Modelagem estatística da série mensal de acidentes aéreos: um algoritmo automatizado para a seleção de modelos de previsão do número de ocorrências em curto prazo

Démerson André Polli[1,3], *Nara Núbia Vieira*[2]

1 Universidade de Brasília, Departamento de Estatística, Brasília-DF
2 Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília-DF
3 polli@unb.br

RESUMO: O presente trabalho propõe um algoritmo para encontrar, automaticamente, o modelo que melhor ajusta à série temporal do número de acidentes aéreos mensais no Brasil, com o objetivo de prever a quantidade de acidentes para os meses subsequentes, a fim de identificar um possível crescimento acentuado e de monitorar as atividades preventivas já realizadas. Para tal, utilizou-se a técnica estatística de análise de séries temporais proposta por Box e Jenkins (1970), selecionando o modelo com o menor AIC [Akaike's Information Criteria – Critério de Informação de Akaike], dentre um conjunto de modelos ajustados e testados (o AIC é uma medida proposta por Akaike (1974) que mede a qualidade de ajuste de um modelo estatístico – quanto menor o AIC melhor é o ajuste do modelo estatístico). Foram considerados acidentes aeronáuticos com aeronaves civis não experimentais, que aconteceram entre janeiro de 2000 e dezembro de 2012, dentro dos limites do território brasileiro, segundos dados fornecidos pelo Centro de Investigação e Prevenção de Acidentes Aeronáuticos [Cenipa]. Este trabalho propõe um algoritmo desenvolvido no software livre R (R Development Core Team, 2013) para a seleção automática de modelos de série temporal de Box e Jenkins.

Palavras chave: Acidentes aeronáuticos. Série temporal. Seleção de modelos

Month series statistical modelling of aeronautical accidents: an automated algorithm to select forecast models for short-term events

ABSTRACT: This paper proposes an algorithm to automatically find the model that best fits the time series of the monthly number of air accidents in Brazil, in order to predict the number of accidents in the subsequent months with the objective of identifying any possible surge, as well as monitoring the preventative activities already undertaken. For this purpose, the statistics time series analysis technique proposed by Box and Jenkins (1970) was utilized, with selection of the model with the lowest AIC [Akaike's Information Criteria] from a set of models adjusted and tested (AIC is a measure proposed by Akaike (1974), which measures the quality of the adjustment of a statistical model - the lower the AIC, the better the adjustment of the statistical model). One took into account the aeronautical accidents with non-experimental civil aviation aircraft occurring within the limits of the Brazilian territory between January 2000 and December 2012 (data provided by the Brazilian Aeronautical Accidents Investigation and Prevention Center – CENIPA). This paper proposes an algorithm developed in the free R software (R Development Core Team, 2013) for automatic selection of Box and Jenkins' time-series models.

Key words: Aeronautical accidents. Time series. Selection of models

Citação: Polli, DA, Vieira, NN. (2015) Modelagem estatística da série mensal de acidentes aéreos – um algoritmo automatizado para seleção de modelos para previsão em curto prazo do número de ocorrências. *Revista Conexão Sipaer*, Vol. 6, No. 1, pp. 551-558.

Recebido 18 outubro 2014; **Aceito** 27 janeiro 2015; **Publicado** 30 abril 2015

1 INTRODUÇÃO

Evitar o aumento no número de acidentes aeronáuticos é interesse de toda a sociedade, em especial, diante do crescimento do setor aéreo. Cada vez mais se torna necessário aperfeiçoar ações preventivas, tais como palestras, cursos, seminários, etc., que possuem o objetivo "de evitar perdas de vidas e de material decorrentes de acidentes aeronáuticos" (Brasil, 1982).

Uma forma de monitorar o efeito das atividades de prevenção é avaliar a evolução da quantidade de ocorrências de acidentes aéreos mensais, o que pode ser feito por meio do estudo de séries temporais. Este estudo se baseia na análise do conjunto de observações de uma variável ao longo do tempo, sendo que dados futuros dependem das informações do passado. Como existe dependência temporal entre as observações, esse trabalho tem o objetivo de encontrar o melhor modelo de série temporal que se ajusta ao número de acidentes aéreos e, com ele, gerar previsões dos valores futuros para os próximos meses da série. Ao extrapolar essa série temporal, é possível comparar os valores esperados com os reais e, assim, verificar se as ações preventivas feitas em um período surtiram os efeitos positivos nos meses subsequentes.

Em Estatística, uma sequência de variáveis aleatórias que evoluem no tempo, e apresentam dependência entre as observações, é denominada uma série temporal (Harvey, 1993; Morettin e Toloi, 2004). Como exemplos de séries temporais pode-se citar a evolução de um valor ativo ou de commodities no tempo, a variação de temperatura ou do

índice pluviométrico em uma determinada localidade, e, em particular, o número de acidentes aéreos ocorridos em um período determinado de tempo. Os modelos usuais para análise de séries temporais são aqueles da família de modelos de Box e Jenkins (1970): o autorregressivo de média móvel, em inglês, Auto Regressive Moving Average [ARMA]; o autorregressivo integrado de média móvel, em inglês, Auto Regressive Integrated Moving Average [ARIMA] e o sazonal autorregressivo integrado de média móvel, em inglês, Seasonal Auto Regressive Integrated Moving Average [SARIMA].

