BETA METEOROLOGICAL TIME SERIES: APPLICATION TO AIR HUMIDITY DATA

Edilberto Cepeda-Cuervo.[1]

Statistics department

Universidad Nacional de Colombia, Bogotá, Colombia

Marinho G. Andrade[2]

Applied mathematics and statistics department

ICMC, Universidade de São Paulo

Sao Carlos, S.P., Brazil

Jorge Alberto Achcar [3]

Social medicine department

FMRP, Universidade de São Paulo

Ribeirão Preto, S.P., Brazil

Abstract

Time series models are often used in the analysis of meteorological phenomena to model levels of rainfall, temperature and levels of air humidity series in order to make forecasting and generate synthetic series which are inputs for the analysis of the influence of these variables on the quality of life. Relative air humidity for example, has great influence on the count increasing of respiratory diseases, especially for some age populations as newly born and elderly people. In this paper we introduce a new modeling approach for meteorological time series assuming a beta distribution for the data, where both the mean and precision parameters are being modeled. Bayesian methods using standard MCMC (Markov Chain Monte Carlo Methods) are used to simulate samples for the joint posterior distribution of interest. An example is given with a time series of 313 air humidity observations, measured by a wether station of Rio Claro, a city localized in São Paulo state, southeastern of Brazil.

[1]email:ecepedac@unal.edu.co

[2]email:marinho@icmc.usp.br

[3]achcar@fmrp.usp.br

Keywords*: Meteorological time series data, beta distribution, Bayesian analysis, MCMC methods.*

1 Introduction

Many aspects of the meteorological cycle could be described by time series data. Meteorologists, usually use time series data to evaluate the climatic conditions and predictions. In many studies, hydrologists also use time series data for displaying the amount of rainfall that has fallen in a region for the past days, years or a period of 10 years (see for example, Guimaraes & Santos ,2011, and Lee & Lee, 2000)

Time series models are often used in meteorology to model relative air humidity times series in order to make forecasting and to generate synthetic series which are inputs for the analysis of meteorological phenomena (see Abu-Taleb et al., 2007 and Barros et al., 2001). Modeling meteorological variability is very important in the planning and management of resources of the city, region or country. Important wether related variables given by a time series model describe and give estimates for the relative air humidity parameters. A time series model estimates the relative air humidity parameters. Different existing time series models as ARMA and higher orders of MA models have been extensively used by some authors to analyze meteorological time series (Katz, et al., 1981). Other modeling approach also have been used by some authors to analyze meteorological time series as spectral analysis models (Jones, R. H., 1964, and Evan et al., 2011).

The weakly relative air humidity series typically have a periodic behavior in the mean and variance and in general periodic autoregressive models are adopted to analyze de data(Chiawa et al., 2010). Usually, standard probability distributions are assumed to analyze meteorological time series as the normal, log-normal, and gamma distributions can be found in the literature (Cepeda-Cuervo et. al, 2012, and Sarraf et al., 2011).

In this paper it is assumed that the air humidity observations are fitted by a beta distribution $B(p_i, q_i)$, with mean and precision given respectively by (2)

and (3), respectively. To illustrate the application of the modeling approach, we consider a time series related to the weekly humidity average measured in the city of Rio Claro, São Paulo state, located in southeastern Brazil, from 18/10/2002 to 08/10/2008. This time series is showed in Figure 1.

