ESTIMAÇÃO CLÁSSICA E BAYESIANA DE MODELOS INAR(1) EM NÚMERO DE DIAS COM PRECIPITAÇÃO NO MUNICÍPIO DE GARANHUNS-PE

Dâmocles Aurélio Nascimento da SILVA[1]
Moacyr CUNHA FILHO[2]
Ana Patrícia Siqueira Tavares FALCÃO[3]
Gabriela Isabel Limoeiro ALVES[2]

▪ RESUMO: O ciclo meteorológico pode ser, muitas vezes, descrito por dados de séries temporais. Os meteorologistas, em geral, usam modelos contínuos. O interesse foi analisar dados meteorológicos discretos com o modelo INAR(1), através de abordagem clássica e bayesiana na estimação dos parâmetros. Sendo assim, neste trabalho é descrito uma sequência de procedimentos para estimar parâmetros de modelos autoregressivos de ordem $p = 1$, para valores inteiros INAR(1), por meio de inferência clássica via estimador de máxima verossimilhança e inferência bayesiana via simulação de Monte Carlo em Cadeias de Markov (MCMC). Duas alternativas são consideradas para a densidade a priori dos parâmetros do modelo. Para o primeiro caso, adota-se uma densidade a priori não-informativa. Para o segundo, adota-se uma densidade conjugada beta-gama. A análise a posteriori é efetuada por meio de algoritmos de simulação MCMC. Avalia-se também a previsão de novos valores da série número de dias com precipitação. O período de análise compreendeu 30/11/1993 à 29/02/2012 e obteve previsões do período de 31/03/2012 à 28/02/2013. Foram utilizados um modelo INAR(1) de estimação clássica dos parâmetros e dois modelos INAR(1) de estimação bayesiana para os parâmetros. Na escolha do modelo mais adequado foi utilizado o critério de informação de Akaike (AIC). A análise dos erros de previsão foi um instrumento utilizado para verificar qual modelo se adequou melhor aos dados. Conclui-se que o uso de simulação MCMC torna o processo de inferência bayesiana mais flexível, podendo ser estendido para problemas

[1]Universidade de Pernambuco – UPE, CEP: 55294-902, Garanhuns, PE, Brasil. E-mail: *damocles_aurelio@hotmail.com*
[2]Universidade Federal Rural de Pernambuco – UFRPE, Departamento de Informática – DEINFO, CEP: 52171-900, Recife, PE, Brasil. E-mail: *moacyr2006@ibest.com.br*; *gabbybel@hotmail.com*
[3]Instituto Federal de Pernambuco – IFPE, CEP: 55600-000, Vitória de Sto Antão, PE, Brasil. E-mail: *apstfalcao@hotmail.com*

de dimensão maior. Os modelos bayesianos apresentaram melhor desempenho do que o modelo clássico.

▪ PALAVRAS-CHAVE: Modelos INAR; inferência Bayesiana; MCMC; modelos mistos

1 Introdução

O continente americano experimentou, nos últimos anos, uma sucessão de acontecimentos radicais: chuvas torrenciais na Venezuela, inundações nos pampas argentinos, secas na Amazônia, tempestades de granizo na Bolívia e uma temporada recorde de furacões no Caribe. Ao mesmo tempo, as chuvas diminuem no Chile, no sul do Peru e no sudoeste da Argentina. Com a elevação de temperaturas já registrada ($+1^oC$ na América Central e na América do Sul em um século, ante a média mundial de $+0,74^oC$), os glaciares andinos estão retrocedendo. A disponibilidade de água destinada ao consumo e à geração de eletricidade já está comprometida e o problema se agravará no futuro, tornando-se crônico caso medidas não sejam tomadas, afirma o relatório do IPCC (Painel Intergovernamental sobre Mudanças Climáticas) para a América Latina (SHIKLOMANOV *et. al.*, 2000).

O Brasil tem posição privilegiada no mundo, em relação à disponibilidade de recursos hídricos. A vazão média anual dos rios em território brasileiro é de cerca de 180 mil m^3/s. Esse valor corresponde a aproximadamente 12% da disponibilidade mundial de recursos hídricos, que é de 1,5 milhão de m^3/s (SHIKLOMANOV *et. al.*, 2000). Se forem levadas em conta as vazões oriundas em território estrangeiro e que ingressam no país (Amazônica: 86.321 mil m^3/s; Uruguai: 878 m^3/s e Paraguai: 595 m^3/s), a vazão média total atinge valores da ordem de 267 mil m^3/s (18% da disponibilidade mundial).

Em relação às chuvas, observa-se a tendência já detectada, em estudos anteriores do IPCC, o aumento de até 30%/década da chuva na bacia do Prata e em algumas áreas isoladas do Nordeste. Essa região do Brasil possui apenas 3% de água doce. Em Pernambuco, existem apenas 1.320 litros de água por ano por habitante (TRENBERTH; DAI, 2007).

Em 2013, várias cidades do Agreste de Pernambuco viveram um cenário difícil. As plantas secas e sem folhas, a terra rachada e os açudes vazios geraram ausência do verde na região. Segundo Agência Pernambucana de Águas e Clima (APAC), essa situação se deu devido ao déficit de chuvas em 2012. Nessa lógica apresentada pelos órgãos, 2014 deveria ter sido um ano difícil, pois o mesmo *déficit* ocorreu em 2013. Fato que até agosto do corrente ano, não aconteceu (APAC, 2013).

Os meteorologistas costumam usar dados de séries temporais para avaliar as condições climáticas e previsões. Em muitos estudos, hidrólogos também usam séries temporais para análise dos dados referentes à quantidade de chuva precipitada em uma região para os últimos dias, anos ou um período de 10 anos (GUIMARÃES; SANTOS, 2011; LEE;LEE, 2000). Sendo assim, abordaremos o problema utilizando séries temporais para dados discretos, sob duas óticas: abordagem de estimação dos

parâmetros clássica e abordagem de estimação dos parâmetros bayesiana.

Dentre os modelos de Séries Temporais, uma classe de modelos ainda pouco explorado pela metodologia bayesiana é a modelagem de séries temporais para variáveis discretas. Esses modelos, conhecidos como autoregressivos de valores inteiros INAR(p), são uma adaptação dos modelos AR a dados inteiros em que a operação multiplicação é substituída pela operação *thinning* definida por Steutel & Harn (1979).

A distribuição comumente utilizada para analisar dados na forma de contagem é a distribuição de Poisson, que apesar de bastante útil apresenta uma restrição forte, a igualdade da média e da variância dos dados. Por outro lado, a aproximação bayesiana fornece uma estrutura coerente que facilita a análise de problemas de decisão sobre incertezas (BERGER, 1995). Inicia-se a análise bayesiana para seleção do modelo através da denominação das probabilidades usadas para cada modelo e, posteriormente, escolhe-se as distribuições a priori para os parâmetros que ainda são desconhecidos. Tipicamente, seleciona-se o modelo ou modelos com a maior probabilidade a posteriori.