O modelo ARMA é usado em séries sem sazonalidade e sem tendências. Enquanto isso, o ARIMA é utilizado apenas em séries com tendências e o SARIMA, em séries com sazonalidade e tendência. Estes modelos permitem o ajuste e a predição da série temporal (Harvey, 1993; Morettin e Toloi, 2004).

2 METODOLOGIA

Uma série temporal é uma sequência $\{X_1, X_2, \dots\}$ de valores observados nos instantes $t \in \mathbb{Z}^+$ (índices inteiros positivos), de modo que exista dependência temporal entre valores observados em instantes distintos (Harvey, 1993; Morettin e Toloi, 2004). Desta forma, X_1 representa o primeiro valor observado, X_2 representa o segundo valor observado, e assim por diante. As séries temporais podem ser modeladas através de uma família de modelos chamadas de modelos de Box e Jenkins (1970) – estes modelos são definidos com base em um operador chamado *backshift* (B).

O operador *backshift* (*B)* é normalmente usado em modelos ARIMA, nos índices da série temporal, para deslocar para trás uma unidade de tempo, de modo a formar uma nova série. (Cryer e Chan, 2008). Representando por $\{X_1, X_2, \dots\}$ a sequência de valores observados nos instantes $t > 0, t \in \mathbb{Z}$ (a série temporal $\{X_t : t = 1,2, \dots\}$) é possível definir o *operador backshift* de ordem *k* por $B^k X_t = X_{t-k}$, e o operador diferença de ordem por $\nabla^d X_t = (1-B)^d X_t$, em que o operador $(1-B)^d$ é desenvolvido como se fosse um polinômio de ordem d e, posteriormente, os respectivos operadores *backshift* são aplicados à série. A Tabela 1 mostra a definição dos modelos de Box e Jenkins usando esta notação para os modelos; em que o polinômio $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ define a parte autorregressiva, $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ define a parte de médias móveis, $\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$ define a parte autorregressiva sazonal e $\Theta(B^s) = 1 + \Theta_1 B^2 + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}$ define a parte de médias móveis sazonais, os parâmetros do modelo são $\phi_1, \phi_2, \dots, \phi_p$ na parte autorregressiva, $\theta_1, \theta_2, \dots, \theta_q$ na parte de médias móveis, $\Phi_1, \Phi_2, \dots, \Phi_P$ na parte autorregressiva sazonal e $\Theta_1, \Theta_2, \dots, \Theta_Q$ na parte de médias móveis sazonais e a ordem do modelo (quantidade de termos no polinômio) são p para a parte autorregressiva, q para a parte de médias móveis, P para a parte autorregressiva sazonal e Q para a parte de médias móveis sazonais. O valor s representa o período no qual ocorrem as sazonalidades, ou seja, qual o intervalo de tempo que a série demora em apresentar comportamentos semelhantes – de aumento ou decréscimo. O componente ω_t do modelo é aleatório, a saber, é gerado por meio de uma distribuição de probabilidades – em geral uma 'normal de média zero', caso no qual este componente é chamado de "ruído branco". O componente X_t do modelo é determinístico e trata dos valores observados na série.

Tabela 1: Definição dos modelos de Box e Jenkins

Modelo	Ordem	Definição
AR	p	$\phi(B)X_t = \omega_t$
MA	q	$X_t = \theta(B)\omega_t$
ARMA	(p,q)	$\phi(B)X_t = \theta(B)\omega_t$
ARIMA	(p,d,q)	$\phi(B)\nabla^d X_t = \theta(B)\omega_t$
SARIMA	(p,d,q) $\times (P,D,Q)_s$	$\Phi(B^s)\phi(B)\nabla_s^D \nabla^d X_t$ $= \Theta(B^s)\theta(B)\omega_t$

A identificação do modelo a ser usado para ajustar uma série temporal exige do analista um grande conhecimento estatístico; isto inibe a aplicação do método em diversas situações. O modelo pode ser identificado através da observação da série temporal (com o objetivo de identificar tendências e/ou sazonalidades) e das funções de autocorrelação e de autocorrelação parcial (com o objetivo de identificar o modelo mais adequado e a ordem de tal modelo). Por se tratar de uma análise gráfica, este método traz alguma subjetividade na escolha do modelo – em geral tal subjetividade não é bem-vinda em análises estatísticas (Morettin e Toloi, 2004). A Tabela 2 mostra como se interpreta as figuras das funções de autocorrelação [FAC] e de autocorrelação parcial [FACP]; as duas funções precisam ser avaliadas em conjunto para escolher o modelo. Além disso, somente é aplicável calcular as funções FAC e FACP em séries sem tendência e nem sazonalidade; caso a série apresente tais características, é necessário eliminá-las por meio da aplicação dos operadores diferença $\nabla^d(\cdot)$ e diferença sazonal $\nabla_s^D(\cdot)$.