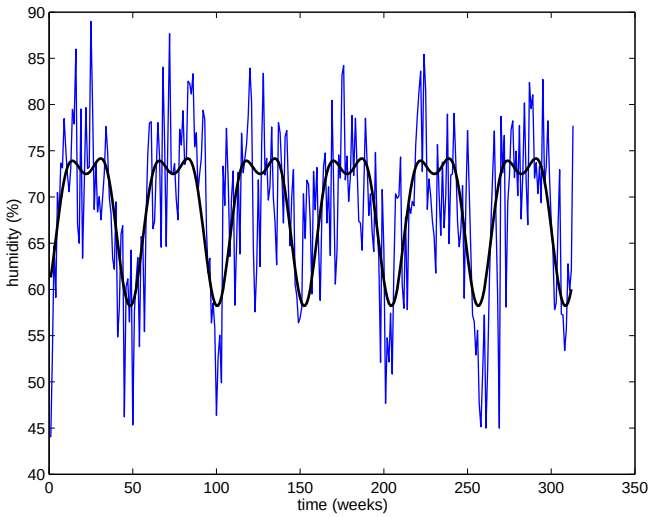


Figure 1: The humidity time series data and the fitted periodical mean

From Figure 1, we can see that the variability in the maximum values of the series is larger than the variability in the valleys. To account for this heteroscedasticity in the time series of streamflows, the model proposed in this paper assumes seasonal and autoregressive models for the mean and for the precision of the models. In this way, we propose a new general periodic and heteroscedastic model, in which the variance also presents a correlation structure between the present data with the past values of the air humidity.

This paper is structured as follows: in section 2, we introduce the proposed model; in section 3, we present a Bayesian methodology to analyze the data; in section 4, we introduce a new proposed model;in section 5, we present an application of the proposed methodology; finally, in section 6, we present some conclusions

2 Beta regression models

A random variable Y has a beta distribution if its probability density function is given by,

$$f(y|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} I_{(0,1)}(y), \tag{1}$$

where $p > 0$, $q > 0$ and $\Gamma(.)$ denotes the gamma function. If Y is a random variable with a beta distribution, the mean $\mu = E(Y)$ and variance $\sigma^2 = Var(Y)$ are given respectively by,

$$\mu = \frac{p}{p+q}, \tag{2}$$

$$\phi = \frac{p\,q}{(p+q)^2(p+q+1)}. \tag{3}$$

Some reparametrizations of the beta distribution given in (1) can be more appropriate to analyze the data. As a first one, let us consider $\phi = p + q$; in this parametrization, we can see that $p = \mu\phi$, $q = \phi(1-\mu)$ and $\sigma^2 = \frac{\mu(1-\mu)}{\phi+1}$. In this case, ϕ can be interpreted as a precision parameter in the sense that, for fixed values of μ, larger values of ϕ correspond to smaller values of the variance of Y. This reparametrization was proposed in Cepeda (2001). With this reparametrization, the density of the beta distribution (1) can be rewritten as in the same way as proposed by Ferrari and Cribari-Neto (2004),by,

$$f(y|\alpha,\beta) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} I_{(0,1)}(y). \tag{4}$$

In this case, the mean and dispersion parameters can be modeled as functions of explanatory variables as it was proposed by Cepeda(2001), where joint modeling of the mean and the dispersion parameter, as proposed initially, are defined, respectively by,

$$\text{logit}(\mu_i) = \mathbf{x}_i^{'}\boldsymbol{\beta} \tag{5}$$

$$\log(\phi_i) = \mathbf{z}_i^{'}\boldsymbol{\gamma} \tag{6}$$

Inferences for the beta regression models have been discussed by many authors under a classical inference approach (see for example, Paolino, 2001, or

Ferrari and Cribari-Neto, 2004; Smithson and Verkuilen, 2006, and Simas et al., 2010) or under a Bayesian approach (see for example, Cepeda, 2001; Branscum et al., 2007). Nonlinear beta regression models were proposed by Cepeda and Achcar (2010).

