TESTE DE ADERÊNCIA E PODER DO TESTE PELA SIMULAÇÃO DE MONTE CARLO: TESTE DE DISTRIBUIÇÃO NOS DADOS CLIMÁTICOS DO BRASIL

Leonardo R. O. Merelles
Pontifícia Universidade Católica de Goiás – MEPROS
Avenida Universitária, nº 1.440, Setor Universitário, Goiânia-GO
merellesleonardo@gmail.com

Maria J. P. Dantas
Pontifícia Universidade Católica de Goiás – MEPROS
Avenida Universitária, nº 1.440, Setor Universitário, Goiânia-GO
mjpdantas@gmail.com

José E. Menezes
Pontifícia Universidade Católica de Goiás – MEPROS
Avenida Universitária, nº 1.440, Setor Universitário, Goiânia-GO
jelmo.maf@gmail.com

Viviane S. Dias
Pontifícia Universidade Católica de Goiás – MEPROS
Avenida Universitária, nº 1.440, Setor Universitário, Goiânia-GO
engvivianedias@gmail.com

RESUMO

A qualidade do ajuste dos dados de entrada em um modelo é importante para sustentar a performance dos resultados. Há vários testes para determinar a aderência dos dados a uma distribuição específica. Foi proposto um algoritmo para verificar se uma amostra aleatória pertence a uma determinada distribuição aplicando a simulação de Monte Carlo. O *p-value* simulado e o poder do teste foram determinados através do teste de Kolmogorov-Smirnov e pela estatística-*t*, nesta ordem. O desempenho do algoritmo foi avaliado com dados climáticos do Brasil. A temperatura média e a umidade relativa do ar possuem distribuição logística. A Simulação de Monte Carlo pode ser aplicada para determinar a aderência de uma amostra e a probabilidade de cometer o erro tipo II na escolha dos parâmetros.

PALAVRAS CHAVE. Bondade do ajuste, Teste hipótese, Simulação de Monte Carlo.
SIM - Simulação / EST - Estatística

ABSTRACT

The goodness-of-fit of the input data in a model is important to sustain the performance of the results. There are several tests to determine the adherence of the data to a specific distribution. An algorithm was proposed to verify if a random sample belongs to a certain distribution applying the Monte Carlo simulation. The simulated p-value and power of the test were determined by the Kolmogorov-Smirnov test and by the t-statistic, in that order. The performance of the algorithm was evaluated using climatic data from Brazil. The mean temperature and relative humidity of the air is logistically distributed. The Monte Carlo simulation can be applied to determine the adhesion of a sample and the probability of making the type II error in the choice of parameters.

KEYWORDS. Goodness-of-fit, Hypothesis test, Monte Carlo Simulation.
SIM - Simulation / EST - Statistic

1. Introdução

A Simulação de Monte Carlo (SMC) é definida como um método que emprega variáveis aleatórias de uma determinada distribuição para resolver problemas determinísticos ou estocásticos. Esta abordagem foi originada durante a Segunda Guerra Mundial no desenvolvimento da bomba atômica [Law, 2007].

Através da simulação podem ser apresentadas diretrizes que conduzam a modelagem de um processo e assim, auxiliar na tomada de decisão [Bertrand e Fransoo, 2002]. Porém, com frequência o método de simulação é questionado devido a problemas que podem ocorrer durante a pesquisa. Para Banks e Chwif [2011] os erros podem ocorrer na coleta dos dados, construção do modelo, verificação e validação, análise dos resultados e em outros pontos.

Na fase de análise de dados, tem-se o objetivo de testar se uma distribuição pertence a uma determinada família paramétrica. Para isso, Pearson introduziu no início do século XX o termo *Goodness-of-Fit* (GoF) que verifica a discrepância entre a amostra aleatória e a função teórica [González-Manteiga e Crujeiras, 2013]. Desde então, GoF vem sendo aplicado em modelo de regressão, análise multivariada de dados [Henseler e Sarstedt, 2013], padrões espaciais [Dao e Genton, 2014] e outros.

