International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 9 Issue 9 Series I || September 2020 || PP 34-40

Time series modeling for death incidence in accidents on federal highways in Minas Gerais state, Brazil

Patricia Mendes dos Santos[1], Rafael Agostinho Ferreira[1], Vânia de Fátima Lemes de Miranda[1], Cristian Tiago Erazo Mendes[1], Luciano Antônio de Oliveira[1], Thelma Sáfadi[2]

[1]*(Students of the Department of Statistics/Federal University of Lavras, Brazil)*
[2]*(Professor of the Department of Statistics/Federal University of Lavras, Brazil)*
CorrespondingAuthor: Patricia Mendes dos Santos

ABSTRACT: *MinasGerais is the state with the highest deaths records by accidents on the roads under the responsibility of the Union. Therefore, we understand that it is of great importance to assess the death incidence in traffic accidents on federal highways in the state, because it can help in control and prevention actions. So, our study aims to use time series theory to model these death incidences, among January 2007 and December 2019. We verified the variability of the series, as well as the presence of the trend and seasonality components. We used AIC for choosing the best model between those that we've considered suitable. The best model to describe our time series' behavior was SARIMA(0.1.1)(0.1.1) model with just one intervention. The fitted model was used to make forecasts about future observations in this time series. According to this forecasting, we observed that deaths tend to stabilize over the months from 2020 to 2021.*

KEYWORDS-*SARIMA model, seasonality, trend, forecasting.*

Date of Submission: 20-08-2020 Date of Acceptance: 05-09-2020

I. INTRODUCTION

A time series is a set of observations indexed on time, resulting from a stochastic process. Because they are time ordered data, it is expected that close observations are dependent and, for this reason, it results in analyzes in which these errors have a time-dependent structure. It makes impossible to apply conventional methods, such as linear regression models, which depends on the assumption that close observations are independently distributed.

Thus, one of the main goals of the time series study is to analyze and model the existing correlation between errors. Thus, the dependence between errors is no longer seen as inappropriate, since this information is incorporated into the model. To model a time series, we can use the Box & Jenkins methodology [1], which is based on the assumption that an observation, and/or its error, at a given moment is influenced by what happened in the past. Such models are known in the literature as Seasonal Autoregressive Integrated Moving Average (SARIMA) models. In this paper, time series analysis was used to describe the observations of the death incidence of traffic accidents on federal highways in the state of Minas Gerais.

Traffic accidents are a major cause of mortality, in addition to having a strong impact on health services, as well as on our society. Several factors have been contributed to this reality, such as drivers' carelessness, the growing number of vehicles in circulation, the urban space and traffic signs precariousness, the lack of inspection, and so on [3].

Minas Gerais is the state with the highest deaths and accidents record on the roads under the responsibility of the Union, and remains ahead when it comes to these disasters cost [3]. Brazil has achieved a slightly higher motorization rate and the same mortality rate in recent years [2].

According to National Transport Confederation, Brazilian Roads (BRs) that cross Minas Gerais state recorded 7214 accidents in 2018, which resulted in the 693 people's death. These informationswere based on data from the Federal Highway Police regarding the events among 2007 and 2018. In this period, 250,140 accidents were recorded, almost half of them with victims (114,673 deaths). Just in 2018, there were 69,206 disasters, of which 53,963 had victims. In almost 10% of them lost their lives (5,269 people) [3].

In addition, Minas Gerais follows the national mean of 81 accidents per 100 km, so that BR-381,tolled since December 19, 2008, lives up to the sad title of "Highway of Death". Just in 2018, BR-381 recorded 2,213 accidents and 171 deaths [3].

Among the various government agencies actions in the Traffic National Week, the State Secretariat of Public Security (SESP) of Minas Gerais has promoted educational blitz, in order to draw drivers' attention about the risks, such as drinking and driving, speed excess and the importance of correct seat belt use.

In this sense, we understand that the use of time series theory is of great importance for modeling death incidence about traffic accidents on federal highways in Minas Gerais state, so that it can help in actions for the accidents' control. Thus, after fitting the behavior of the death incidence by SARIMA model, the results were used to make forecasting for the death incidence from January 2020 to December 2021.

II. MATERIALS AND METHODS

Our time series are observations regarding the death incidence in traffic accidents on federal highways in Minas Gerais state, covering the period from January 2007 to December 2019. The data were obtained by consulting the database of the Federal Highway Police Department, at https://portal.prf.gov.br/.