3 Beta autoregressive models

An autoregressive model is defined as $Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \nu_t$, where $\beta_0, \beta_1, \ldots, \beta_p$ are the regression parameters of the model and ν is a white noise. Some restrictions are imposed in this model to be wide-sense stationary: the roots of the polynomial,

$$P(z) = z^p - \sum_{i=1}^{p} \beta_i z^{p-i} \tag{7}$$

should fall within the unit circle. Thus, in the case of $AR(1)$ models, $|\beta_1|$ should have a value at least equal to one. The variance and auto-covariance of Y_t, $t = 1, 2, \ldots, n$ are given by:

$$Var(Y_t) = \frac{\beta_0}{1-\beta_0^2} \text{ and } E(Y_t, Y_{t-k}) = \frac{\sigma_\nu^2}{1-\beta^2}|\beta|^k \tag{8}$$

In the context of beta regression models, let us assume that the mean follows the model $g(\nu) = h(t, x_t, \boldsymbol{\beta}) + \eta_t$ where h is an appropriate real function, x_t is a vector of explanatory variables values, at time t, and η_t is a random variable, that follows a stochastic stationary process. At this point, it is assumed that η_t follows an autoregressive process, where,

$$\eta_t = \alpha_0 + \sum_{i=1}^{p} \beta_i \eta_{t-i} + \nu_t \tag{9}$$

and where ν_t follows a normal distribution, $\nu_t \sim N(0, \sigma^2)$. Thus, the beta autoregressive models are defined as follows: we assume that Y_t has a beta distribution, that is, let $Y_t \sim Beta(\mu_t, \phi)$, $t = 1, 2, \ldots$, where $\mu_t = E(Y_t|Y_{t-k})$ and $Var(Y_t|Y_{t-k}) = \mu_t/\phi$. The random process Y_t follows a beta regression process if

$$g(\mu_t) = h(t, x_t, \boldsymbol{\beta}) + \alpha_0 + \sum_{i=1}^{p} \alpha_i \eta_{t-i} + \nu_t \tag{10}$$

with $\eta_t = g(Y_{t-1})$. In this definition, if the model follows the restriction given by (7) , we say that the beta stochastic process is stationary.

4 The proposed model

For the modeling of meteorological data, with values in the open interval $(0, 1)$, we assume that the observations of the interest are generated from a conditional beta probability distribution function, with conditional means $\mu_t = \mathrm{E}(Y_t|H_{t-1})$ and conditional precision parameters given respectively by,

$$\mu_t = \alpha_0 + \sum_{i=1}^{q}(\alpha_{1i}\cos(2\pi f_i t) + \alpha_{2i}\sin(2\pi f_i t)) + \sum_{i=1}^{p}\phi_i y_{t-i} \tag{11}$$

$$\log(\phi_t) = \lambda_0 + \sum_{i=1}^{s}(\lambda_{1i}\cos(2\pi f_i t) + \lambda_{2i}\sin(2\pi f_i t)) + \sum_{i=1}^{r}\theta_i y_{t-i} \tag{12}$$

In these models, $\boldsymbol{\beta} = \{\phi_0, \phi_1, \ldots, \phi_p, \alpha_{11}, \ldots, \alpha_{1q}, \alpha_{21}, \ldots, \alpha_{1q}\}$ is the vector of the mean parameters model and $\gamma = \{\theta_0, \theta_1, \ldots, \theta_r, \lambda_{11}, \ldots, \lambda_{1s}, \lambda_{21}, \ldots, \lambda_{1s}\}$ is the vector of the variance parameters model, respectively. The parameters are estimated using a Bayesian aproach.

Special models can be proposed easily from this general model. As first one, we can assume a seasonal model, with mean given by (11) and autoregressive precision without seasonal terms. As second one,we can assume a seasonal mean model, with mean given by (11) and seasonal precision without autoregressive terms. As a third one, we can assume a model with an autoregressive mean and a variance model, without seasonal terms in the mean and in the precision. As a fourth one, we can assume an autoregressive model, with a constant precision parameter.

5 Rio Claro city air humidity time series analysis

5.1 A period model

In this section we introduce the seasonality present in the time series of monthly air humidity. Thus, after the detrains of the time series, a spectral analysis is

developed to determine the time periods to be considered in the mean and variance model formulation.