A investigação de diferentes distribuições de probabilidade a priori é objeto deste estudo, já que estas distribuições são características intrínsecas dos parâmetros dos modelos. Segundo Berger (1995), a primeira vantagem do procedimento bayesiano é a simplicidade na interpretação das conclusões relacionadas às probabilidades a posteriori.

A segunda vantagem é a consistência, quando se aumenta o número de dados, pois garante a seleção do modelo mais apropriado ou de um modelo próximo. Os métodos clássicos, tipicamente, falham neste critério mínimo já que os modelos selecionados se tornam complexos quando há uma grande quantidade de dados. E, a terceira vantagem, apresentada pelo autor, é a possibilidade de se poder incluir a incerteza na seleção do modelo bayesiano. Além das vantagens, anteriormente, apresentadas, ainda existe a possibilidade da seleção do modelo bayesiano ser aplicada na comparação de diversos modelos e aplicada, de modo genérico, ou seja, sem que estes modelos pertençam à família padrão e nem estar agrupados segundo o tipo.

Conforme Berger e Insua (1996), as duas dificuldades na utilização do modelo Bayesiano são a escolha da distribuição a priori e a computação do modelo escolhido. Porém, a escolha da distribuição a priori é considerada o maior problema. Este pode ser o caso em que o conhecimento subjetivo sobre os parâmetros desconhecidos é avaliado e pode ser incorporado à subjetividade própria das densidades a priori para estes parâmetros. Isto é, claramente, desejável se puder ser realizado. Algumas ferramentas computacionais recentes têm permitido a aplicação de métodos bayesianos para modelos de alta complexidade e não padronizados. Na verdade, para modelos mais complicados, a análise bayesiana tenha, talvez, se tornado o mais simples, e frequentemente o único, método de análise.

O objetivo deste trabalho foi modelar uma série temporal de dados discretos,

com abordagem clássica e comparar com o modelo sob abordagem bayesiana, na estimação dos parâmetros, analisando a partir dos critérios de seleção de modelo e os tipos de erros. Os dados discretos utilizados referem-se ao número de dias com precipitação no município de Garanhuns, no período de 11/1993 à 02/2013, totalizando 220 observações.

2 Material e métodos

Referenciando-se nos trabalhos anteriores objetivamos estudar o comportamento das séries de número diário de precipitações (valores inteiros não negativos) referente ao município de Garanhuns, agreste pernambucano. Utilizou-se para isso, técnicas para séries temporais de contagem, afim de modelar o comportamento e prever esse número, comparando-o entre os modelos Poisson e Poisson-Normal e contribuindo na elaboração de políticas públicas ambientais.

Os dados foram ajustados em quatro modelos autoregressivos para valores inteiros, dentre os quais tiveram abordagem bayesiana na estimação dos parâmetros. Utilizou-se a distribuição Poisson e Poisson-Normal, combinadas a prioris Normal e Beta.

A seleção dos modelos se deu através dos critérios AIC, AICc e BIC, não para escolher a ordem e sim qual modelo melhor se ajustou aos dados. Esses são critérios de informação que baseiam-se fundamentalmente na qualidade do ajuste, através da quantidade de termos de n, mas também penalizam a inclusão de parâmetros extras.

2.1 Descrição dos dados

Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE, 2013), o município de Garanhuns, dista 288 km da capital Recife e é constituído de 135.138 habitantes, correspondendo a 22% da população do agreste meridional de Pernambuco. Sua área é de 458.552 km^2, sendo o nono maior município pernambucano e um dos principais pólos de turismo do Estado (Figura 1).

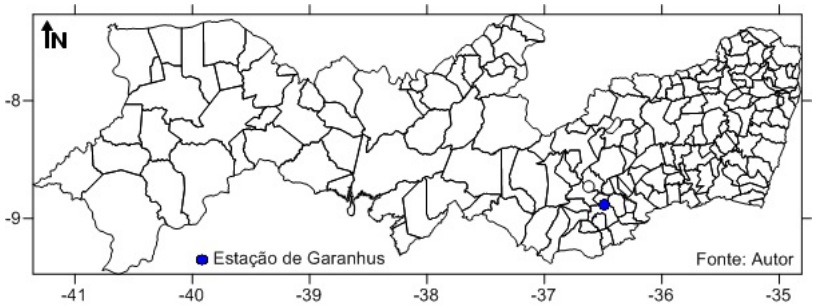


Figura 1 - Disposição geográfica do Município de Garanhuns-PE.

Utilizou-se uma série de tempo obtida da estação meteorológica n° 82893 - Garanhuns - PE, relacionada ao número de dias com precipitação, localizada no nordeste do Brasil, com Latitude: -8,88, Longitude: -36,51, altitude: 822,76 m, a partir de 30/11/1993 à 27/02/2012. Os valores de número de dias de precipitação por mês, do Instituto Nacional de Meteorologia (INMET, 2014).

2.2 Inferência bayesiana para processos INAR(1) de Poisson

Considerando um processo INAR(1), se se pretender obter uma distribuição de Poisson para a distribuição de $\{Y_t\}$ é necessário que $\{\epsilon_t\}$ também tenha uma ditribuição de Poisson. Mais concretamente, $\{Y_t\}$ tem distribuição de Poisson com média $\lambda/(1-\alpha)$ se somente se $\{\epsilon_t\}$ tem distribuição de Poisson de média λ. Portanto, a distribuição de Poisson tem um papel análogo à distribuição Normal do modelo AR. Considera-se o processo autoregressivo de valor inteiro de primeira ordem INAR(1) dado por:

$$y_t = \alpha \circ y_{t-1} + \epsilon_t, t \geq 2, \tag{1}$$

onde a operação "$\circ$"é a operação *thinning* binomial, $\alpha \in]0,1]$ e $\{\epsilon_t\}$ é uma sucessão de variáveis aleatórias de Poisson de parâmetros λ, não correlacionadas e independentes de y_{t-1}.