Tabela 2: Interpretação das funções de autocorrelação e de autocorrelação parcial

Modelo	FAC	FACP
AR	Zero para lags maiores de q.	Decaimento exponencial
MA	Decaimento exponencial	Zero para lags maiores que p.
ARMA	Decaimento exponencial	Decaimento exponencial

Cleveland et al. (1990) propõem um método de decomposição da sazonalidade de uma série temporal através de regressões locais do tipo LOWESS (Cleveland, 1979). Tal método permite investigar se na série original existem tendências e, principalmente, sazonalidades e qual o período em que tal sazonalidade ocorre. Isto permite identificar tais componentes na série original dos dados e efetuar a

eliminação dos mesmos, para que seja possível identificar o modelo gerador por meio das funções de autocorrelação.

A seleção do melhor modelo para ajuste da série temporal e a respectiva ordem requer um grande conhecimento da teoria de séries temporais. Os parágrafos acima apresentam um pequeno resumo desta teoria. Este trabalho propõe um algoritmo desenvolvido no software livre R (R Development Core Team, 2013) para a seleção automática de modelos de série temporal de Box e Jenkins. Tal seleção se baseia em selecionar o modelo que apresenta o menor AIC [*Akaike's Information Criteria* – Critério de Informação de Akaike], dentre um conjunto de modelos ajustados e testados. O AIC é uma medida proposta por Akaike (1974) e que mede a qualidade de um modelo estatístico – quanto menor o AIC melhor é o modelo.

2.1 ALGORITMO PARA SELEÇÃO DO MODELO ESTATÍSTICO

As Figuras 1 e 2 mostram o código escrito em R para selecionar os modelos ARIMA (figura 1) e SARIMA (figura 2) a partir de um conjunto de 75 modelos ARIMA ($ARIMA(0,0,0)$ ao $ARIMA(4,2,4)$) e 1200 modelos do SARIMA ($SARIMA(0,0,0) \times (1,1,1)_6$ ao $SARIMA(4,2,4) \times (4,1,4)_6$). Os códigos são bastante parecidos, a diferença é apenas a criação da matriz de ordens dos modelos e a definição do modelo. A sazonalidade está fixada em 6 meses (semestralidade).

```
td = data.frame(ar = replicate(75, NA),
                delta = replicate(75, NA),
                ma = replicate(75, NA),
                aic = replicate(75, NA))
td$ar    = rep(0:4, each = 15)
td$delta = rep(0:2, each = 5)
td$ma    = 0:4
ARIMA.step = function(td, data) {
 model.fit = NA; aic = NA
 model.fit = try(expr = ARIMA(data, order = c(td["ar"], td["delta"], td["ma"])), silent = TRUE)
 if(class(model.fit) == "ARIMA") {if((!any(is.na(model.fit$coef))) & (!any(is.na(model.fit$var.coef))))
    aic = with(model.fit, aic)}
 aic}
td$aic = apply(X = td, MARGIN = 1, FUN = ARIMA.step, data = cnt)
td = na.omit(td)
tdo = td[with(td, order(aic)),]
head(tdo); tail(tdo)
```

Figura 1: Código em R para a seleção do modelo ARIMA

Ambos os códigos testam os modelos e exibem uma seleção dos melhores (os menores valores de AIC) e piores (maiores valores de AIC) modelos analisados. Após a execução destes códigos, é possível fazer previsões usando a(s) série(s) selecionada(s). Observe que os dados (contagem mensal de acidentes aéreos) estão no vetor numérico cnt.

2.2 PREVISÃO DA CONTAGEM MENSAL DE ACIDENTES

As Figuras 3 e 4 apresentam, respectivamente, os códigos usados para a previsão do número de acidentes a partir dos modelos ARIMA e SARIMA, selecionados pelo algoritmo de seleção de modelos apresentado na seção anterior.

3 DISCUSSÃO / ANÁLISE DOS DADOS

Como aplicação para os algoritmos apresentados neste trabalho, foi analisada a série dos acidentes aéreos ocorridos com aeronaves civis não experimentais, investigados pelo Centro de Investigação e Prevenção de Acidentes Aeronáuticos [Cenipa], no período entre janeiro de 2000 e dezembro de 2012. As seções 3.1, 3.2 e 3.3 a seguir apresentam, respectivamente, o ajuste do modelo; a avaliação do ajuste e as previsões para a contagem de acidentes aéreos entre janeiro de 2013 a junho de 2013 feitas pelos modelos ARIMA e SARIMA.