The classic time series model consists of decomposing the Z_t observations as the sum of three components:

$$Z_t = T_t + S_t + a_t \ \#(1)$$

where$T_t$is the trend component, that is, a gradual increase or decrease in observations over time;$S_t$is the seasonal component, that is, the fluctuations that occurred in periods and$a_t$is the random component, with zero mean and constant variance. For time series analysis, we initially have constructed its graph. From its behavior, it was observed whether the series has a trend, seasonality or atypical observations.

However, before inferring about the series' components, it was necessary to verify whether the model is additive or multiplicative. If so, the following Box-Cox transformation was used:

$$Z_t^*(\lambda) = \begin{cases} \frac{Z_t^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ log\, Z_t, & \text{if } \lambda = 0 \end{cases} \ \#(2)$$

In order to verify the need to transform the series data, we observed the graph of the amplitude versus mean and we applied t-Student test for the slope of the fitted linear regression. If this coefficient is statistically significant, the Box-Cox transformation will be necessary. In addition, the transformation parameter estimationλ was performed following the Guerrero's method [6].

To use the SARIMA models, the time series must be stationary, that is, it develops randomly around a constant mean. However, many time series show trend and seasonality. In these cases, a stationary series is obtained through successive differences from the original series. In addition, it's necessary to apply some tests to verify the presence of these components. Trend effect presence is tested by Cox-Stuart test, known as the signal test [7].

For the application of this test, the observations must be grouped by pairs (Z_1, Z_c), $(Z_2, Z_{c+1}), \ldots, (Z_{N-c}, Z_N)$, wherec=$\left(\frac{N}{2}\right)$if N is even and c=$\left(\frac{N+1}{2}\right)$if Nisodd. For each pair$(Z_i, Z_{i+c})$a "+" signal is associatedwhether $Z_i < Z_{i+c}$or"–"otherwise [7]. If the proportion of"+"is statistically equal to$\left(\frac{1}{2}\right)$, the series won't present trend. However, if this proportion is different from$\left(\frac{1}{2}\right)$, this series will present trend, both increasing (more "+" signs) and decreasing (more "-" signs) one.

In order to check for the presence of seasonality, we used a regression fitting for the time series, considering seasonal dummies. If at least one of the coefficients is significant, the series will present component seasonality.After identifying and removing the model components, through simple and seasonal differences, we obtained a stationary series. Box & Jenkins models have been fitted for this series.

When time series has trend and seasonality components, a SARIMA model can be used for fitting it. A SARIMA model SARIMA$(p, d, q)(P, D, Q)_s$, can be written as

$$\Phi(B^s)\phi(B)(1 - B^s)^D(1 - B)^d Z_t = \Theta(B^s)\theta(B)a_t \ \#(3)$$

where$\Phi(B^s) = 1 - \Phi_1(B^s) - \ldots - \Phi_p(B^{sP})$is the seasonal autoregressive polynomial of order P, $\phi(B) = 1 - \phi_1(B) - \ldots - \phi_p(B^p)$ is the non-seasonal autoregressive polynomial of order p, $\Theta(B^s) = 1 - \Theta_1(B^s) - \ldots - \Theta_q(B^{sQ})$ is the seasonal moving averages polynomial of order Q, $\theta(B) = 1 - \theta_1(B) - \ldots - \theta_q(B^q)$is the non-seasonal moving averages polynomial of order q, $(1 - B^s)^D$and$(1 - B)^d$is the seasonal and non-seasonal difference operator, where D and d are the number of differences applied to the series[7].

In addition, a time series may have points of intervention. It consists of a discrepancy that occurred with the data at a certain moment in time, that is, there is an interference in the time series' behavior, for some known or unknown cause [7].

For intervention analysis, the proposed model is given by

$$Z_t = \sum_{i=1}^{K} \nu_i\,(B)x_{i,t} + n_t, \#(4)$$

where$Z_t$is the time series, $K$is the intervention numbers, $\nu_i(B)$is the transference function value, $x_{i,t}$ is the binaryvariablethatindicatesthe intervention point or interval and$n_t$is the model noise, given by SARIMA.

In order to verify the order of the model to be fitted, we analyzed the autocorrelation and partial autocorrelation functions graphs of the stationary series. After defining the fitting models, the parameters were estimated using maximum likelihood.