Sob estas condições e dado y_1 a função verossimilhança da amostra $y = (y_2, ..., y_n)$ é dada por,

$$\begin{array}{ll} l(y, \alpha, \lambda | y_1) & = P(Y_2 = y_2, ..., Y_n = y_n | y_1) = \prod_{t=2}^{n} P(Y_t = y_t | Y_{t-1} = y_{t-1}) \\ & = \prod_{t=2}^{n} P(\alpha \circ Y_{t-1} + \epsilon_t = y_t | Y_{t-1} = y_{t-1}) = \prod_{t=2}^{n} p(y_t | y_{t-1}). \end{array} \tag{2}$$

A variável $Y_t|Y_{t-1}$ é a convolução da distribuição binomial de parâmetros Y_{t-1} e α com a distribuição de Poisson de parâmetro λ, portanto a sua função massa de probabilidade é dada por,

$$\begin{array}{rcl} p(y_t|y_{t-1}) & = & \displaystyle\sum_{i=0}^{min(y_t, y_{t-1})} P(\epsilon_t = y_t - i) \times P(\alpha \circ Y_{t-1} = i | Y_{t-1} = y_{t-1}) \\ & = & \displaystyle\sum_{i=0}^{min(y_t, y_{t-1})} e^{-\lambda} \frac{\lambda^{y_t - i}}{(y_t - i)!} \times \sum_{i=0}^{min(y_t, y_{t-1})} C_i^{y_{t-1}} \alpha^i (1-\alpha)^{y_{t-1} - i}. \end{array}$$

Deste modo, a função de verossimilhança condicional a y_1é dada por:

$$l(y,\alpha,\lambda|y_1)=\prod_{t=2}^{n}e^{-\lambda}\sum_{i=0}^{M_t}\frac{\lambda^{y_t-i}}{(y_t-i)!}C_i^{y_{t-1}}\alpha^i(1-\alpha)^{y_{t-1}-i}, \tag{3}$$

onde $M_t=min(y_t,y_{t-1})$, $t=2,...,n$.

Consideremos a distribuição beta como distribuição a priori para o parâmetro α e a distribuição gama como distribuição a priori para o parâmetro λ, isto é,

$$\begin{aligned} h(\alpha)&\propto\alpha^{a-1}(1-\alpha)^{b-1} \quad , \quad \alpha\sim Be(a,b), a,b>0 \\ h(\lambda)&\propto\lambda^{c-1}exp(-d\lambda) \quad , \quad \lambda\sim Ga(c,d), c,d>0. \end{aligned} \tag{4}$$

A escolha das distribuições de beta e gama para as distribuições a priori dos parâmetros prende-se com o fato de serem conjugadas da binomial e poisson, respectivamente.

Supondo α e λ independentes, a distribuição a priori conjunta é dada por,

$$h(\alpha,\lambda)\propto\lambda^{c-1}exp(-d\lambda)\alpha^{a-1}(1-\alpha)^b-1, \tag{5}$$

onde $\lambda>0$, $0<\alpha<1$ e os hiperparâmetros a,b,c e d são conhecidos e positivos. Nota-se que se $a\to 0$, $b\to 0$, $c\to 0$ e $d\to 0$ tem-se o caso de uma distribuição a priori não informativa.

Assim, a distribuição a posteriori conjunta é dada por,

$$h(\alpha,\lambda|y)\propto exp[-(d+n)\lambda]\lambda^{c-1}\alpha^{a-1}(1-\alpha)^{b-1}\times\prod_{t=2}^{n}\sum_{i=0}^{min(y_t,y_{t-1})}\frac{\lambda^{y_t-i}}{(y_t-i)!}C_i^{y_{t-1}}\alpha^i(1-\alpha)^{y_{t-1}-i}$$

Integrando a distribuição marginal a posteriori para λ em ordem α é dada por:

$$h(\lambda|y)\propto\int\left[exp[-(d+n)\lambda]\lambda^{c-1}\alpha^{a-1}(1-\alpha)^{b-1}\times\prod_{t=2}^{n}\sum_{i=0}^{min(y_t,y_{t-1})}\frac{\lambda^{y_t-i}}{(y_t-i)!}C_i^{y_{t-1}}\alpha^i(1-\alpha)^{y_{t-1}-i}\right]d\alpha.$$

Integrando a distribuição marginal a posteriori para α em ordem λ é dada por:

$$h(\alpha|y)\propto\int\left[exp[-(d+n)\lambda]\lambda^{c-1}\alpha^{a-1}(1-\alpha)^{b-1}\times\prod_{t=2}^{n}\sum_{i=0}^{min(y_t,y_{t-1})}\frac{\lambda^{y_t-i}}{(y_t-i)!}C_i^{y_{t-1}}\alpha^i(1-\alpha)^{y_{t-1}-i}\right]d\lambda.$$

Como se pode verificar estas integrais são muito complexas e portanto é necessário utilizar a metodologia de Gibbs para obter as estimativas de λ e α. Para

tal devem-se calcular as distribuições condicionais completas para os parâmetros λ e α

A distribuição condicional completa para o parâmetro λ é

$$h(\lambda|\alpha, y) \propto exp[-(d+n)\lambda]\lambda^{c-1} \times \prod_{t=2}^{n} \sum_{i=0}^{M_t} L(t,i)\lambda^{y_t - i}, \tag{6}$$

onde

$$L(t,i) = \frac{1}{(y_t - i)!} C_i^{y_{t-1}} \alpha^i (1-\alpha)^{y_{t-1}-i}, \lambda > 0. \tag{7}$$

A distribuição dada em 6 é uma combinação linear de funções de densidade de probabilidade de variáveis aleatórias com distribuição gama.

Analogamente a distribuição condicional completa para α é dada por:

$$h(\alpha|\lambda, y) \propto \alpha^{a-1}(1-\alpha)^{b-1} \times \prod_{t=2}^{n} \sum_{i=0}^{M_t} K(t,i)\alpha^i (1-\alpha)^{y_{t-1}-i}, \tag{8}$$

onde

$$K(t,i) = \frac{\lambda^{y_t - i}}{(y_t - i)!} C_i^{y_{t-1}}, 0 < \alpha < 1. \tag{9}$$

A distribuição dada em 8 é uma combinação linear de funções densidade de probabilidade de variáveis aleatórias com distribuição beta.

Agora que são conhecidas as distribuições condicionais completas para os parâmetros pode-se utilizar o algoritmo de Gibbs para determinar as estimativas.

3 Implementação computacional

A implementação computacional foi realizada usando-se o programa WinBUGS 1.4 (SPIEGELHALTER *et. al.*, 2002). Para cada modelo, foram geradas cadeias com 5000 iterações para cada parâmetro. As cadeias foram inicializadas em um mesmo ponto e tiveram a convergência monitorada pelo critério de convergência de Gelman e Rubin (1992) existente no programa WinBUGS.

Em todas as simulações os valores utilizados foram $\alpha = 0,25$ e $\lambda = 1$. O modelo Clássico foi analisado no *software* R.