3.1 AJUSTE DOS MODELOS ARIMA E SARIMA

A Figura 5 apresenta a decomposição da série de acidentes aéreos nos componentes sazonalidade (*seasonal*), tendência (*trend*) e resíduo (*remainder*), de acordo com o método proposto por Cleveland (1990). Observe que a série de sazonalidade apresenta picos aproximadamente a cada 12 meses, indicando haver sazonalidade anual. No entanto, em um período de um ano se observa dois padrões distintos, um menor e outro maior, ambos em formato de 'M'. Isto indica que pode haver também sazonalidade semestral. A modelagem da sazonalidade semestral também ajusta a anual, pois estes períodos são múltiplos.

```
td = data.frame(ar = replicate(1200, NA),
                delta = replicate(1200, NA),
                ma = replicate(1200, NA),
                aic = replicate(1200, NA),
                sar = replicate(1200, NA),
                sma = replicate(1200, NA))
td$ar    = rep(0:4, each = 240)
td$delta = rep(0:2, each = 80)
td$ma    = rep(0:4, each = 16)
td$sar   = rep(1:4, each = 4)
td$sma   = 1:4
SARIMA.step = function(td, data, t = 6) {
 model.fit = NA; aic = NA
 model.fit = try(expr = ARIMA(data, order = c(td["ar"], td["delta"], td["ma"]), seasonal = list(order = c(td["sar"], 1, td["sma"]),
period = t)), silent = TRUE)
 if(class(model.fit) == "ARIMA") {
  if((!any(is.na(model.fit$coef))) & (!any(is.na(model.fit$var.coef))))
   aic = with(model.fit, aic)} aic}
td$aic = apply(X = td, MARGIN = 1, FUN = SARIMA.step, data = cnt)
td = na.omit(td)
tdo = td[with(td, order(aic)),]
head(tdo); tail(tdo)
```

Figura 2: Código em R para a seleção do modelo SARIMA

```
# O modelo com o menor AIC

(aic.min = td[with(td, which.min(aic)),])

(tss.min = ARIMA(d, order = with(aic.min, c(ar, delta, ma)),
seasonal = list(order = with(aic.min, c(sar, 1, sma)), period
= 6)))

tsdiag(tss.min); predict(tss.min, n.ahead = 6)
```

Figura 3: Código em R para a previsão de seis meses do modelo ARIMA

```
# O modelo com o menor AIC

(aic.min = td[with(td, which.min(aic)),])

(ts.min = ARIMA(d, order = with(aic.min, c(ar, delta, ma))))

tsdiag(ts.min); predict(ts.min, n.ahead = 6)
```

Figura 4: Código em R para a previsão de seis meses do modelo SARIMA

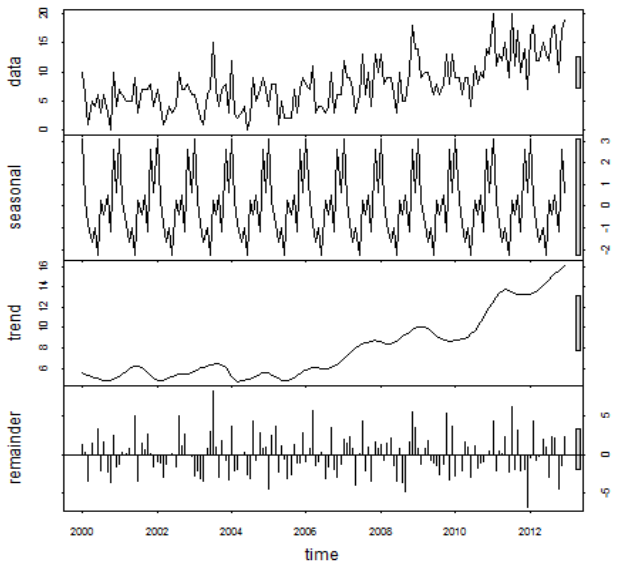

Figura 5: Decomposição da série original em sazonalidade e tendência

Na Figura 6 se observa que a série apresenta tendência de crescimento e picos em intervalos regulares, mostrando que também há sazonalidade.

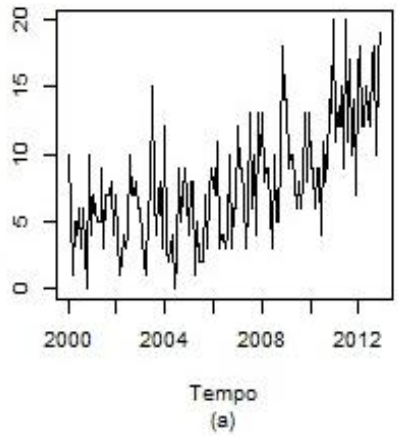

Figura 6: Série original, série livre de tendência e sazonalidade e autocorrelações

A Figura 7 é a série diferenciada pelo operador $\nabla_6^1 \nabla^1 X_t$ que gera a série livre de tendências e sazonalidades; observe que esta série apresenta um aumento na variabilidade posterior a 2010, indicando que o processo gerador da série pode ter mudado após este período.