In the spectral analysis, if the number of observations is $T = 2q + 1$, the Fourier series, given by (13) is fitted. In this equation, $f_i = i/T$ is the ith harmonic of the fundamental frequency $1/T$ and, α_0, α_{1i} and α_{2i}, $i = 1, \ldots, q$, are the coefficients.

$$z_t = \alpha_0 + \sum_{i=1}^{q} (\alpha_{1i} \cos(2\pi f_i t) + \alpha_{2i} \sin(2\pi f_i t)) + e_i \tag{13}$$

The least square estimates of the coefficients α_0 and $(\alpha_{1i}, \beta_{2i})$, $i = 1, 2, \ldots q$, are given by,

$$\widehat{\alpha}_0 = \frac{1}{T} \sum_{t=1}^{T} z_t, \tag{14}$$

$$\widehat{\alpha}_i = \frac{2}{T} \sum_{t=1}^{T} z_t \cos(2\pi f_i t), \tag{15}$$

$$\widehat{\beta}_i = \frac{2}{T} \sum_{t=1}^{T} z_t \sin(2\pi f_i t). \tag{16}$$

The periodogram then consists of the $q = (T-1)/2$ values of the intensities at frequencies f_i, $i = 1, 2, \ldots q$, given by

$$I(f_i) = \frac{T}{2} (\widehat{\alpha}_i^2 + \widehat{\beta}_i^2) \tag{17}$$

When T is even, we set $T = 2q$ and (14), (15), (16) and (17) can be applied for $i = 1, 2, \ldots, (q-1)$, but we need to estimate a_q, b_q and $I(f_q)$ by,

$$\widehat{\alpha}_q = \frac{1}{T} \sum_{t=1}^{T} (-1)^t z_t, \tag{18}$$

$$\widehat{\beta}_q = 0, \tag{19}$$

$$I(f_q) = T\widehat{\alpha}_q^2. \tag{20}$$

$$\tag{21}$$

Observe that the highest frequency is given by 0.5 cycle per day (time interval) since the smallest period is 2 intervals. Figure 3 shows the periodogram

for the frequencies f_i, $i = 1, 2, \ldots, n$. In this figure we observe that the harmonics of higher intensities correspond to the cycle $(1/f_i)$ of 26 and 52 days. The coefficients of these harmonics are presented in Table 1.

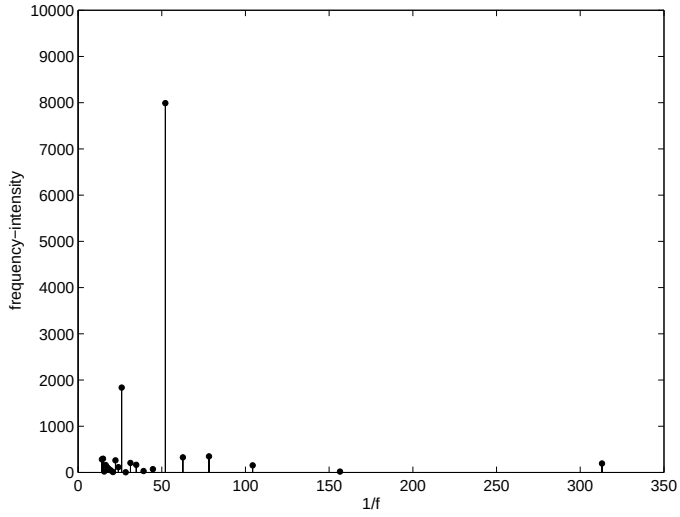


Figure 2: Periodogram of the air humidity time series

Table 1: Ranges of sines and cosines of two harmonics of the mean of air humidity time series.

i	$1/f_i$	$\widehat{\alpha}_i$	$\widehat{\beta}_i$
0	-	68.7664	-
6	52	-6.4829	3.0063
12	26	-2.1038	2.7019

The mean equation of the air humidity time series is given by,

$$\overline{z}_t = \alpha_0 + \alpha_6 \cos\left(\frac{2\pi}{52}t\right) + \beta_6 \sin\left(\frac{2\pi}{52}t\right) + \alpha_{12} \cos\left(\frac{2\pi}{26}t\right) + \beta_{12} \sin\left(\frac{2\pi}{26}t\right) \quad (22)$$

5.2 A Seasonal mean autoregressive model

A beta seasonal model with an order one for the autoregressive term has been fitted to the air humidity data. In this way, the mean and dispersion models are given respectively, by equations (23) and (24), respectively, where $e_i \sim N(0, \tau_e^2)$.