Há dois possíveis testes, o paramétrico e não-paramétrico, para dados contínuos e discretos. O teste de ajuste paramétrico mais bem conhecido é o qui-quadrado χ^2, enquanto o teste não-paramétrico mais popular é o teste proposto por Kolmogorov e Smirnov (KS) [Arnold e Emerson, 2011]. Entretanto, é importante corrigir a probabilidade de ocorrer o erro tipo II, que pode diminuir a potência do teste. Assim, a SMC pode ser aplicada para auxiliar estes testes [Micheaux e Tran, 2016].

Para realizar simulações com dados dependentes de variáveis climáticas, faz necessário conhecer suas distribuições e seus parâmetros. Por este motivo se adota métodos computacionais para aplicação dos testes [R-Team, 2017]. Entretanto não é fácil trabalhar com dados climáticos [Ferrari e Ozaki, 2014; Vicente-Serrano et al., 2010]. Com isto, o objetivo deste trabalho é aplicar o teste KS em variáveis climáticas para testar a sua aderência e o poder do teste, aplicando a SMC no *software* R.

Este artigo está estruturado da seguinte forma: seção 2, teste de hipótese; seção 3, simulação de Monte Carlo no *software* R; seção 4, teste de aderência e poder do teste simulado; seção 5, teste de aderência com poder do teste simulado em dados climáticos; e por último, são apresentas as conclusões.

2. Teste de hipótese

A ideia do GoF se limita a comparar um estimador piloto não paramétrico de distribuição F ou densidade f de uma variável aleatória X, com um estimador da função objetivo. O estimador para teste, pode ser gerado por uma função de distribuição cumulativa empírica para função F e por um estimador de Kernel para densidade f [Bickel e Rosenblatt, 1973; Durbin, 1973]. Para mais detalhes da função de densidade de Kernel pode ser consultada a obra de Rosenblatt [1956] e Parzen [1962].

Dada uma amostra $\{X_1, X_2, \ldots, X_n\}$ de uma variável aleatória X, o objetivo é testar a seguinte hipótese para função de distribuição:

$$H_o\colon F \in F_{dist} = \{F_{\theta_o}\}, \theta_o \in \Theta \subset R^q, versus \;\; H_a\colon F \notin F_{dist},$$

para função de densidade:

$$H_o\colon F \in f_{dens} = \{f_{\theta_o}\}, \theta_o \in \Theta \subset R^q, versus \;\; H_a\colon f \notin f_{dens},$$

onde θ_o é o vetor de parâmetro da distribuição e deve ser levado em consideração que esta função existe. A veracidade da hipótese H_o é questionada, a um nível de significância α, então, aplica-se a estatística do teste $T_n = T(X_1, \ldots, X_n)$ para medir a discrepância dos estimadores sobre a hipótese nula H_o. Para a distribuição, a estatística do teste pode ser escrita como:

$$T_n = T(F_n, F_{\widehat{\theta}}) \equiv T(\alpha_n) \tag{1}$$

em que $F_n(x) = n^{-1}\{j; X_j \leq x\}$ é a função de distribuição acumulativa empírica e $F_{\widehat{\theta}}$ é o parâmetro estimado sobre a H_o. $\widehat{\theta}$ é o estimador de $\theta_o \in \Theta$. A estatística do teste para função de densidade pode ser escrita com:

$$T_n = T(f_{nh}, E_{\widehat{\theta}}(f_{nh})) \equiv T(\tilde{\alpha_n}) \tag{2}$$

em que $f_{nh}(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i)$ é a distribuição de Kernel e h é o parâmetro de suavização; K_h é o redimensionamento de Kernel; e $E_{\widehat{\theta}}(f_{nh})$ é a estimativa de densidade não-paramétrica.

Observado T_n, confrontando-o com um (C_α) ou dois ($C_L\alpha$ e $C_R\alpha$) valores críticos, que são as regiões de rejeição de H_o, definida como R_α, então se $T_n \in R_\alpha$ a hipótese nula é rejeitada. Também pode ser considerado a R_α com os casos a seguir $\{T_n > C_\alpha\}$, $\{T_n < C_\alpha\}$ ou $\{T_n < C_L\alpha/2\} \cup \{T_n > C_R\alpha/2\}$ [Micheaux e Tran, 2016].