Box & Pierce test [8] was used to check the adequacy of the fitted model, that is, if the residues are white noise. If the model is appropriate, the test statistic

$$Q(k) = N(N+2)\sum_{j=1}^{K} \frac{\hat{r}^2}{N-j} \#(5)$$

is approximately distributed by a Chi-squared with$K-m$degrees of freedom, where$N$is the observations number in the series,$K$is the lag numbers considered in the autocorrelation and partial autocorrelation functions, $m$is the parameter numbersof the model and$\hat{r}$the estimated residuals. Therefore, if$Q(K) > \chi^2_{K-m}$ the null hypothesis of white noise is rejected.

For model selection, we used the Akaike Information Criterion (AIC). The value of the AIC is given by

$$AIC(m) = \ln(\hat{\sigma}_a^2)\frac{2m}{N} \#(6)$$

where$\hat{\sigma}_a^2$is the residual variance estimated by the model.By this criterion, the best fitted model is that one with the lowest value of AIC.

The model chosen for fitting will be used to make forecasting about future observations. The prediction of$Z_{t+\hbar}$for h=1,2,..., is defined as the conditional expectation of$Z_{t+\hbar}$ given all past values. Thus,

$$Z_t(\hbar) = E[Z_{t+\hbar}|Z_t, Z_{t-1}, \dots]\#(7)$$

This equation will be used to predict the death incidence of accidents in Minas Gerais state among January 2020 to December 2021.All analysis were performed using R software [9], with the support of the forecast package [10].

III. RESULTS AND DISCUSSIONS

Fig. 1 shows the plot that describes the behavior of the death incidence in traffic accidents on federal highways in Minas Gerais state over time. The time series plot shows that there is a trend behavior in the series, which makes it non-stationary. Thus, the Cox-Stuart test [7] was applied in order to verify, statistically, the presence of this component. In addition, there was no evidence of seasonality in the series.

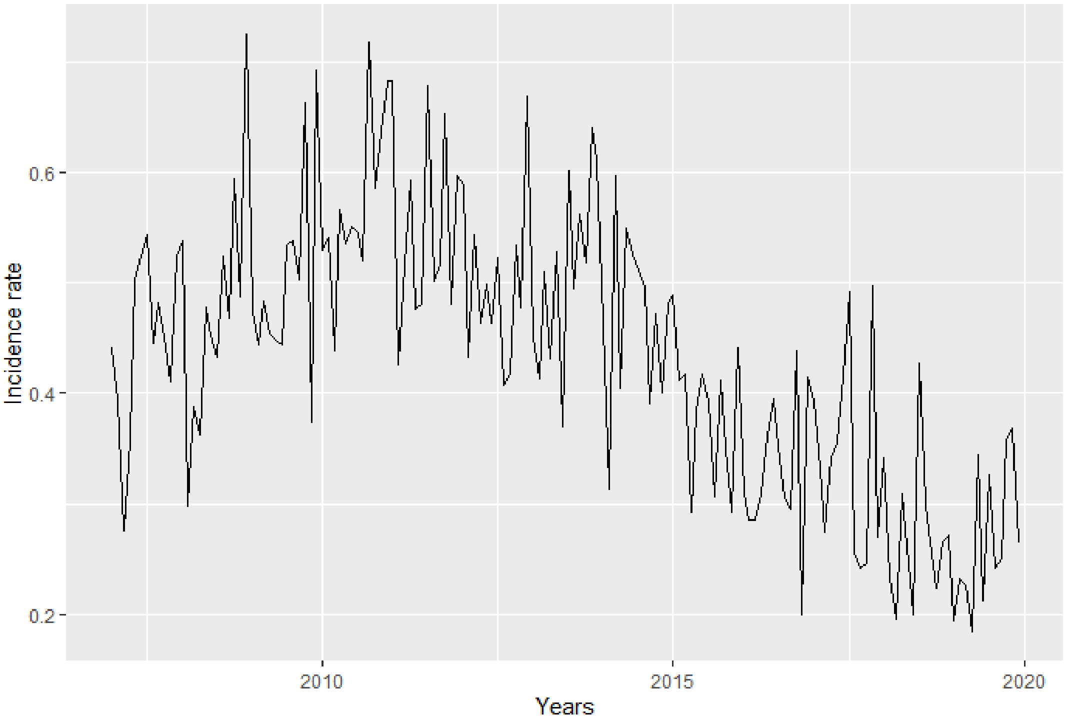

Figure 1: Monthly series of the death incidence of accident in Minas Gerais from January 2007 to December 2019.