3.1 Diagnóstico de Convergência

Uma das questões mais importantes relativas à aplicação de todas as técnicas MCMC é a questão de quantas repetições iniciais têm de ser descartadas de modo a evitar a possibilidade de viés na estimativa média dos parâmetros causada pelo efeito dos valores iniciais. O parâmetro é analisado, do período de tempo inicial até um número de iteração definido pelo pesquisador, objetivando alcançar a convergência. O tempo que se leva do ponto inicial até o início da convergência é comumente chamado de *burn-in*. Infelizmente, para quase todas as configurações a priori de técnicas utilizando o MCMC, para determinar a taxa de convergência da cadeia, faz-se necessário o *burn-in*. Portanto, é indispensável a realização de algumas análise estatística baseada nos resultados para avaliar a convergência. Para a análise da convergência, usou-se neste trabalho o método de convergência informal.

4 Método de Anderson-Darling

Para confirmar o ajuste gráfico, alguns teste de hipóteses não paramétricos podem ser utilizados. Estes testes consideram a forma da distribuição da população em lugar dos parâmetros (ROMEU, 2003). Por este motivo são chamados de testes não-paramétricos. As medidas de ajuste dependem do método de estimação utilizado, sendo o teste de Anderson-Darling, usado para os métodos de máxima verossimilhança e de mínimos quadrados. É uma medida da proximidade dos pontos e da reta estimada no gráfico de probabilidade. O teste de Anderson-Darling é um teste alternativo aos teste de aderência de Chi-quadrado e Kolmogorov-Sminov, o qual tem a vantagem de ser mais sensível que os dois mencionados, pois dá mais peso aos pontos das caudas de distribuição. Assim, valores pequenos da estatística de Anderson-Darling indicam que a distribuição estima melhor os dados (STEPHENS, 1974).

Para estabelecer um critério de rejeição ou não rejeição do modelo (distribuição de probabilidade), é formulada o seguinte teste de hipótese:

$$\begin{cases} H_0, & \text{Y segue uma distribuição de probabilidade} \\ H_1, & \text{não segue uma determinada distribuição de probabilidade proposta.} \end{cases}$$

A estatística do teste para tomar a decisão é dada por:

$$A^2 = -n - \sum_{i=1}^{n} \frac{(2i-1)}{n} [lnF(x_i) + ln(1 - F(x_{n+1-i}))] \tag{10}$$

em que F é a função de distribuição acumulada da distribuição especifica. Observe que $x_{(i)}$ são os dados ordenados. Os valores críticos ou de rejeição para o teste de Anderson-Darling dependem da distribuição específica que está sendo testada. O

teste de Anderson-Darling é um teste unicaudal e a hipótese nula H_0 é rejeitada se o teste estatístico fornecer valor superior ao crítico. Cabe observar que este teste pode ser ajustado, multiplicando-no por uma constante que depende do tamanho da amostra (NIST, 2002).

4.1 Medidas de erros de previsão

No estudo das técnicas de previsão as medidas de precisão são uma aplicação de extrema importância. Os valores futuros das variáveis tornam-se bastante difíceis de prever dada a complexidade da grande maioria dessas variáveis na vida real. Assim, é fundamental incluir informação acerca da medida em que a previsão pode desviar-se do valor real da variável. Este conhecimento adicional fornece uma melhor percepção sobre o quão precisa pode ser a previsão, ver (STEVENSON, 1996).

A diferença entre o valor real e a previsão do valor dá origem ao erro de previsão:

$\epsilon_t = A_t - P_t$, onde:

ϵ_t = Erro no período t;

A_t = Valor real no período t;

P_t = Previsão para o período t.

Erro quadrático médio (EQM)

O erro quadrático médio (EQM) também pode ser usado como uma medida do erro de previsão. O EQM é determinado somando os erros de previsão ao quadrado e dividindo pelo número de erros usados no cálculo, (FERREIRA; VASCONCELOS; ADEODATO, 2005). O erro quadrático médio pode ser expresso pela seguinte equação:

$$EQM = \frac{\sum_{t=1}^{n} \epsilon_t^2}{n}. \tag{11}$$

Deseja-se que essa métrica seja a menor possível, em uma previsão perfeita o EQM será zero.

Erro percentual (EPT)

O erro percentual mede a percentagem do erro em relação ao valor real. Calcula-se, subtraindo ao valor real no período t a previsão no respectivo período e divide-se o resultado pelo valor real utilizado anteriormente.

$$EPT = \frac{(A_t - P_t)}{A_t} \times 100 \tag{12}$$

MAPE

De acordo com Ferreira, Vasconcelos e Adeodato (2005) o erro médio percentual absoluto (MAPE - *Mean Absolute Percent Error*).

$$MAPE = \frac{1}{n}\left|\sum_{t=1}^{n}\frac{(A_t - P_t)}{A_t}\right|, \tag{13}$$

Para esta métrica quanto menor o valor, melhor são as previsões geradas pelo modelo. Na previsão perfeita irá assumir valor zero.

5 Resultados e discussões

Considerou-se que o número de dias mensais com precipitação em um mês está relacionado probabilisticamente com número de dias com precipitação em um mês anterior, assim como com o número de dias com precipitação em um mês posterior.

5.1 Análise descritiva dos dados

A série completa inicia-se no mês de novembro de 1993 até fevereiro de 2013, totalizando 232 observações. Para o ajuste do modelo retirou-se os 12 últimos meses de observações a fim de comparar com as medidas de previsão ao fim do ajuste.

A Tabela 1 contém as medidas descritivas da série do número de dias com precipitação. Tem-se o valor mínimo de 0, ou seja, houve meses em que não ocorreu precipitação. Já em outros meses houve precipitação em todos os dias do mês (máximo = 30), considerando meses com 30 dias.

Tabela 1 - Medidas descritivas da série de número de dias com precipitação

	Nº Observações	Mínimo	Máximo	Média	Desvio padrão
Série Temporal	220	0	30	12,61	8,53

Os valores das médias por mês, da série temporal número de dias com precipitação, estão apresentadas na Tabela 2

Na Figura 2 apresenta-se a distribuição da média mensal do período analisado, observando um comportamento próximo ao da distribuição normal.

A função autocorrelação parcial amostral da série número de dias com precipitação estão representados na Figura 3.

Tabela 2 - Medidas descritivas da série de número de dias com precipitação por mês

Meses	Mínimo	Máximo	Média	Desvio padrão
Janeiro	0	21	6,68	5,78
Fevereiro	2	13	7,00	3,23
Março	3	16	8,50	3,96
Abril	5	21	13,67	4,06
Maio	4	25	19,11	5,30
Junho	17	29	23,11	3,77
Julho	15	30	25,33	3,51
Agosto	14	27	22,50	3,07
Setembro	2	21	12,67	5,12
Outubro	0	12	5,50	2,96
Novembro	0	11	4,37	3,32
Dezembro	0	14	4,37	3,47

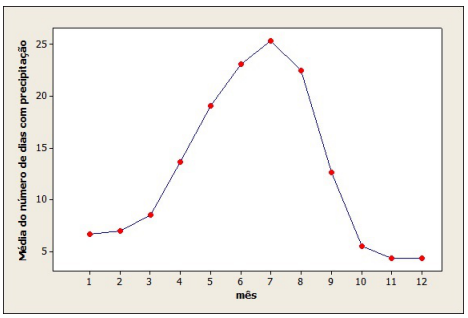


Figura 2 - Média mensal do número de dias com precipitação.

