in order to analyze the effect of the possible seasonality in the forecast. It will be necessary to verify whether the model residues are autocorrelated, and for that, we will use the Ljung-Box test. Its null hypothesis is that the observed model's residues are jointly uncorrelated through time. Thus, the non-independency of the residues would indicate a flaw in the structure of proposed model. The ARIMA(1,0,1) model's p-value was 0,4557 and the ARIMA(1,0,1)(0,0,1)[12] model's one was 0,8859. Therefore, considering a 5% significance rate, there are evidences to conclude that both models' residues are correlated.

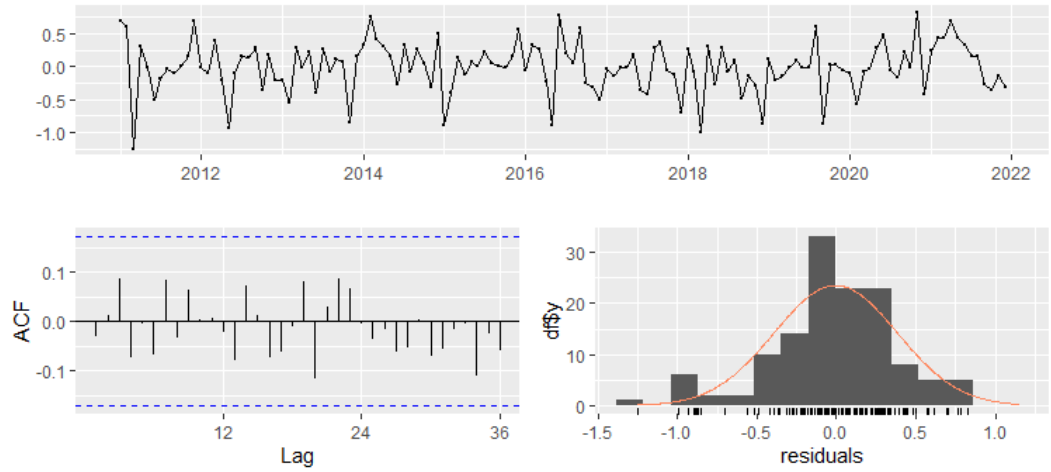
**Figure 8: Residual Analysis for the ARIMA(1,0,1) Model**




Source: Research Data.

**Figure 9: Residual Analysis for the ARIMA(1,0,1)(0,0,1)[12] Model**

ISSN: 2316-2341




Source: Research Data.

To further complement the tests, the residues' correlograms are exposed in Figures 8 and 9, represented above. From them, it is possible to observe that both series residues are well distributed over time. Also, there are no lags in the ACFs that surpass the limit lines. Lastly, we can conclude from the histograms that both models, with and without seasonality, show symmetry surrounding zero.

### 4.4.2 Forecasts

Following the former subsection's logic, two forecast sequences were carried out, each using an ARIMA model for the data from the 1st semester of 2022, as seen in Table 5 below.

**Table 5: Loss ratio - Real versus Forecast by ARIMA Models**

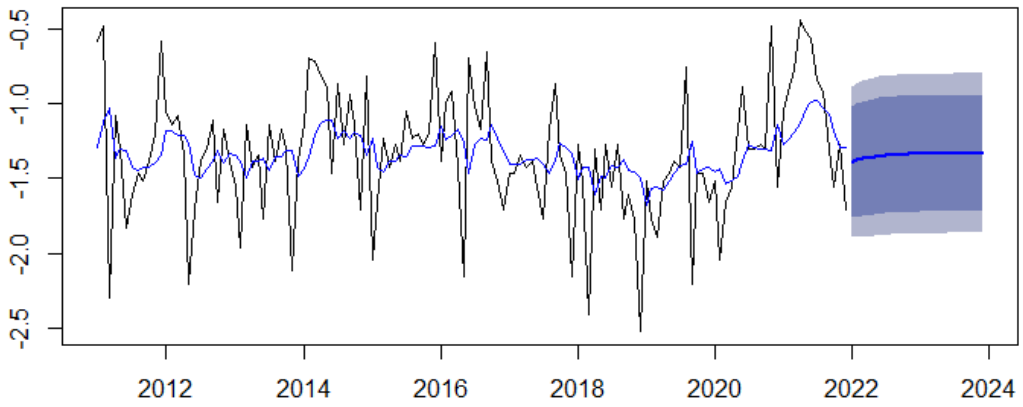
| Month | Real | Forecast ARIMA(1,0,1) | Forecast ARIMA(1,0,1)(0,0,1)[12] |
|---|---|---|---|
| Jan/22 | 0,23 | 0,39 | 0,39 |
| feb/22 | 0,29 | 0,40 | 0,40 |
| Mar/22 | 0,38 | 0,41 | 0,41 |
| Apr/22 | 0,34 | 0,41 | 0,41 |
| May/22 | 0,22 | 0,41 | 0,41 |
| Jun/22 | 0,29 | 0,41 | 0,41 |

Source: Research Data

ISSN: 2316-2341

The time series projection where the blue line represents the adjusted values will be presented along with confidence intervals, considering 65% pictured in light gray) and 80% (in dark gray) levels. By analyzing the ARIMA(1,0,1) model's forecast, it is possible to notice, both from Figure 10 and Table 5, that this model projects a constant increase series. The disadvantage of this model is that the projection does not take into consideration the ups and downs in the observed series.