However, before testing the trend component's presence in the series, it was necessary to check out the need for data transformation. So, the mean X amplitude graph was constructed, as shown in Fig. 2.

As the slope of the regression between the amplitude and mean of subgroups of the series presented a value statistically different from zero (p-value <0.0001), the components of the series have shown multiplicative and non-additive effects. In this case, it was necessary to make a Box-Cox transformation in the series in order to stabilize its variance, considering $\lambda = 0.6099$.

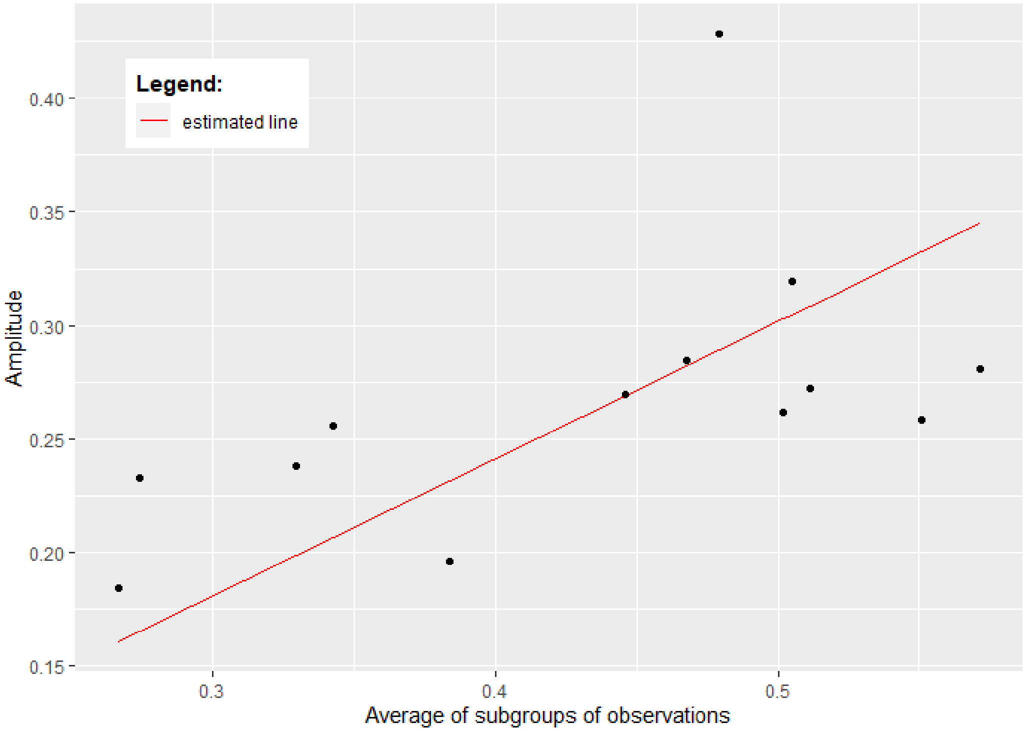

Figure 2: Dispersion graph of the amplitude x average of subgroups of observations in the series (Z_t). The red line represents the estimated regression function for the points.

Cox-Stuart test showed the presence of a trend in our time series (null hypothesis rejected), with 95% confidence, since the p-value of this test was less than 5% (p-value < 0.0001). In addition, the regression fitting with seasonal dummies showed significant coefficients for the months of February, March and April, at the level of 5% of significance. Thus, it was verified the presence of seasonality ($s = 12$ months).

In order to make the series stationary, it was necessary to remove the trend and seasonality through seasonal and non-seasonal differences of order 1 in the time series. Fig. 4 presents the differentiated series of death incidence, as well as their autocorrelation (ACF) and partial autocorrelation (PACF) functions.