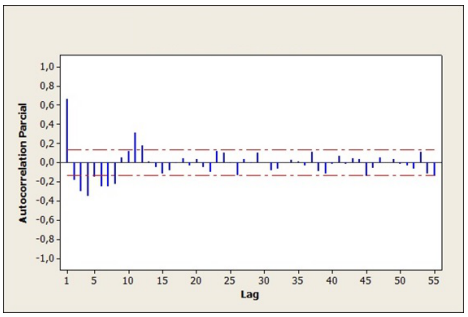


Figura 3 - Função de autocorrelação parcial, com 5% de limites de significância.

5.2 Modelos ajustados

Neste trabalho, aplicou-se o modelo INAR (1) aos dados. A abordagem foi feita sob o ponto de vista clássico e bayesiano, comparando-se assim os resultados encontrado.

Os dados foram ajustados com as seguintes abordagens:

- Clássica - Modelo INAR(1) Poisson (Modelo 1),
- Bayesiana - Modelo INAR(1) Poisson, priori Beta (Modelo 2),
- Bayesiana - Modelo INAR(1) Poisson, priori Normal (Modelo 3).

Considera-se então um modelo INAR(1) dado por:

$$y_t = \alpha \circ y_{t-1} + \epsilon_t, t \geq 2. \tag{14}$$

O Modelo 1, terá abordagem clássica. Na abordagem bayesiana usaremos distribuições a priori conjugadas e não informativas. Assim, assumiremos que:

- Modelo 2: $\alpha \sim dbeta(10^{-4}, 10^{-4})$ e $\lambda \sim dgamma(10^{-4}, 10^{-4})$ para valores iniciais $\alpha = 0,25$ e $\lambda = 1$, priori conjugada;
- Modelo 3: $\alpha \sim dnorm(10^{-4}, 10^{-4})$ e $\lambda \sim dnorm(10^{-4}, 10^{-4})$ para valores iniciais $\alpha = 0,25$ e $\lambda = 1$, priori não informativa.

Encontrou-se assim os seguintes resultados:

5.3 Clássica - Modelo INAR(1) Poisson (Modelo 1)

A modelagem gerou os valores $\hat{\alpha} = 0,6629$ e $\hat{\lambda} = 12,3289$, com distribuição de probabilidade apresentada na Figura 4 e boxplot da série na Figura 5

Na Figura 5, 50% dos dados estão entre 5 e 20. 25% dos dados estão abaixo de 5 e 25% acima de 20. Dentre os valores situados entre 5 e 20, existe uma menor dispersão dos dados no intervalo 5 e 12, em comparação ao intervalo 12 e 20.

5.4 Bayesiana - Modelo INAR(1) Poisson, priori Beta (Modelo 2)

Resultando um valor para $\hat{\alpha} = 0,6549$ e $\hat{\lambda} = 12,6500$, com função marginal a posterior para cada parâmetros, apresentada na Figura 6.

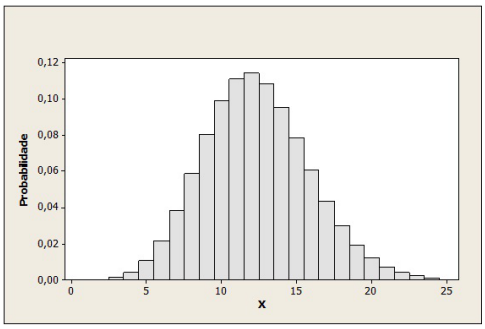


Figura 4 - Histograma da distribuição de probabilidade dos resultados do Modelo 1.

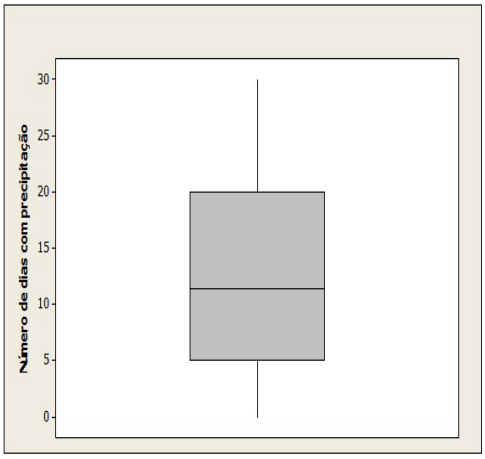


Figura 5 - Box-plot da série temporal.

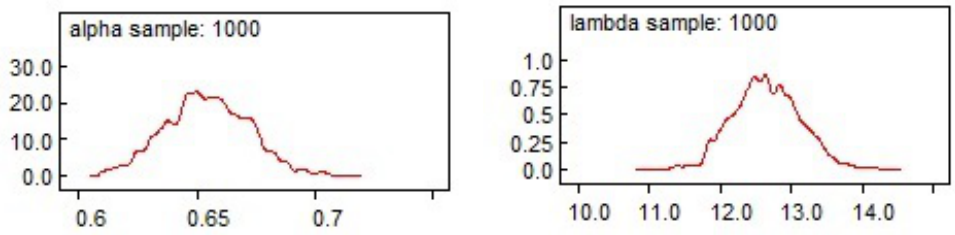


Figura 6 - Distribuição a posteriori marginal dos parâmetros α e λ do Modelo 2.

Foram realizadas 5000 iterações para os parâmetros do modelo 2, e verificamos na Figura 6, que a função marginal a posteriori usou apenas as 1000 iterações finais, descartando assim as 4000 primeiras iterações (*burn-in*). Observa-se que a distribuição dos parâmetros tem uma concentração de informação em torno da média. Considerando os valores iniciais para $\alpha = 0,25$ e $\lambda = 1$ a Figura 7 mostra a dinâmica da convergência dos parâmetros do Modelo 2.