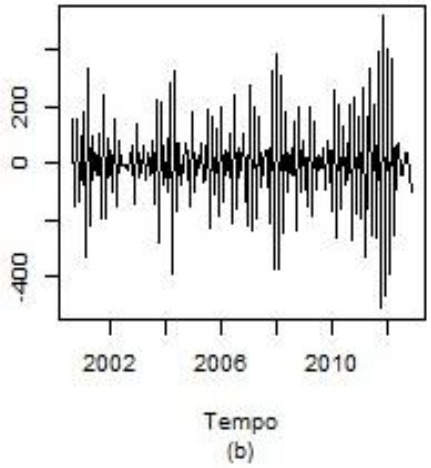

Figura 7: Série original, série livre de tendência e sazonalidade e autocorrelações

A Figura 8 mostra a função de autocorrelação para a série livre de tendência e sazonalidade, apresentando decaimento exponencial indicando (Tabela 2) que esta série modificada deve seguir ou o modelo MA ou o modelo ARMA.

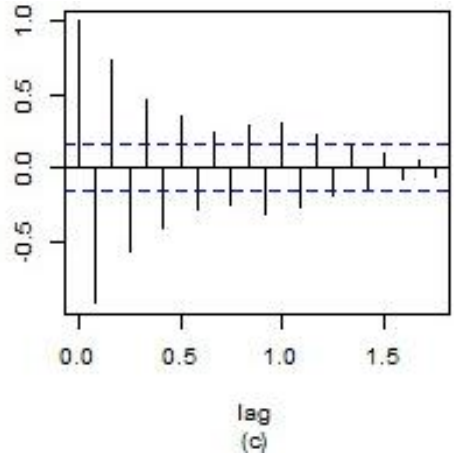

Figura 8: Série original, série livre de tendência e sazonalidade e autocorrelações

A Figura 9 apresenta valores da função de autocorrelação parcial para a série livre de tendência e sazonalidade; observe que a função decai e os valores após o 4º *lag* (barras verticais) estão entre as linhas pontilhadas indicando que tais valores podem ser considerados nulos – neste caso, há indicação que a série livre de tendência e sazonalidade deve ser modelada pela série $MA(4)$. Desta análise gráfica, surge, então, a sugestão de que a série original seja modelada por um processo SARIMA, cujo componente MA tem ordem 4.

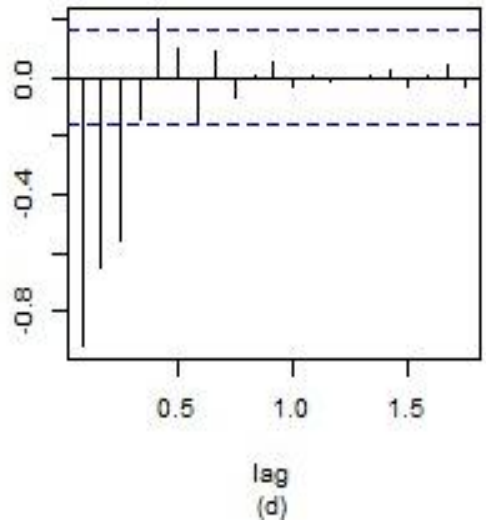

Figura 9: Série original, série livre de tendência e sazonalidade e autocorrelações

O ajuste dos modelos à série dos acidentes aéreos foram obtidos executando os algoritmos dos Quadros 1 e 2. Para o modelo ARIMA foram testados 75 modelos do $ARIMA(0,0,0)$ ao $ARIMA(4,2,4)$ em 3,79 segundos; e para o modelo SARIMA foram testados 1200 modelos do $SARIMA(0,0,0) \times (1,1,1)_6$ ao $SARIMA(4,2,4) \times (4,1,4)_6$ em 59 minutos e 57,17 segundos. O algoritmo foi executado no R versão 3.0.1 para Windows em um HP Compaq com processador AMD Phenom II X4 B97 3.2 GHz e 4GB de memória.

O modelo ARIMA selecionado como o menor AIC (*Akaike's Information Criteria* – Critério de Informação de Akaike) foi o $ARIMA(4,2,4)$ com $AIC = 810{,}8986$ e o modelo SARIMA selecionado como o menor AIC foi o $SARIMA(4,1,4) \times (2,1,3)_6$ com $AIC = 779{,}0282$. Observe que o modelo sazonal apresentou um AIC menor que o modelo apenas com tendência, isto indica que considerar sazonalidade no modelo melhora o ajuste à série em estudo. A seguir são mostrados os ajustes para ambos os modelos.