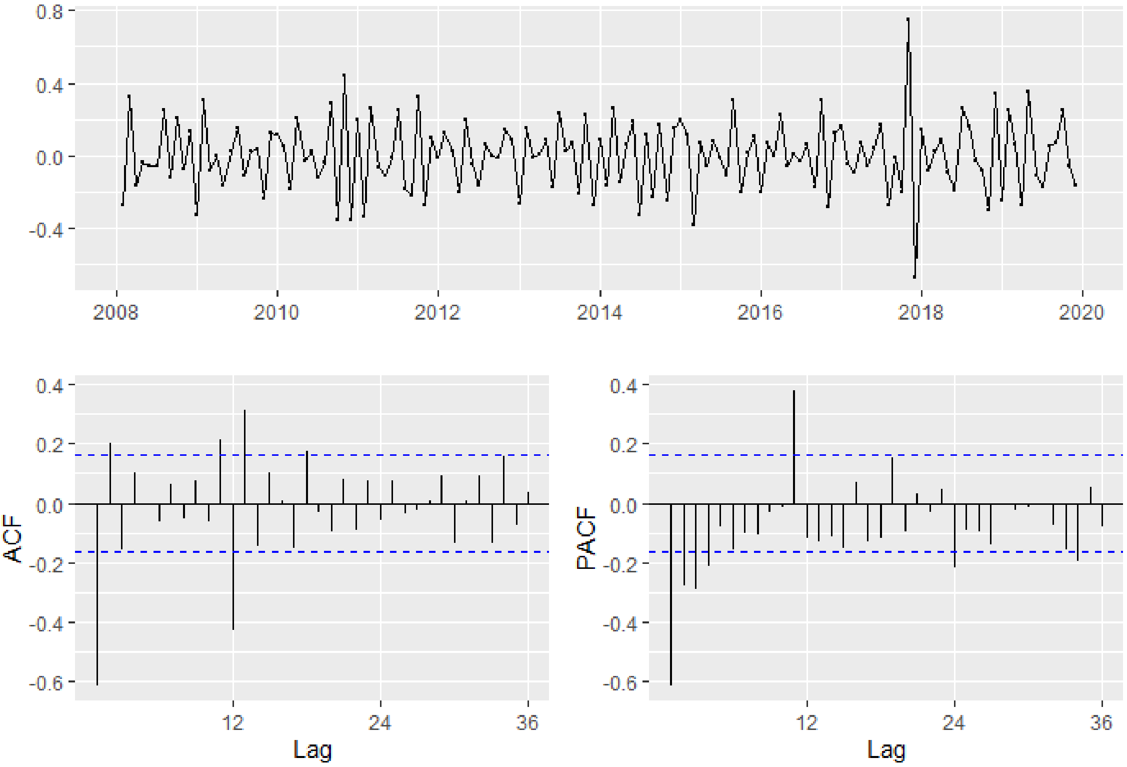

Figure 3: Stationary series of death incidence and their autocorrelation and partial autocorrelation functions.

Based on the ACF and PACF graphs analysis of the stationary series showed in Fig. 3, several possible models were proposed for the fitting of the series of death incidence, looking for the one with greater parsimony and with errors following a white noise.

Based on the diagnostic statistics of the models, SARIMA$(0,1,1)(0,1,1)_{12}$model was the one that presented the best fitting to the data behavior, when we are based on the Box & Pierce test. That is, it was the only model, among those proposed, that presented white noise for the residues.

In order to verify possible points with intervention in the time series, we analyzed the estimated residuals from the chosen model. Those residues that presented, in absolute terms, a value greater than or equal to 2.5 times the standard error, were considered as discrepant. We noted that the observation for November 2017 was discrepant.Therefore,SARIMA$(0,1,1)(0,1,1)_{12}$model was fitted with just one interventionw_1,corresponding to the observationNovember 2017.In this month, the series showed a significant decline in the death incidence by accident. However, no cause was found during this time to justify this behavior.

Table 1 shows theSARIMA $(0,1,1)(0,1,1)_{12}$ model considering the fitting with and without intervention, as well as their respective AIC values, p-values of the Box & Pierce test and the mean squared error (MSE) of forecast.

Table 1: Diagnostic statistics for the proposed models.

Models	AIC	Box & Pierce (p-value)	MSE
SARIMA$(0,1,1)(0,1,1)_{12}$ without intervention	-207.3976	0.8251	0.0043
SARIMA$(0,1,1)(0, 1,1)_{12}$with intervention	-216.0023	0.8224	0.0051

Based on the diagnostic statistics, we can see that the model with the inclusion of an intervention presented a better fit when it is compared by the AIC value. On the other hand, the model without intervention was better in terms of the MSE. However, the difference between the MSE of the two models is small. Therefore, we considered the model with intervention as being the most appropriated for our data.Table2shows the estimated values for the coefficients, as well as their respectivestandard errors.