Há dois tipos de erros associados a este teste, a rejeição da hipótese nula quando ela é verdadeira, erro tipo I, e a não rejeição da hipótese nula quando ela é falsa, erro tipo II [Walpole et al., 2014]. A probabilidade de cometer o erro tipo I pode ser controlada e é determinada por α. Com o valor crítico definido, a probabilidade da região de rejeição da hipótese nula é determinada pela $P_{Ho}[T_n \in R_\alpha] \leq \alpha$. Quanto ao erro tipo II para estatística do teste T_n pode ser determinado pelo poder do teste definido por $(1 - \beta_\alpha)$, com $\theta \in \Theta$, e diferente de θ_o, então a hipótese alternativa é a $P_{Ha}[T_n \in R_\alpha] = (1 - \beta_\alpha)$, em que, β_α é a probabilidade de ocorrer o erro tipo II.

2.1. Teste de Kolmogorov-Smirnov

O teste Kolmogorov-Smirnov é baseado na mensuração da discrepância vertical entre as funções de distribuição cumulativa $F_n(x)$ hipotética e a distribuição empírica $\widehat{F}(x)$ dos dados observados [Razali e Wah, 2011], definido por:

$$D_n = \sup_x\{|F_n(x) - \widehat{F}(x)|\} \tag{3}$$

em que $F_n(x)$ pode ser continua ou discreta; sup indica maior; e D_n não depende da distribuição hipotética.

A estatística do teste KS se destina a testar:

H_o: $\widehat{F}(x) = F_n(x)$, x de $-\infty$ a ∞, versus H_a: $\widehat{F}(x) \neq F_n(x)$, para pelo menos um x. Se D_n excede $1 - \alpha$ quartis, então, rejeita-se H_o ao nível de significância α.

3. Simulação de Monte Carlo no software R

O *software* R permite realizar tarefas como: aplicar funções; elaborar pacotes estatísticos; desenvolver rotinas; aplicar métodos de simulação; e realizar testes estatísticos [Crawley, 2007; R-Team, 2017]. Assim, dada uma amostra aleatória, é comum testar qual distribuição melhor se adere. Em geral, nesse processo faz necessário determinar a distribuição, estimar o vetor de parâmetros e a qualidade da avaliação [Delignette-Muller e Dutang, 2015]. O pacote MASS [Venables e Ripley, 2013] e tdistrplus [Delignette-Muller et al., 2014] auxiliam a determinar o vetor de parâmetros para amostra uni-variada. O ks.test determina o valor da estatística do teste de Kolmogorov-Smirnov [Arnold e Emerson, 2011].

O teste KS com o poder do teste podem ser avaliados pela SMC para verificar se uma amostra aleatória iid de n observações vieram de uma distribuição F_Θ, com nível de significância α. A Simulação de Monte Carlo é apresentada no Algoritmo 1. Na aplicação do poder do teste, a região crítica pode ser determinada fazendo $rc = \{x\}$, em que $x = 0$ é o teste bilateral da região crítica igual a $\{T_n < C_L\alpha\} \cup \{T_n > C_R\alpha\}$, $x = 1$ é o teste unilateral para região crítica igual a $\{T_n < C_\alpha\}$, e $x = 2$ é o teste unilateral para região crítica igual a $\{T_n > C_\alpha\}$. O poder do teste

Algorithm 1 $Test_{ad}$ *simulado* (X, DF, nreps, alpha, rc)

Input:

X ▷ $\{X_1, X_2, \ldots, X_n\}$
DF ▷ normal, log-normal, exponencial, weibull, gamma, beta, uniforme, logística)
$nreps$ ▷ número de replicações
$alpha$ ▷ probabilidade do erro tipo I
rc ▷ região critica {0, 1 e 2}

Output:

$test_{ad}$ ▷ teste de aderência simulado

1: **function** PARAM(X, DF)
2: $\theta_o \leftarrow fitdist(X, distr = \text{“}DF\text{”})$ ▷ [Delignette-Muller et al., 2014]
3: **return** (θ_o)
4: **end function**