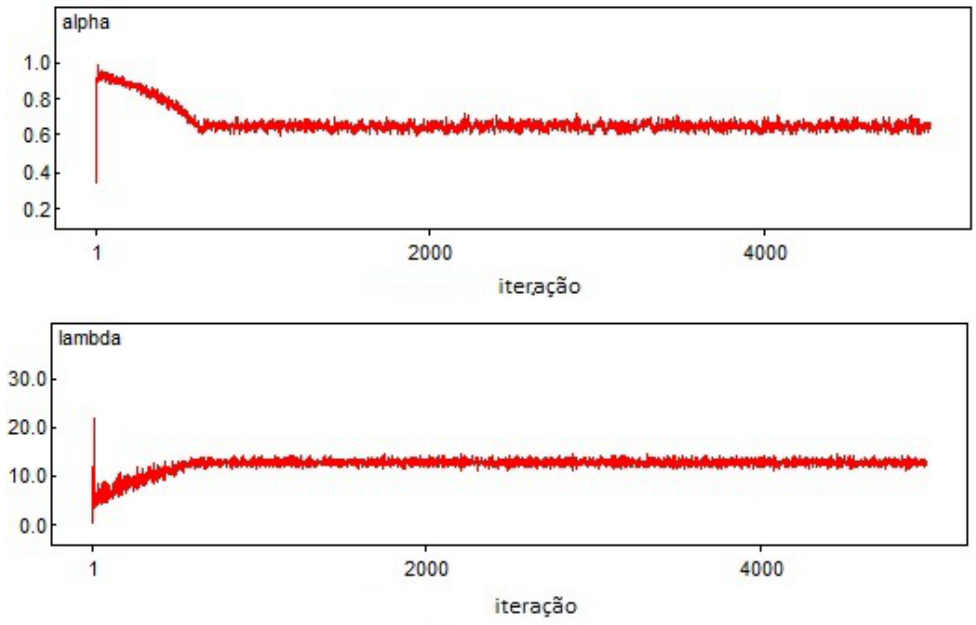


Figura 7 - Convergência dos parâmetros do Modelo 2.

Na Figura 8, verificamos o intervalo de variação do parâmetros, nas 1000 iterações finais.

Na Figura 9, apresentamos as funções de autocorrelação dos parâmetros, onde a suavidade no decaimento indica um comportamento autoregressivo.

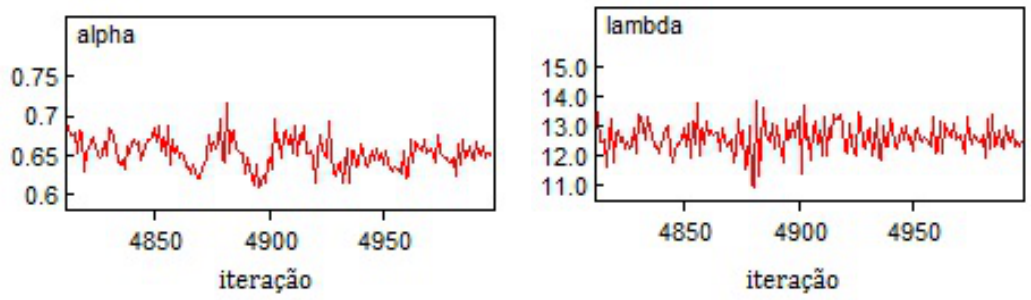


Figura 8 - Traço da variabilidade dos parâmetros nas 1000 iterações finais (Modelo 2).

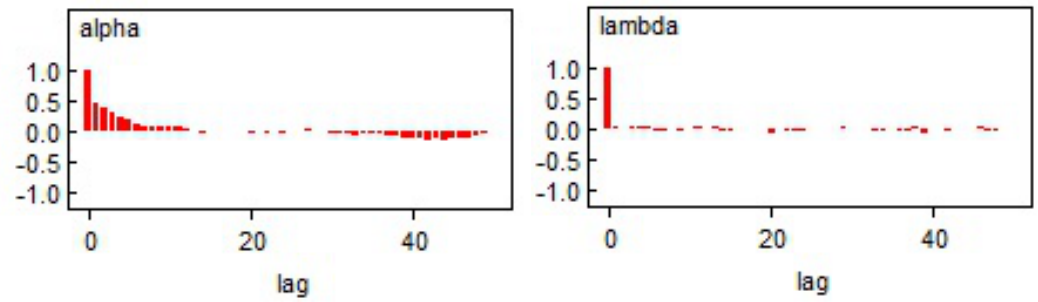


Figura 9 - Função de autocorrelação parcial para os parâmetros α e λ do Modelo 2.

5.5 Bayesiana - Modelo INAR(1) Poisson, priori Normal (Modelo 3)

Foram realizadas 5000 iterações para os parâmetros do modelo 3, e verificamos na Figura 10, que a função marginal à posteriori usou apenas as 1000 iterações finais, descartando assim as 4000 primeiras iterações (*burn-in*). Verificamos na Figura 10, que a distribuição dos parâmetros tem uma concentração de informação em torno da média. Resultando um valor para $\hat{\alpha} = 0,6509$ e $\hat{\lambda} = 12,6200$, com função marginal a posterior para cada parâmetros, apresentada na Figura 10

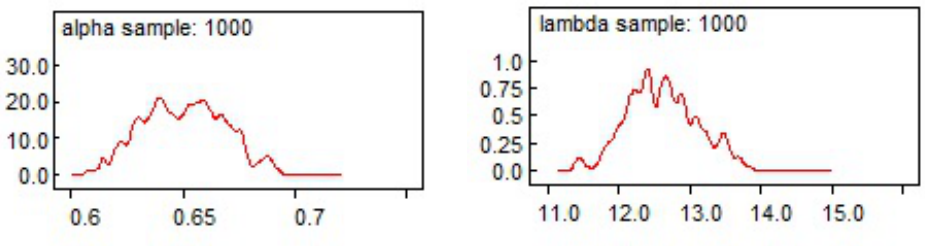


Figura 10 - Distribuição a posteriori marginal dos parâmetros α e λ do Modelo 3.

Considerando os valores iniciais para $\alpha = 0,25$ e $\lambda = 1$ a Figura 11 mostra a dinâmica da convergência dos parâmetros do Modelo 3.

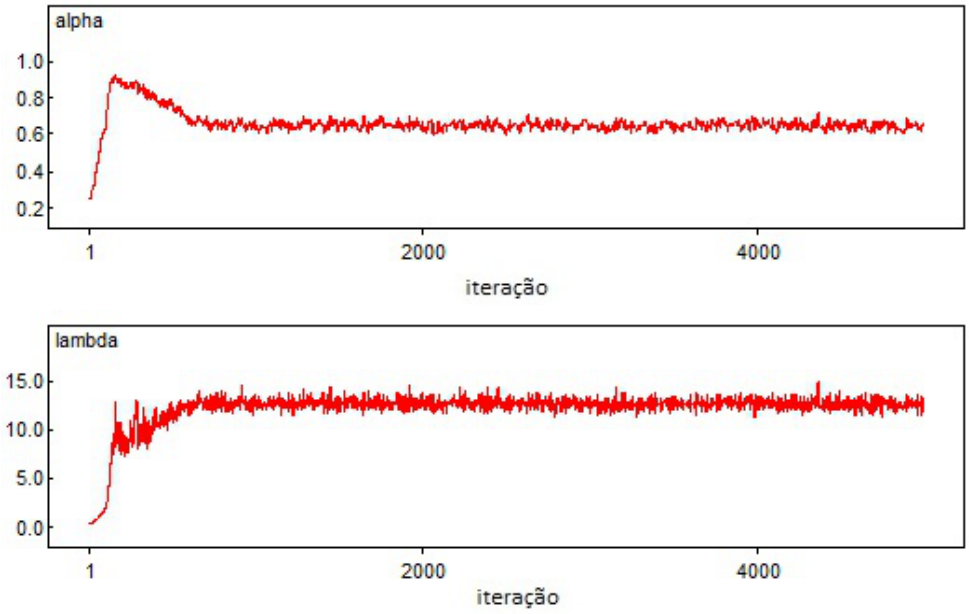


Figura 11 - Convergência dos parâmetros do Modelo 3.