Tabela 3: Estimativas para os modelos ARIMA e SARIMA selecionados

Coeficiente	Estimativa (Erro Padrão)	
	ARIMA	SARIMA
AR1 (ϕ_1)	-0,9623 (0,2257)	0,9916 (0,0886)
AR2 (ϕ_2)	-0,4629 (0,1739)	-0,7533 (0,0835)
AR3 (ϕ_3)	0,3121 (0,1129)	1,0629 (0,0856)
AR4 (ϕ_4)	0,3115 (0,1039)	-0,3016 (0,0872)
MA1 (θ_1)	-0,8633 (0,2470)	-1,8504 (0,0602)
MA2 (θ_2)	-0,4382 (0,2929)	1,6321 (0,1056)
MA3 (θ_3)	-0,5165 (0,1896)	-1,7029 (0,0966)
MA4 (θ_4)	-0,8263 (0,2094)	0,9258 (0,0546)
SAR1 (Φ_1)	–	-0,9974 (0,1590)
SAR2 (Φ_2)	–	-0,5739 (0,2075)
SMA1 (Θ_1)	–	0,0422 (0,1533)
SMA2 (Θ_2)	–	-0,1182 (0,1018)
SMA3 (Θ_3)	–	-0,8807 (0,2093)

3.2 AVALIAÇÃO DA QUALIDADE DO MODELO

A avaliação da qualidade do ajuste do modelo pode ser feito pelas Figuras 10 e 11. Nestes, estão apresentados a série dos resíduos do modelo, a função de autocorrelação dos resíduos e o teste de Ljung-Box (1978), para a autocorrelação dos resíduos. Um modelo está bem ajustado quando a série dos resíduos não apresenta indícios de tendência e nem sazonalidade. Isto é verificado pela função de autocorrelação dos resíduos que deve ter todos os lags (pontos) entre as linhas pontilhadas (exceto o primeiro que sempre é igual a 1); e os p-valores do teste de Ljung-Box devem todos estar acima da linha pontilhada indicando que não existem autocorrelações entre observações do resíduo.

O Figura 10 apresenta a análise de qualidade de ajuste para o modelo ARIMA selecionado. Observe que não há indícios de tendência e nem sazonalidade na série dos resíduos, a função de autocorrelação é considerada nula (valor entre as linhas pontilhadas) para todos os *lags* exceto o *lag* equivalente à 12ª haste – isto ocorre, como foi visto, pois os dados apresenta sazonalidade, e o modelo ajustado [ARIMA] não captura tal sazonalidade. Os p-valores dos testes de *Ljung-Box* mostram que não ocorrem autocorrelações significantes.

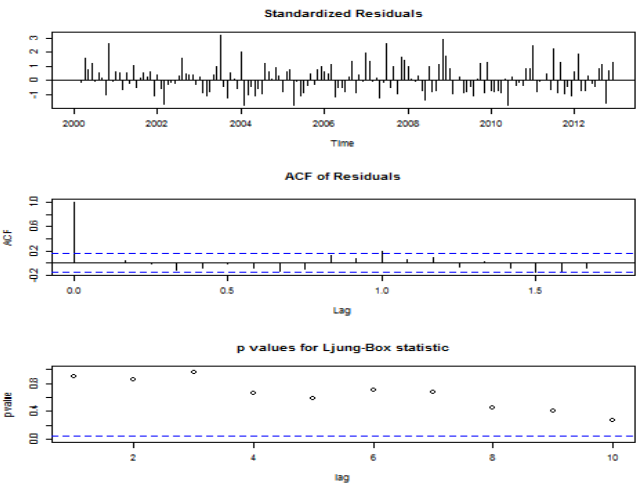

Figura 10: Diagnóstico do modelo ARIMA com o menor AIC

O Figura 8 apresenta a análise de qualidade de ajuste para o modelo SARIMA selecionado. Observe que, da mesma forma que no caso anterior, não há indícios de tendência e nem sazonalidade na série dos resíduos, todos os lags da função de autocorrelação são considerados nulos (entre as linhas pontilhadas) e todos os p-valores dos testes de Ljung-Box mostram que não ocorrem autocorrelações. No entanto, observe que os p-valores do teste de Ljung-Box são maiores que aqueles apresentados para o modelo ARIMA, mostrando um melhor ajuste do SARIMA se comparado com o ARIMA.

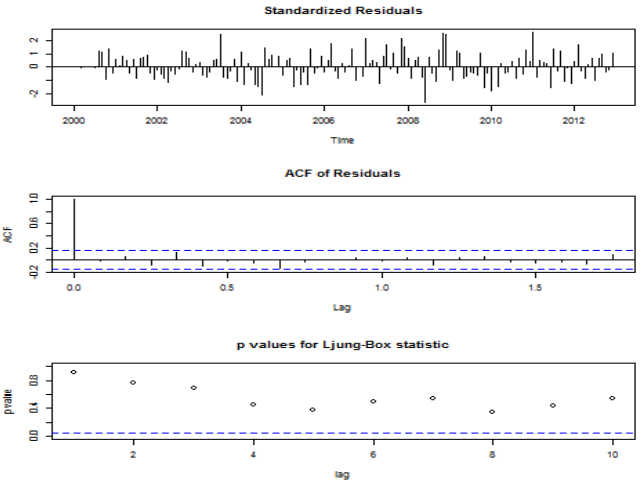

Figura 11: Diagnóstico do modelo SARIMA com o menor AIC

O diagnóstico do modelo é feito pelo software R com o comando ts.diag(ts.min).