**Figure 10: Loss Ratio Projection - ARIMA(1,0,1) Model**

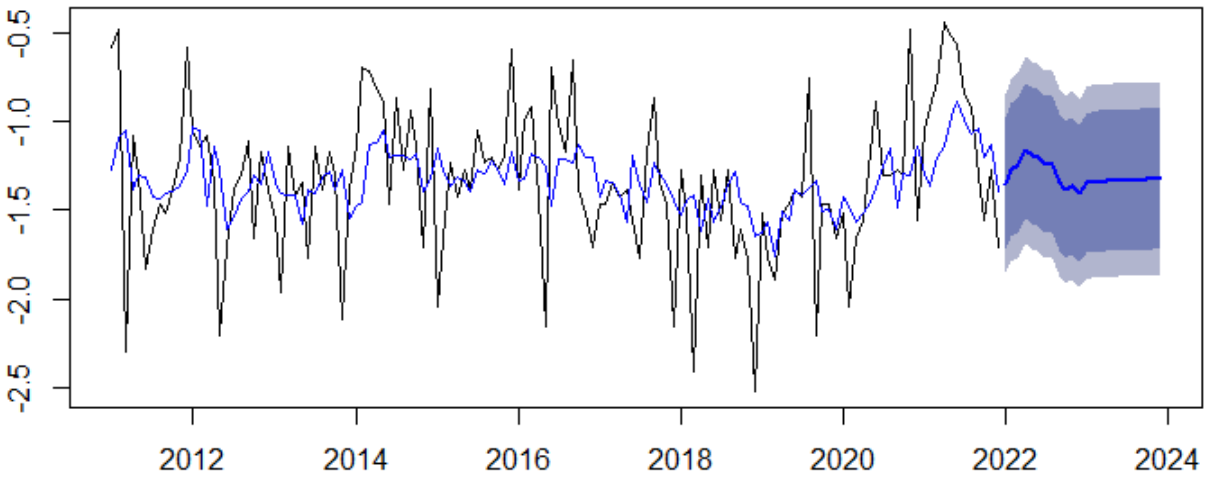



Source: Research Data.

In the ARIMA(1,0,1)(0,0,1)[12] model's analysis is possible to observe, based on Figure 11, that the projection of this model seems to capture more similarly the variations present in the series' observed data. However, it is important to underline that, given that only the real data of the first semester of 2022 were available for comparing with the forecast, there is no difference between both models' predictions, as shown in Table 5.

**Figure 11: Loss Ratio Projection – ARIMA(1,0,1)(0,0,1)[12] Model**

ISSN: 2316-2341




Source: Research Data.

In relation to the forecasts, it was noticed that the smallest difference between the forecast and observed values was 0,03, while the biggest difference was 0,19. This fluctuation is expected, once we are dealing with a non-deterministic variable. The variation that is not explained by the model can be attributed to several factors, many of which are hard to measure (latent), including specific characteristics of the analyzed period, insurance companies' particularities as well as their subscription processes, among others. Generally, the predictions are coherent, especially after the Covid-19 crisis and the ratio increase in 2021.

### 4.5 Comparison between Regression and ARIMA Models

The statistic metric of Mean Square Errors (EQMs) is amply used in the literature to compare the performance of models in relation to their predictive capacity. Hence, to compare and select between the proposed models for loss ratio forecast in this study, we have opted for the EQM method. The lower the EQM, the better the model adjustment quality, resulting in a more precise forecast.

**Table 6: Mean Square Error for the Regression and ARIMA Models**

| Statistic | Regression | ARIMA(1,0,1) | ARIMA(1,0,1)(0,0,1)[12] |
|---|---|---|---|
| EQM | 0,00283 | 0,01629 | 0,01595 |

Source: Research Data.

From the statistics presented in Table 6, the regression model has the lowest MSE. Consequently, we can conclude that this model best predicted the loss ratio for the 1st semester of 2022, according to the adjustment made with the data from January 2011 to December 2021. It is important to underline the proximity between the ARIMA(1,0,1) and the ARIMA(1,0,1)(0,0,1)[12] MSEs, which was expected, given that the projections were similar, as observed in the 4.4.2 subsection on forecasts.

## 5. FINAL CONSIDERATIONS

After an extensive analysis of the data observed between 2011 and 2021, this paper has indicated three different models to forecast loss ratio in life insurances for the 1st semester of 2022. In general, the values are contained within the confidence interval, though the comparative analysis of the models suggest that the regression model predicted the loss ratio taxes for the analyzed period more accurately, thus reaching the goals of this research. We hope this study contributes to the literature on the theme, given that there are few studies on the insurance market.