5: **function** POWER_TEST(X, alpha, nreps, rc, dist)
6: $mu \leftarrow mean(X)$ ▷ X é uma matriz de n linhas e m colunas
7: $sd \leftarrow standard\ deviation(X)$
8: $dist \leftarrow mu \cdot dist$
9: **for** $(m = 1, \ldots, length(dist))$ **do**
10: $t_{nreps} \leftarrow mu - mean(x_{1,n}, \ldots, x_{nreps,n}) - dist[m]$
11: $t_{nreps} \leftarrow t_{nreps} / (\sqrt{sd^2 + standard - deviation(x_{1,n}, \ldots, x_{nreps,n})^2)/n}$
12: **end for**
13: $\beta_\alpha \leftarrow t_{nreps} \in rc$ ▷ rc é a região critica R_α
14: **return** (β_α)
15: **end function**

16: **function** KS_SIM(X, DF, nreps, alpha, rc)
17: $\theta_o \leftarrow$ **goto** Param(X,DF)
18: $D_n \leftarrow p_{value} \leftarrow ks.test(X, DF, \theta_o)$ ▷ [Arnold e Emerson, 2011]
19: $X_{sim} \leftarrow rng_{F_\theta}(x_{1,1}, \cdots, x_{n,nreps})$ ▷ gerador de números aleatórios
20: $p_{test} \leftarrow$ **goto** Power_test(X_{sim}, alpha, nreps, rc, dist)
21: **for** $(m = 1, \ldots, nreps)$ **do**
22: $D_{sim} \leftarrow ks.test(X_{sim_{nreps,n}}, DF, \theta_o)$
23: **end for**
24: $p_{sim} \leftarrow sum(D_{sim} > D_n + 1) / (nreps + 1)$
25: **return** $(D_n, D_{sim}, p_{value}, p_{sim})$
26: **end function**

27: **for** $(i = 1, \ldots, length(DF))$ **do**
28: $test_{ad} \leftarrow$ **goto** ks_sim(X, DF[i], nreps, aplha,rc)
29: **end for**
30: **return** $(test_{ad})$

é calculado de acordo o valor do vetor t_{nreps} na sub-rotina power_test na linha (5). O erro tipo II pode ser testado a distâncias determinadas pelo vetor *dist*(0.01, 0.05, 0.10, 0.15, 0.20, 0.25), que pode ser facilmente configurado para testar outras distâncias.

4. Teste de aderência e poder do teste simulado

Inicialmente, com objetivo de manter a replicação, foi aplicado o algoritmo de Wichmann e Hill [1982] para gerar números pseudoaleatórios. Este algoritmo gera um período de $7\text{x}10^{12}$ variáveis sem repetição. A partir deste ponto, todas as amostras aleatórias geradas para teste foram iniciadas com a semente *set.seed*(123, kind = "Wichmann-Hill"). Foram gerados 200 números aleatórios para as distribuições normal $N(5,1)$, uniforme $U(0,2)$ e exponencial $EXP(0,5)$. Após a geração dos números aleatórios, eles são somados de forma a resultar em uma única amostra. Para compreender o comportamento desta amostra foi plotado o histograma dos dados com a função de densidade, densidade acumulada, Q-Q plot e P-P plot. A análise da amostra aleatória está na Figura 1.