3.3 PREVISÃO DO NÚMERO DE ACIDENTES ENTRE JAN/2013 A JUN/2013

A previsão dos modelos de Box e Jenkins (1974) no software R é feita pelo comando *predict(ts.min, n.ahead = 12)*. A Tabela 4 mostra a previsão do número de acidentes

aéreos entre janeiro a junho de 2013. Os valores são apresentados com duas casas decimais, apesar de se tratar da contagem de eventos – isto porque o valor apresentado é exatamente o previsto pelo modelo. Para efeitos práticos, pode se considerar a parte inteira caso a fração seja inferior a 0,50 ou a parte inteira acrescida de 1, caso a fração seja superior ou igual a 0,50. Os valores entre parênteses mostram os erros-padrão correspondentes; observe que, à medida que se avança no tempo, o erro padrão da previsão aumenta – isto é esperado, pois não se pode prever com segurança um futuro muito distante.

Tabela 4: Previsões do número de acidentes aéreos entre jan/2013 a jun/2013

Mês	Estimativa (Err. Padrão)	
	SARIMA	ARIMA
Jan/2013	16,05 (2,772)	15,87 (3,059)
Fev /2013	16,52 (2,803)	15,05 (3,113)
Mar /2013	17,81 (2,846)	17,19 (3,184)
Abr/2013	14,61 (2,907)	15,47 (3,192)
Mai/2013	14,80 (2,911)	15,21 (3,293)
Jun/2013	16,78 (2,918)	16,97 (3,304)
TOTAL	96,57	95,76

Um resultado interessante que se pode obter dos modelos ARIMA e SARIMA é um intervalo aproximado de confiança para a quantidade de acidentes. Isto é obtido, da tabela acima, somando e subtraindo da estimativa da quantidade mensal de acidentes duas vezes o valor do erro padrão. Desta forma, para efeitos práticos, pode-se considerar que pelo modelo ARIMA a previsão dos acidentes é: 16 (esperado entre 10 e 22) em janeiro; 17 (esperado entre 9 e 21) em fevereiro; 18 (esperado entre 11 e 24) em março; 15 (esperado entre 9 e 22) em abril; 15 (esperado entre 9 e 22) em maio e 17 (esperado entre 10 e 24) em junho de 2013 – totalizando 98 acidentes no período (58 no melhor cenário e 135 no pior cenário). De modo análogo para o modelo SARIMA a previsão da contagem de acidentes é 16 (esperado entre 11 e 22) em janeiro; 15 (esperado entre 11 e 22) em fevereiro; 17 (esperado entre 12 e 23) em março; 15 (esperado entre 9 e 20) em abril; 15 (esperado entre 9 e 21) em maio e 17 (esperado entre 11 e 23) em junho de 2013 – totalizando 97 acidentes no período (63 no melhor cenário e 131 no pior cenário).

Para avaliação da qualidade do modelo, o período de previsão (janeiro a junho de 2013) é comparado na Tabela 5 com os valores tabulados pelo Cenipa (2014). Observa-se que ambos os modelos superestimaram o total de acidentes no período entre janeiro e junho de 2013 – o modelo ARIMA superestimou a quantidade de acidentes em 13,9% enquanto que o modelo SARIMA superestimou a quantidade de acidentes em 12,8%. Ao observar as estimativas mensais, ambos os modelos apresentaram uma estimativa abaixo do observado em janeiro de 2013, uma estimativa idêntica ao observado em maio de 2013 e estimativas acima do observado nos demais meses. Considerando os desvios mensais, a maior diferença entre o observado e o esperado ocorreu para ambos os modelos no mês de junho, com estimativa 41,7% maior que o observado. Por se tratar de um modelo estatístico, é natural que os valores preditos não sejam idênticos aos valores observados – além disso, em cada um dos meses, a quantidade observada de acidentes estava dentro dos intervalos estimados (mínimo esperado – máximo esperado).

Tabela 5: Previsões do número de acidentes aéreos entre jan/2013 a jun/2013

Mês	Estimativa		Publicado no site do Cenipa
	SARIMA	ARIMA	
Jan/2013	16 (11 – 22)	16 (10 – 22)	18
Fev /2013	15 (11 – 22)	17 (09 – 21)	14
Mar /2013	17 (12 – 23)	18 (11 – 24)	13
Abr/2013	15 (09 – 20)	15 (09 – 22)	14
Mai/2013	15 (09 – 21)	15 (09 – 22)	15
Jun/2013	17 (11 – 23)	17 (10 – 24)	12
TOTAL	97 (63–131)	98 (58–135)	86

Com o passar do tempo, a série naturalmente é acrescida das novas observações. Neste caso, é necessário que o modelo seja ajustado novamente à nova série dos dados.