In this application, independent normal prior distributions N(0, 10k), with k = 2, was assumed for the parameters of the models. The initial values were considered to be equal to zero, except for β_0 and γ_0. For the parameter γ_0 a negative values for the initial values, for example -4 is a good choice. For τ_0^2 we assume a gamma distribution ($G(0.001, 0.001)$).

$$\begin{aligned} \text{logit}(\mu_i) &= \beta_0 + \beta_1 \cos(2\pi t_i/52) + \beta_2 \sin(2\pi t_i/52) \\ &\quad + \beta_3 \cos(2\pi t_i/26) + \beta_4 \sin(2\pi t_i/26) + \beta_5 \text{logit}(Y_{i-1}) \qquad (23) \\ \phi_i &= \exp(\gamma_0 + e_i) \qquad (24) \end{aligned}$$

The parameter estimates (Monte Carlo estimates of the posterior means) for the parameters obtained using a simulated sample of the joint posterior distribution for the parameters of the models using the WinBugs software(Spielgelhalter et al., 2003) for the assumed model (23) and their respective standard deviations are given in Table (2). Let us denote this model as "Model I". Convergence of the Gibbs sampling simulation algorithm was observed from standard trace plots for simulated samples for each parameter. For this model the logarithm of the likelihood function evaluated at the obtained estimates for the parameters of interest is given by $2logL = 834.852$ and the respective DIC value by -820.766. The estimate for the variance of the random effect is given by $\hat{\sigma}_e^2 = 0.020(0.054)$.

Table 2: Beta mean seasonal regression model estimates.

Parameter	β_0	β_1	β_2	β_3	β_4	β_5	γ_0
Mod. I	0.5871 (0.0469)	−0.2084 (0.0296)	0.1029 (0.0256)	−0.0647 (0.0250)	0.0864 (0.0253)	0.2675 (0.0535)	3.919 (0.0842)
Mod. II	−0.1175 (0.1805)	−0.2002 (0.0310)	0.1027 (0.0246)	−0.0615 (0.0251)	0.0837 (0.0249)	1.342 (0.2615)	3.935 (0.0876)

As a second model, denoted as "model II" we assume a model similar to the model defined by (23)and (24), but with the original variables Y_{t-1} instead of the transformation $logit(Y_{t-1})$. The estimates of the mean parameters of "Model II", are also given in Table 2. The posterior mean and standard deviation of the variance for the random effects are given by $\sigma_e = 0.04979(0.07365)$. For

this model, (see Table 2) the logarithm of the likelihood function evaluated at the obtained estimates for the parameters of the model is given by $2log(L) = 841.451$ and the DIC value is given by -821.419. From the DIC values we can see that the "Model II" is better fitted by the data than the "Model I", given that its DIC value is the smallest one.

6 Seasonal mean and dispersion models

In this section, double seasonal beta regression models are proposed to analyze the air humidity data. Thus, beta regression models with mean model given by equation (23) and dispersion model given by

$$\log(\phi_i) = \gamma_0 + \gamma_1 \cos(2\pi t_i/52) + \gamma_2 \sin(\pi/52) + \gamma_3 \sin(\pi/26) + e_i \qquad (25)$$

are considered, where $e_i \sim N(0, \sigma_e^2)$. Let us denote this model as "Model III" . Assuming independent normal prior distributions as in the last section, samples of the posterior distribution were generated using the WinBugs software. The estimates of the mean parameters of "model III" are given in Table 3. The estimates of the precision parameters and their standard deviations are given in Table 4. For this model the logarithm of the likelihood function evaluated at the estimates of the parameters of interest is given by $2LogL = 846.669$ and the DIC criterion gives the value -824.450.