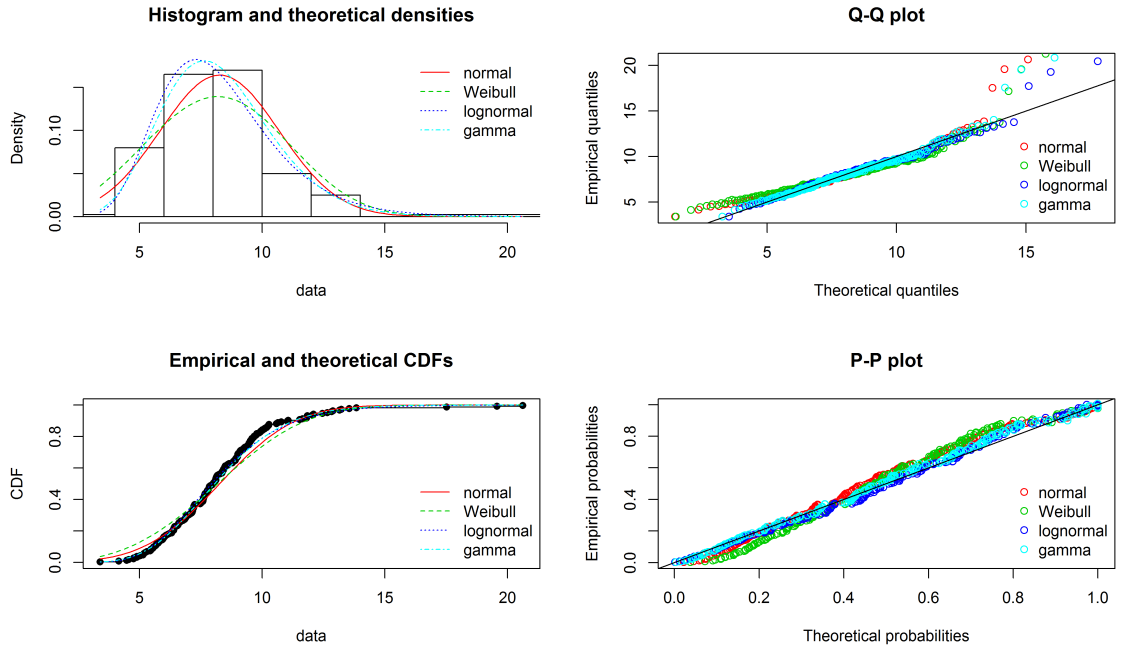

Figura 1: Análise de dados aleatórios para distribuições normal, weibull, log-normal e gamma

Em seguida, foi calculada a estatística do teste KS tradicional e o *p-value* simulado para determinar a distribuição que mais adere a amostra gerada. Esta análise foi realizada com $n = 200$ (tamanho da amostra aleatória) e $M = 2.000$ replicações. O resultado do teste de aderência simulado está na Tabela 1. Nota-se que o valor de *p-value* estatístico e *p-sim* simulado tem valores aproximados. A distribuição beta retornou NA, o que era esperado, pois ela pertence a $0 > x < 1$.

Tabela 1: Teste de aderência Kolmogorov-Smirnov

Distribuição	D	*p-value*	*p-sim*
Normal	0,10934	0,01675	0,01549
Log-normal	0,05863	0,49748	0,46926
Uniforme	0,46573	0,00000	0,00049
Exponencial	0,41432	0,00000	0.00049
Logística	0,07228	0,24691	0,22338
Beta	NA	NA	NA
Gamma	0,07297	0,23720	0,22038
Weilbull	0,11152	0,01381	0,01449

Esta amostra testada pode pertencer a distribuição gamma (0,22 *p-sim*), logística (0,22 *p-sim*) e log-normal (0,46 *p-sim*), no entanto, a distribuição log-normal é a melhor. Assim, o poder do teste para distribuição log-normal foi determinado pela Simulação de Monte Carlo para verificar se esta distribuição pertence ao vetor Θ, utilizando o Algoritmo 1. Os dados de entrada foram: vetor de parâmetro θ_o (meanlog = 2,073 e sdlog = 0,274); $\alpha = 0,05$; $rc = 0$; M = 2.000 replicações; e $n = 10, ..., 400$. As n variáveis foram geradas pela mesma semente. A probabilidade do erro tipo II está na Tabela 2. O poder do teste pode ser determinado por $(1 - \beta_\alpha)$. Nota-se que quando aumentada a distância (d é a distância em porcentagem na média) e o tamanho da amostra a probabilidade de ocorrer o erro tipo II diminui.

Tabela 2: *P-value* teórico, simulado KS e a probabilidade do erro tipo II para dados log-normal

n	*p-sim*	*p-value*	α = 0,05 e rc = 0					
			d=0,01	d=0,05	d=0,10	d=0,15	d=0,20	d=0,25
10	0,938	0,945	0,998	0,985	0,926	0,754	0,473	0,192
20	0,952	0,945	0,996	0,982	0,887	0,604	0,262	0,054
30	0,663	0,669	0,992	0,963	0,764	0,358	0,520	0,001
50	0,402	0,373	0,995	0,954	0,636	0,194	0,009	0,001
75	0,854	0,867	0,992	0,871	0,278	0,008	0,000	0,000
100	0,728	0,759	0,991	0,813	0,128	0,000	0,000	0,000
150	0,570	0,579	0,992	0,757	0,063	0,000	0,000	0,000
200	0,563	0,587	0,991	0,604	0,005	0,000	0,000	0,000
300	0,395	0,397	0,984	0,428	0,000	0,000	0,000	0,000
400	0,649	0,649	0,979	0,239	0,000	0,000	0,000	0,000