3.4 CONSIDERAÇÕES PARA USO FUTURO DO ALGORITMO PELO CENIPA

Este artigo apresenta um código desenvolvido em R para a previsão da quantidade acidentes aéreos no Brasil, mensalmente, baseado na série histórica apurada pelo Cenipa a partir de janeiro de 2000. As previsões podem (e devem) ser atualizadas periodicamente e, para isto, é suficiente acrescentar os novos dados no script e executá-lo no software livre R (disponível em www.r-project.org). O código faz o ajuste e a escolha dos modelos ARIMA e SARIMA, através do critério de Akaike, que apresenta a previsão 6 meses adiante, além de criar arquivos no formato PNG com as figuras apresentadas neste trabalho.

4 CONCLUSÃO

A previsão da contagem do número de acidentes aéreos pode ser uma informação estratégica, pois, além de permitir a identificação de um possível aumento acentuado em um determinado período, pode ser um importante instrumento para monitorar a qualidade das ações de prevenção, já que, depois de passado o período de extrapolação da série, é possível verificar se houve uma diferença significativa entre o valor esperado e o observado. Caso a quantidade de acidentes tenha crescido, é um alerta de que essas ações foram insuficientes ou pouco efetivas. Os modelos de séries temporais, devido à sua complexidade, em geral, exige o apoio de um profissional especializado em tais modelos. Este trabalho, no entanto, apresenta uma proposta de algoritmo computacional que pode facilmente ser implementado no software R – um software livre (e, portanto, gratuito, disponível em *www.r-project.org*) – para a seleção de modelos de Box e Jenkins (1974) para modelagem da série de acidentes aéreos, sem que seja necessário um conhecimento aprofundado destes modelos. Em particular, para os dados observados entre janeiro de 2000 e dezembro de 2012 os modelos selecionados pelo critério de Akaike (1974) foram, respectivamente, para os modelos com tendência e com tendência e sazonalidade, o $ARIMA(4,2,4)$ com $AIC = 810{,}8986$ e o $SARIMA(4,1,4) \times (2,1,3)_6$ com $AIC = 779{,}0282$. Ambos os modelos ficaram bem ajustados, com melhora da qualidade do ajuste ao se considerar o modelo sazonal. Com a evolução da série temporal, é interessante que o modelo seja atualizado novamente, neste caso pode ocorrer do modelo selecionado não coincidir com os mostrados neste trabalho, pois a série futura traz informação nova ao ajuste do modelo, o que pode permitir ao software selecionar modelos ainda melhores.

A seleção do modelo ARIMA considera 75 propostas de modelos do $ARIMA(0,0,0)$ ao $ARIMA(4,2,4)$ e a busca ocorre aproximadamente em 5 segundos; e a seleção do modelo SARIMA considera 1200 modelos do $SARIMA(0,0,0) \times (1,1,1)_6$ ao $SARIMA(4,2,4) \times (4,1,4)_6$ com uma busca de cerca de 1 hora. O algoritmo foi executado no R versão 3.0.1 para Windows em um HP Compaq com processador AMD Phenom II X4 B97 3.2 GHz e 4GB de memória.

Aqueles interessados em obter o código original do algoritmo em R podem contatar o autor principal.

REFERENCIAS BIBLIOGRÁFICAS

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp. 716 – 723.

Box, GEP; JENKINS, G. (1970) *Time series analysis: Forecasting and control.* São Francisco: Holden-Day.

Ministério da Defesa, Comando da Aeronáutica, *Decreto nº 87.249, de 07 de junho de 1982*. Disponível em: http://www.cenipa.aer.mil.br/cenipa/index.php/legislacao/category/5-outros [06 jan 2013].

------. Centro de Investigação e Prevenção de Acidentes Aeronáuticos (Cenipa) (2013). *Acidentes Civis 2013*, Aviação Civil Brasileira, pp. 10, Disponível em: http://www.cenipa.aer.mil.br/cenipa/Anexos/article/18/Acidentes_Civis_2013.pdf [28 Jan 2014].

Cleveland, RB; Clevelend, WS; McRae, JE; Terpenning, I. (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess, *Journal of Official Statistics*, Vol. 6, No. 1, pp. 3 – 73.

Cleveland, WS. (1979) Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistician*, Vol. 74, No. 368, pp. 829 – 836.

Cryer, JD; Chan, KS. (2008) *Time series analysis with applications in R.* 2. Ed, Nova Iorque: Springer.

Harvey, AC. (1993) *Time series models*, 2. ed. Harvester Wheatsheaf.

Ljung, GM; Box, GEP. (1978) On a measure of a lack of fit in time series models, *Biometrika*, Vol. 65, No. 2. pp. 297 – 303.

Morettin, PA; Toloi, CMC. (2004) *Análise de séries temporais*, São Paulo: Blucher.

R Development Team. (2013) *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, 2013. Disponível em: http://www.R-project.org/ 16 de maio de 2013.