Table 2: Parameters estimates of $\mathrm{SARIMA}(0,1,1)(0,1,1)_{12}$model with one intervention and their standard errors.

Parameters	Estimates	Standard error
θ_1	-0.8080	0.0462
Θ_1	-0.8158	0.0959
w_1	0.3251	0.0978

Based on the estimates, the fitted model can be written by

$$Z_t = \frac{(1 \; + \; 0.8158 \, B^{12})(1 \; + 0.8080 \, B) a_t}{(1 - B^{12})(1 - B)} + 0.3251 \; \mathbb{I}^{(Nov/2017)} \; \#(8)$$

where $\mathbb{I}$ is an indicatorvariable.

Considering the estimates of this model, whose AIC was -216.002, forecasting were made for the death incidence of traffic accidents in Minas Gerais state from January 2020 to December 2021. The estimates are presented in Table 3:

Table 3: Forecasting of the death incidence of accidents in Minas Gerais from 2020 to 2021 and their respective confidence intervals.

Months	PointForecast	Lo95%	Hi95%
Jan/2020	0.2737695	0.15963783	0.4105348
Feb/2020	0.2194749	0.11475232	0.3488635
Mar/2020	0.2307422	0.12189253	0.3648520
Apr/2020	0.2301372	0.11980953	0.3666114
May/2020	0.2817564	0.15881801	0.4305101
Jun/2020	0.2495409	0.13169710	0.3947701
Jul/2020	0.3288905	0.19324458	0.4910793
Aug/2020	0.2404812	0.12158806	0.3887162
Sep/2020	0.2441725	0.12291628	0.3955306
Oct/2020	0.2797203	0.14888700	0.4405811
Nov/2020	0.2481539	0.12301248	0.4050034
Dec/2020	0.2940358	0.15692869	0.4624852
Jan/2021	0.2556550	0.12182808	0.4251382
Feb/2021	0.2028889	0.08172027	0.3629119
Mar/2021	0.2138213	0.08780069	0.3794832
Apr/2021	0.2132341	0.08578069	0.3816071
May/2021	0.2634333	0.12022739	0.4468210
Jun/2021	0.2320830	0.09586795	0.4108888
Jul/2021	0.3094048	0.15086386	0.5089853
Aug/2021	0.2232789	0.08659446	0.4054176
Sep/2021	0.2268654	0.08757344	0.4126828
Oct/2021	0.2614500	0.11031914	0.4588118
Nov/2021	0.2307347	0.08729104	0.4229904
Dec/2021	0.2753989	0.11707935	0.4817689

From the graph and the forecast table of the series, we can see that the model presented a satisfactory forecast for the series, since the MSE of forecast calculated for the year 2019 was low. In this case, we considered that the forecasts for the months from 2020 to 2021 also will be satisfactory. In addition, the model forecasts that for the next two years the death incidence will slightly stabilize in Minas Gerais state.