Table 3: Double seasonal models: mean parameter estimates.

Parameter	β_0	β_1	β_2	β_3	β_4	β_5
Model-III	0.5847 (0.04677)	−0.2068 (0.03)	0.09414 (0.0255)	−0.06676 (0.02576)	0.08844 (0.02487)	0.2708 (0.05421)
Model-IV	−0.09075 (0.1816)	−0.2042 (0.03064)	0.09492 (0.02541)	−0.0643 (0.02551)	0.08558 (0.02499)	1.304 (0.2635)

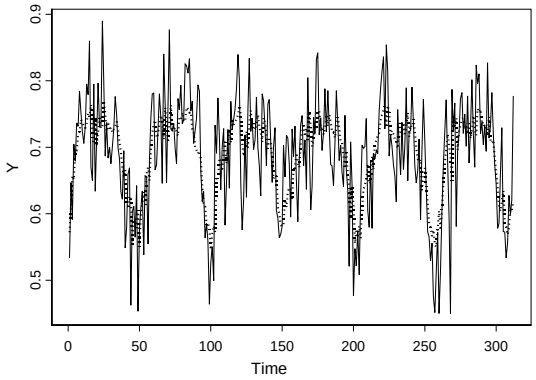
Another model denoted as "Model IV" is considered assuming the mean and dispersion beta regression model, with mean given by (23) and precision given by (25), but with Y_{t-1} instead of $\text{logit}(Y_{t-1})$. For this model, the parameter estimates of the mean model also are given Table 3. The parameter estimates

for the precision parameter model considering, denoted as "Model-VI" also are given in Table 4. The logarithm of the likelihood function evaluated at their parameter estimates is given by $logL = 847.460$ and the DIC criterion gives the value -824.520.

Table 4: Beta mean seasonal models: precision parameter estimates.

Mod	γ_0	γ_1	γ_2	γ_3	σ_e^2
Model III	3.937 (0.08246)	−0.1127 (0.1129)	−0.2812 (0.1162)	0.1819 (0.1165)	0.008216 (0.01724)
Model IV	3.945 (0.0849)	−0.1082 (0.1153)	−0.2769 (0.1172)	0.1761 (0.1157)	0.02926 (0.06146)

(a)



(b)

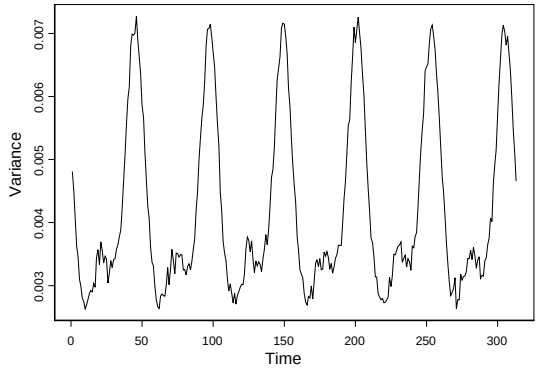


Figure 3: (a) Air humidity series data (continuous line) and fit mean (diseased line) for double seasonal model. (b) Fitted variance time series.

The results of this statistical analysis show that the models that include seasonal modeling in the mean and precision are the best ones, given that their DIC values are smaller than the DIC values of the models without seasonal terms in the precision model. Figure 3(a) is a good graphical illustration of the agreement between the data and the fitted mean, showing the performance of the proposed models. From Figure 3(b), obtained using the equation $\sigma^2 = \frac{\mu(1-\mu)}{\phi+1}$, we can observe the behavior of the fitted variance.

7 Conclusions

In this paper, new beta regression models are proposed, to model meteorological data series as levels of rainfall, temperature or levels of air humidity. In these new modeling approaches, both the mean and precision parameters were modeled, which gives a great flexibility of fit for the data, as it was observed in the data application considering the air humidity time series. The Bayesian methodology used to find posterior estimates of the parameters showed good performance to analyze the proposed data set. The use of available, gives a great computational simplification in the analysis of the data. It is important to point out that the use of a Bayesian approach also permits the use of existing prior opinion, usualy avaliable in the meteorological time series. These results could be of great interest for statisticians and environmental researchers working with meteorological data series.