For possible future works we suggest studying the applicability of other forecast models on individual life insurance data. Additionally, the use of the proposed models for other types of insurance, like Group Life Insurance, which have other specificities. Finally, we also suggest a more detailed investigation into the data concerning the Covid-19 pandemic, to analyze the impact of the pandemic in loss ratio rates and verify whether the forecasts would be affected somehow.

## REFERENCES

Antunes, J. L. F.; Cardoso, M. R. A. (2015). Uso da análise de séries temporais em estudos epidemiológicos. Epidemiol. Serv. Saúde, Brasília, 24(3), 565-576. <http://scielo.iec.gov.br/pdf/ess/v24n3/v24n3a24.pdf>

Box, G. E.; Jenkins, G. M. (1994). Time Series Analysis: Forecasting and Control. Prentice Hall, Englewood Cliffs.

Carvalho, G. Q. F; Gonçalves, J. N. (2022). Estimação do Índice de Desempenho da Saúde Suplementar (IDSS): Inferência Utilizando a Classe de Modelos GAMLSS. In: Anais do Congresso de Ciências Contábeis e Atuariais da Paraíba - CONCICAT, João Pessoa. <https://www.concicatufpb.com.br/anaisconcicat-2022>

Chaves, M. E. D.; Mataveli, G. A. V.; Justino, R. C. (2014). Uso da modelagem estatística para monitoramento da vegetação no Parque Nacional da Serra da Canastra, Minas Gerais. Caderno de Geografia, 24(1), 120-132. <https://doi.org/10.5752/P.2318-2962.2014v24nespp120>

CNSP. Resolução CNSP Nº 117, de 22 de dezembro de 2004; Resolução CNSP Nº 348, de 25 de setembro de 2017.

Correa, B. (2022). Os impactos causados nas seguradoras do ramo vida pelo aumento da sinistralidade devido a pandemia de Covid-19. UFRGS.

Diniz, H.; Andrade, L. C. M.; Carvalho, A. C. P.; Andrade, M. G. (1998). Previsão de séries temporais utilizando redes neurais artificiais e modelos de Box e Jenkins. In: Anais do Simpósio Brasileiro de Redes Neurais, 173-178.

Gonçalves, J.N., Barreto-Souza, W. (2020). Flexible regression models for counts with high-inflation of zeros. METRON 78, 71-95. <https://doi.org/10.1007/s40300-020-00163-9>

Granger, C. W. J.; Newbold, P. (1974). Spurious Regressions in Econometrics. Journal of Econometrics, 2, 111-120. <https://doi.org/10.1016/0304-4076(74)90034-7>

Guimarães, A. L. S.; Alves, W. O. (2009). Prevendo a insolvência de operadoras de planos de saúde. Revista de Administração de Empresas, 49, 459-471. <https://doi.org/10.1590/S0034-75902009000400009>

Huebner, S. S.; Black, K. J. (1976). El seguro de vida. Madrid: Editorial Mapfre. <https://documentacion.fundacionmapfre.org/documentacion/publico/es/bib/1110.do>

Hyndman, R. J; Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 27(1), 1-22. <https://www.jstatsoft.org/article/view/v027i03>

Lima, C. A. (2018). A história do seguro. Revista Apólice. <https://revistaapolice.com.br/2018/10/a-historia-do-seguro/>

Machado, M. A. L. (2012). Modelos de previsão aplicados à optimização da gestão das actividades de um Call Center. Universidade de Lisboa, Lisboa. <http://hdl.handle.net/10451/9422>

Morettin, P. A.; Toloi, C. M. C. (2006). Análise de séries temporais. São Paulo: Edgard Blucher.

Mori, F. T. M.; Gonçalves, L. R. (2016). Aplicação da metodologia de séries temporais ao sistema de saúde do Brasil e dos Estados Unidos (2000-2012). Revista Debate Econômico, 4(1). <https://publicacoes.unifal-mg.edu.br/revistas/index.php/revistadebateeconomico/article/view/288>

Nusair, S. (2003). Testing the validity of purchasing power parity for asian countries during the current float. Journal of Economic Development, 28(2), 129-147.

Shumway, R. H., Stoffer, D. S., & Stoffer, D. S. (2000). Time series analysis and its applications, 3, New York: Springer.

SUSEP. Circular SUSEP N° 302, de 19 de setembro de 2005; Circular SUSEP N° 395, de 3 de dezembro de 2009.

SUSEP. Sistema de Estatísticas da Superintendência de Seguros Privados.

Vianna, F. G. (2018). Sinistralidade das operadoras de planos privados de assistências à saúde médico-hospitalar: determinação dos fatores associados a esse índice e seus efeitos. UFMG.

ISSN: 2316-2341

Wooldridge, J. M. (2015). Introductory econometrics: A modern approach. Cengage learning.