5. Teste de aderência com poder do teste simulado em dados climáticos

Os dados climáticos diários foram acessados no banco de dados do Instituto Nacional de Meteorologia (INMET) e do *National Oceanic and Atmospheric Administration* (NOAA). Para baixar os dados no INMET foi realizada uma adaptação no algoritmo de Tatsch [2016] e para os dados do NOAA foi elaborado um código no R com as informações de Piccirilli [2015]. Foram acessadas todas as 265 Estações Meteorológicas (EMs) convencionais disponíveis no INMET e também, as 922 EMs no NOAA. No total foram analisadas 1.187 EMs no Brasil.

As variáveis analisadas foram temperatura média e umidade relativa do ar. Os dados de 2010 a 2016 foram selecionados para o teste de aderência. As séries históricas possuíam dados faltantes, entretanto, foi aplicado o método de Vicente-Serrano et al. [2010] e Ferrari e Ozaki [2014] para completar as falhas. Após completar os dados, restaram 263 EMs com dados completos.

Para aplicar o teste de aderência com poder do teste simulado, primeiro foram rotuladas as variáveis por mês (jan, ..., dez). Os dados foram rotulados porque podem pertencer a mesma

Tabela 3: Teste de aderência simulado para variáveis climáticas do Brasil

Mês	Normal	Log-normal	Exponencial	Weibull	Gamma	Beta	Uniforme	Logística
Temperatura média								
jan	0.4426	0.4221	0.0020	0.2285	0.4298	NA	0.0020	0.4750
fev	0.3331	0.3509	0.0020	0.3136	0.3411	NA	0.0060	0.4680
mar	0.5396	0.5030	0.0020	0.2938	0.5196	NA	0.0025	0.5537
abr	0.4582	0.3652	0.0020	0.4048	0.4006	NA	0.0022	0.5479
mai	0.3814	0.2818	0.0020	0.4527	0.3090	NA	0.0020	0.4980
jun	0.3530	0.3162	0.0020	0.3761	0.3255	NA	0.0020	0.5838
jul	0.3554	0.2832	0.0020	0.3240	0.3163	NA	0.0029	0.5102
ago	0.3189	0.2684	0.0020	0.2470	0.2910	NA	0.0020	0.4594
set	0.3871	0.3531	0.0020	0.2831	0.3748	NA	0.0021	0.5151
out	0.3626	0.3057	0.0020	0.3520	0.3363	NA	0.0021	0.5150
nov	0.4417	0.4490	0.0020	0.2034	0.4630	NA	0.0021	0.6184
dez	0.3887	0.3481	0.0020	0.2405	0.3730	NA	0.0023	0.4811
Média	0.3969	0.3539	0.0020	0.3100	0.3733	NA	0.0025	0.5188
Umidade relativa								
jan	0.2692	0.2363	0.0020	0.2643	0.2480	NA	0.0071	0.3398
fev	0.3272	0.2737	0.0020	0.3309	0.2977	NA	0.0023	0.4180
mar	0.3831	0.3885	0.0020	0.2297	0.3949	NA	0.0051	0.4294
abr	0.4090	0.4094	0.0020	0.2643	0.4338	NA	0.0032	0.4896
mai	0.3912	0.4199	0.0020	0.2035	0.4159	NA	0.0027	0.5150
jun	0.3849	0.4225	0.0020	0.1984	0.4286	NA	0.0020	0.5536
jul	0.3712	0.4255	0.0020	0.1707	0.4165	NA	0.0020	0.5644
ago	0.3590	0.4367	0.0020	0.1920	0.4195	NA	0.0021	0.5643
set	0.3230	0.4198	0.0020	0.1478	0.3916	NA	0.0026	0.4628
out	0.2452	0.2900	0.0020	0.1454	0.2795	NA	0.0060	0.3463
nov	0.2758	0.2484	0.0020	0.2663	0.2574	NA	0.0105	0.4215
dez	0.2581	0.2361	0.0020	0.2349	0.2424	NA	0.0128	0.3441
Média	0.3331	0.3506	0.0020	0.2207	0.3521	NA	0.0049	0.4541

distribuição e possuir parâmetros θ diferentes ao longo dos meses. Na sequência, foi determinado o *p-sim* para todas as EMs. O *p-sim* médio para as variáveis temperatura média e umidade relativa do ar está na Tabela 3. Nota-se que os dados melhor aderiram a distribuição logística e não há hipótese dos dados pertencerem a distribuição exponencial, beta e uniforme.