Na Figura 12, verificamos o intervalo de variação do parâmetros, nas 1000 iterações finais.

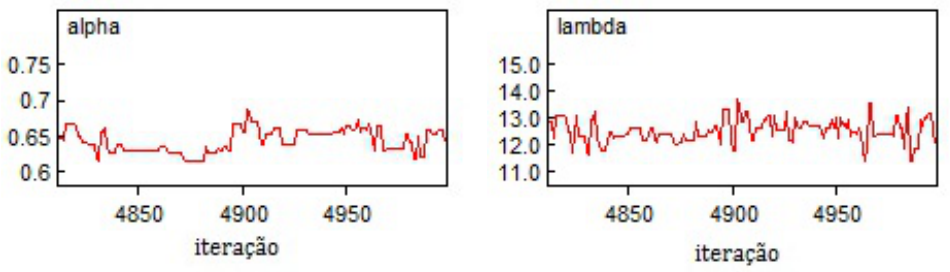


Figura 12 - Traço da variabilidade dos parâmetros nas 1000 iterações finais (Modelo 3).

Na Figura 13 são apresentadas as funções de autocorrelação dos parâmetros, onde a suavidade no decaimento indica um comportamento autoregressivo.

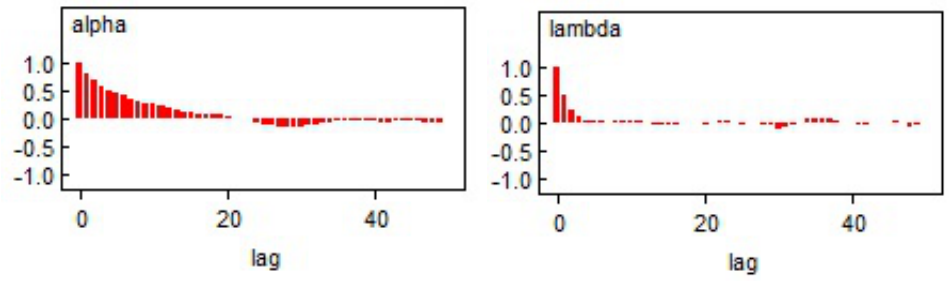
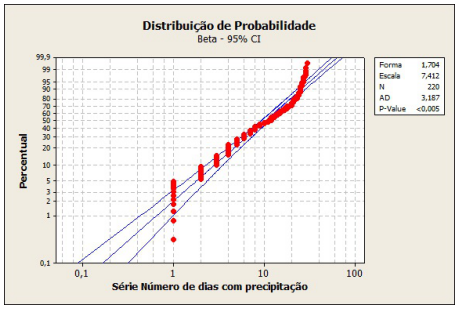


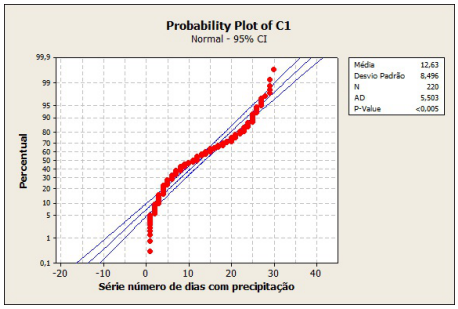
Figura 13 - Função de autocorrelação parcial para os parâmetros α e λ do Modelo 3.

5.6 Distribuição de probabilidade

Segundo Steulen e Harn (1979), ao adaptar o modelo AR com o operador *thinning*, compara-se o resultado obtido com essa abordagem e o efeito de utilizar a parte inteira do dado contínuo, relatando que o usual era aplicar modelos contínuos mesmo em dados discretos. Considerando a forma de modelar dados discretos antes de Steulen e Harn (1979), aplicamos os modelos Normal e Beta na série e verificamos qual das duas se ajustam melhor aos dados, objetivando analisar se essa abordagem influência no resultado obtido com a proposta de abordar modelos de dados discretos com modelos mistos. As Figuras 14(a) e 14(b), apresentam as distribuições de probabilidade Beta e Normal, respectivamente.



(a) Distribuições Beta



(b) Distribuições Normal

Figura 14 - Teste de aderência (Anderson-Darling) para as distribuições Beta e Normal.

Usando o teste de Anderson-Darling, encontrou-se para a distribuição Beta o valor de 3,187, e, para distribuição Normal o valor de 5,503. Assim segundo esse teste, a distribuição Beta apresenta um ajuste melhor aos dados se comparado com a distribuição Normal.

5.7 Previsão

Nesta seção, será considerado os cálculos das previsões para o processo autoregressivo de valores inteiros definido em 1, seja y_{n+h} a previsão de um valor futuro.

Uma vez estimados parâmetros dos modelos utilizados neste trabalho através dos dados valores observados da série (n=220), foi realizado a comparação da previsão dos 12 meses primeiros meses após último mês dos dados utilizados na modelagem da série.

Na Tabela 3 é apresentado os valores reais e as previsões encontradas a partir da aplicação de cada modelo, utilizando os tipos de erro: Erro Percentual Total (EPT), Erro Absoluto Médio Percentual (MAPE) e Erro Quadrático Médio (EQM).

Observando os resultados indicados na Tabela 3, verificou-se que os modelo 2 e 3 apresentaram menores erro nos métodos de análise de erro utilizado. Logo os modelos onde os parâmetros foram estimados via abordagem bayesiana apresentaram uma melhor previsão que o modelo via abordagem clássica

Tabela 3 - Valores previstos pelos modelos (12 passos à frente) e erros de previsão

Valor observado	Previsão Clássica	Previsão Bayesiana Modelo 2	Previsão Bayesiana Modelo 3
4	3	3	3
5	2	2	3
12	3	3	3
22	7	7	8
29	14	14	14
22	19	18	19
8	14	14	14
8	5	5	5
1	5	5	5
4	0	0	1
7	2	2	3
1	4	4	5
EPT	36,59%	37,18%	32,79%
MAPE	1,064	1,068	1,095
EQM	23,61	23,86	22,07

Conclusões

Este trabalho tratou da utilização de técnicas de modelagem e previsão em séries temporais de valores discretos de ordem 1, com abordagem clássica e abordagem bayesiana. A série é referente ao número de dias com precipitação no município de Garanhuns, agreste meridional de Pernambuco.