References

[1] Abu-Taleb, A. A., Alawneh, A. J., and Smadi, M. M. (2007) Statistical analysis of recent changes in relative humidity in Jordan. *American Journal of Environmental Sciences*, **3**(2), 75û77.

[2] Andrade, M., Cepeda-Cuervo, E., and Achcar, J. (2012). A seasonal and heteroscedastic gamma model for hydrological time series: A Bayesian approach. *AIP Conf. Proc.*, 1490, pp. 97-107

[3] Barros, A. P., and Lang, T. J. (2003). Monitoring the Monsoon in the Himalayas: Observations in Central Nepal. *Montly Weather Review*, **131**(7), 1408-1427.

[4] Branscum A. J., Johnson W. O., Thurmond M. C. (2007). Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian and New Zealand Journal of Statistics*, **49**, 287-301.

[5] Cepeda, E.C. (2001). Variability Modeling in Generalized Linear Models, *Unpublished Ph.D. Thesis. Mathematics Institute, Universidade Federal do Rio de Janeiro.*

[6] Cepeda-Cuervo, E. and Achcar J. A. (2010). Heteroscedastic Nonlinear Regression Models. *Communications in Statistics - Simulation and Computation*, **39**(2):405-419

[7] Cuervo, E. C., Andrade, M. G., and Achcar, J. A. (2012, October). A seasonal and heteroscedastic gamma model for hydrological time series: A Bayesian approach. *In AIP Conference Proceedings* (Vol. **1490**, p. 97).

[8] Chiawa, M.A., Asare B.K. and Audu B. (2010). Short and Long Memory Time Series Models of Relative Humidity of Jos Metropolis, *Research Journal of Mathematics and Statistics*, **2**(1), 23-31, 2010.

[9] Evan, A. T., and Camargo, S. J. (2011). A Climatology of Arabian Sea Cyclonic Storms. *Journal of Climate*, **24**(1), 140-158.

[10] Ferrari, S., Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions, *Journal of Applied Statistics*, **31**, 799-815.

[11] Guimarães, R., and Santos, E. G., (2011). Principles of stochastic generation of hydrologic time series for reservoir planning and design: A case study. *Journal of Hydrologic Engineering.* **16** (11), 891-898.

[12] Jones, R. H., (1964). Spectral analysis and linear prediction of meteorological time series, *J. Appl. Meteor.*, **3**, 45-52.

[13] Katz, R. W., and Skaggs, R. H., (1981). On the Use of Autoregressive-Moving Average Processes to Model Meteorological Time Series. *Mon. Wea. Rev.*, **109**, 479û484.

[14] Lee, J. Y., and Lee, K. K., (2000). Use of hydrologic time series data for identification of rechargemechanism in a fractured bedrock aquifer system. *Journal of Hydrology.* **229**, 190û201.

[15] Sarraf, A., Vahdat,S. F. and Behbahaninia, A., (2011). Relative Humidity and Mean Monthly Temperature Forecasts in Ahwaz Station with ARIMA Model in time Series Analysis, *International Conference on Environment and Industrial Innovation*, IPCBEE vol.12 IACSIT Press, Singapore.

[16] Simas A. B., Barreto-Souza W., and Rocha A. V. (2010). Improved Estimators for a General Class of Beta Regression Models, *Computational Statistics & Data Analysis*, **54**(2), 348-366.

[17] Smithson M.,and Verkuilen J. (2006). A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables. *Psychological Methods*, **11**(1),54-71.

[18] Spiegelhalter, D. J., Thomas, A., Best N. G., Gilks, W. R. (2003). WinBUGS User Manual (version 1.4). MRC Biostatistics Unit, Cambridge, U.K.