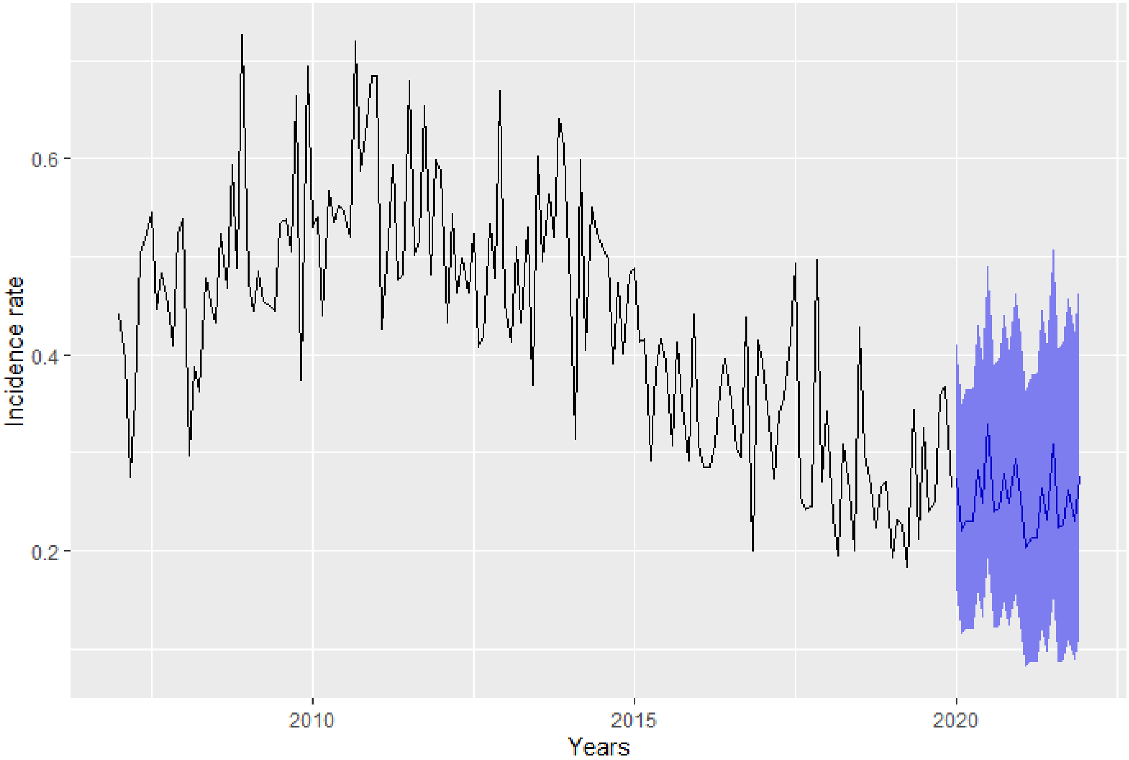

Figure 4: Forecasting graph for death incidence of traffic accidents in Minas Gerais state from January 2020 to December 2021.

IV. CONCLUSION

Time series models were adequate to describe the series of deaths that occurred in traffic accidents on federal highways in the state of Minas Gerais.

In addition, we observed the existence of a trend and seasonality component. There was also anintervention componentassociated to the period of November, 2017.A possible reason for this component in the model would be the presence of some holidays in this month. As these holidays occurred near the weekend, vehicle traffic on the highway tends to grow up, leading to an increased risk of accidents.

Among the suggested models, SARIMA $(0,1,1)(0,1,1)_{12}$ model with just one intervention was considered the most adequate model to describe the time series behavior, as well as to be used for make forecasting about future observations.

The model predicted that the death incidence tends to stabilize in the next two years, from January 2020 to December 2021. These forecasts can be considered as important information for possible actions taken by the State Secretariat for Public Security.

ACKNOWLEDGEMENTS

The authors would like to thank the Coordenação de AperfeiçoamentoPessoal de Nível Superior (CAPES).

REFERENCES

[1] G. E. P. BOX; G. M. JENKINS, Time series analysis: forecasting and control. (San Francisco: Holden-Day, 1976).

[2] Associação Brasileira de Prevenção dos Acidentes de trânsito. 2020. Available in: http://vias-seguras.com.

[3] J. OLIVEIRA, Minas é estado campeão em mortes e acidentes em BRs; veja números. Estado de Minas Gerais, 19 set. 2019.Available in: https://www.em.com.br/app/noticia/gerais/2019/09/19/interna_gerais,1086391/minas-e-estado-campeao-em-mortes-e-acidentes-em-brs-veja-numeros.shtml.

[4] Confederação Nacional do Transporte (CNT). 2020. Available in: https://www.cnt.org.br/.

[5] Departamento de Informática do Sistema Único de Saúde (DATASUS). 2017. Available in: http://www.data-sus.gov.br.

[6] V.M. GUERRERO, Time-series analysis supported by power transformations, *Journal of Forecasting*, *12*, 1993, 37-48.

[7] P. A. MORETTIN; C. M. C. TOLOI, Análise de séries temporais (São Paulo: E. Blucher, 2006).

[8] G. M. LJUNG; G. E. P. BOX, On a measure of lack of fit in time series models, *Biometrika*, *65*, 1978, 297–303.

[9] R Core Team. 2020. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria.

[10] R. HYNDMAN;G. ATHANASOPOULOS; C. BERGMEIR;G.CACERES;L. CHHAY;M. O'HARA-WILD;F. PETROPOULOS;S. RAZBASH;E. WANG; F. YASMEEN. 2020. forecast: Forecasting functions for time series and linear models. R package version 8.12.Available in: http://pkg.robjhyndman.com/forecast.