Os modelos INAR(1) utilizados consideram a característica de contagem dos dados, retornando em suas simulações apenas valores inteiros e positivos, obtendo em suas modelagens e previsões valores que acompanharam as séries de contagem. Os modelos INAR explicaram as características da série apenas com os dados da própria série, baseados nas características probabilísticas dos mesmas. A comparação entre os modelos INAR(1) Poisson, foi realizada através da análise de erros de previsão e pelos critérios AIC, AICc e BIC.

Com a distribuição da média aproximando de uma distribuição Normal, utilizou-se o teste de Anderson-Darling e comparou-se com a distribuição Beta proposta por Silva, Pereira e Silva (2009). Verificou-se que os dados se adequaram melhor a distribuição Beta. Assim, ajustou-se a Poisson com a Normal usando prioris Normal e Beta, tendo em todos os critérios de informação (AIC, AICc e BIC) o Modelo 4 como sendo o melhor modelo ajustado.

Conclui-se que na comparação de abordagem clássica e bayesiana, os modelos bayesianos apresentaram resultados mais consistentes. Na comparação entre os

modelos bayesianos, concluímos que os modelos mistos apresentaram melhores ajustes que os indicados na literatura.

Por fim, deve-se enfatizar que o uso de técnicas de simulação MCMC torna o processo de inferência bayesiana mais poderoso e flexível. Além disso pode ser estendido para problemas de dimensão maior.

SILVA, D. A. N.; CUNHA FILHO, M.; FALCÃO, A. P. S. T.; ALVES, G. I. L. Classical and Bayesian estimation for INAR (1) models in number of precipitation days in Garanhuns-PE. *Rev. Bras. Biom.*, Lavras, v.34, n.1, p.63-83, 2016.

▪ *ABSTRACT: Many aspects of the weather cycle could be described by time series data. Meteorologists often use time series data to assess climate conditions and forecasts. Such models are generally continuous models. The interest was to analyze discrete weather data with the INAR (1) model, using classical and Bayesian approach to parameter estimation. The proposal is to analyze the data series utiizando mixed models with Bayesian approach. Thus, this work is described a sequence of procedures for estimating parameters of autoregressive models of order $p = 1$, for integer values INAR(1), by classical inference via maximum likelihood estimator and Bayesian inference via simulation Monte Carlo Markov Chain (MCMC). Two alternatives are considered for the a priori density of the model parameters. For the former case is adopted a density non-priori information. For the second, we adopt a density combined beta-gamma. A posteriori analysis is performed by algorithms of MCMC simulation. Also evaluates the prediction of new values of the series number of days with precipitation. The period of analysis comprised* 30/11/ *1993 to* 29/02/2012 *and obtained estimates of the period of* 31/03/2012 *to* 28/02/2013. *One INAR (1) model of classical parameter estimation and two models INAR (1) Bayesian estimation for the parameters were used. The choice of the most appropriate model the Akaike information criterion (AIC) was used. The analysis of forecast errors was an instrument used to determine which model is best suited to the data. We conclude that the use of MCMC simulation makes the process more flexible Bayesian inference and can be extended to larger problems. Bayesina models showed better performance than the classical model.*

▪ *KEYWORDS: INAR models; Bayesian inference; MCMC; mixed models*

Referências

APAC. *Agência Pernambucana de Águas e Clima.* 2013. Disponível em: http://www.apac.pe.gov.br/. Acessado em: 18/09/2013.

BERGER, J. O. *Statistical decision theory and Bayesian analysis (2nd edition).* [S.l.]:Springer, 1995.

BERGER, J. O..; INSUA, D. R. *Recent developments in Bayesian inference with applications in hydrology.* [S.l.]: Censiglio and nazionale delle ricerche, 1996.

FERREIRA, T. A. E.; VASCONCELOS, G. C.; ADEODATO, P. J. L. A new evolutionary methos for time series forecasting, In: *Proceedings of the 2005*

conference on Genelir and evolutionary computations. New York, NY, USA; ACM, 2005. (GECCO'05), p.2221-2222.

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. *Journal Statistical Science*. JSTOR. p.457-472. 1992.

GUIMARÃES, R; SANTOS, E. G. Principles of stochastic generation of hydrologic: time series for reservoir planning and design: A case study. *Journal of Hydrologic Engineering*, v.16, n.11, p.891-898, 2011.

IBGE. *Instituto Brasileiro de Geografia e Estatística*. 2013. Disponível em: http://www.ibge.gov.br/home/. Acessado em: 18/09/2013.

INMET. *Instituto Nacional de Meteorologia*. 2014. Disponível em : http://www.inmet.gov.br/portal/. Acessado em: 10/05/2014.

LEE, J. Y.; LEE, K. K. Use of hydrologic time series data for identification of rechargemechanism in a fractured bedrock aquifer system. *Journal of Hydrologic*, Elsevier Science BV, n.229, p.190-201, 2000.

NIST. Engineering Statistics and Handbook of statistical methods. [S.l.:s.n.], 2002.

ROMEU, J. L. Anderson-darling: A goodness of fit test for small samples assumptions. *Selected Topics in Assurance Related Technologies*, v.10, n.5, p.1-6, 2003.

SHIKLOMANOV, I, et. al. The dynamics of river water inflow to the aretic ocean. In: *The Freshwater Budget of the Arctic Ocean.* [S.l.]: Springer, 2000. p.281-296.

SILVA, N.; PERERIRA, I. ; SILVA, M. E. Forecasting in Inar (1) model. Instituto Nacional de Estatística, 2009.

SPIEGELHALTER, D. J. et. al. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* Wiley Online Library, v.64, n.4, p.583-639, 2002.

STEPHENS, M. A. Edf statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, Taylor & Francis. v.69, n.347, p.730-737, 1974.

STEUTEL, F.; HARN, K. V. Discrete analogues of self-decomposability and scability. *The Annals of Probability.* JSTOR, p.893-899, 1979.

STEVENSON, W. J. *Producition / operations management.*[S.l.]; Chicago: Irwin, 1996. (The Irwin series in production operations management). ISBN 9780256197235.

TRENBERTH, K. E.; DAI, A. Effects of mount pinatubo volcanic eruption on the hydrological cycle as an analog of geoengineering. *Geophysical Research Letters*, Wiley Online Library, v.34, n.15, 2007.

Recebido em 03.02.2015.

Aprovado após revisão em 10.11.2015.