Sendo os dados da distribuição logística, a probabilidade de ocorrer o erro tipo II com os parâmetros θ_o pode ser testado com parâmetros Θ. O erro tipo II simulado está na Tabela 4 e o poder do teste pode ser determinado por $(1 - \beta_\alpha)$. Observa-se que em média, na distância de 5%, a probabilidade de ocorrer o erro tipo II bilateral é de 1,05 e 9,13% para temperatura média e umidade relativa do ar, respectivamente.

Tabela 4: Probabilidade de ocorrer o erro tipo II na distribuição logística

Mês	d=0,01	d=0,05	d=0,10	d=0,15	d=0,20	d=0,25
Temperatura média						
jan	0.6463	0.0000	0.0000	0.0000	0.0000	0.0000
fev	0.7006	0.0000	0.0000	0.0000	0.0000	0.0000
mar	0.6407	0.0001	0.0000	0.0000	0.0000	0.0000
abr	0.6247	0.0025	0.0000	0.0000	0.0000	0.0000
mai	0.5718	0.0061	0.0000	0.0000	0.0000	0.0000
jun	0.5314	0.0296	0.0001	0.0000	0.0000	0.0000
jul	0.5309	0.0363	0.0013	0.0000	0.0000	0.0000
ago	0.5693	0.0357	0.0009	0.0000	0.0000	0.0000
set	0.5985	0.0141	0.0000	0.0000	0.0000	0.0000
out	0.6142	0.0019	0.0000	0.0000	0.0000	0.0000
nov	0.6263	0.0002	0.0000	0.0000	0.0000	0.0000
dez	0.6141	0.0000	0.0000	0.0000	0.0000	0.0000
Média	0.6057	0.0105	0.0002	0.0000	0.0000	0.0000
Umidade relativa						
jan	0.8963	0.0864	0.0001	0.0000	0.0000	0.0000
fev	0.8927	0.0665	0.0001	0.0000	0.0000	0.0000
mar	0.8581	0.0440	0.0000	0.0000	0.0000	0.0000
abr	0.8445	0.0414	0.0000	0.0000	0.0000	0.0000
mai	0.8434	0.0136	0.0000	0.0000	0.0000	0.0000
jun	0.8356	0.0105	0.0000	0.0000	0.0000	0.0000
jul	0.8703	0.0276	0.0000	0.0000	0.0000	0.0000
ago	0.8786	0.0503	0.0000	0.0000	0.0000	0.0000
set	0.8935	0.2078	0.0158	0.0002	0.0000	0.0000
out	0.8966	0.2614	0.0225	0.0002	0.0000	0.0000
nov	0.8858	0.1588	0.0027	0.0000	0.0000	0.0000
dez	0.8917	0.1272	0.0015	0.0000	0.0000	0.0000
Média	0.8739	0.0913	0.0036	0.0000	0.0000	0.0000

6. Conclusões

O teste KS aplicando a Simulação de Monte Carlo retorna valores de *p-value* simulado próximos a função ks.test do software R. A SMC pode ser aplicada para determinar qual distribuição melhor se adere ao dado de uma amostra aleatória. Também é possível determinar a probabilidade do erro tipo II simulado. Os dados de temperatura média e umidade relativa do Brasil de 2010 a 2016 foram testados e melhor se aderiram a distribuição logística. Apesar de haver outras formas de estimar o *p-value* e a probabilidade do erro tipo II, a SMC apresentou excelente desempenho.

Em trabalhos futuros, as demais distribuições de probabilidade poderiam ser testadas, assim como outras variáveis climáticas.

Agradecimentos – Os autores agradecem a Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro para a realização deste estudo e de outras atividades do Mestrado em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica de Goiás.

Referências

Arnold, T. B. e Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2):34–39.

Banks, J. e Chwif, L. (2011). Warnings about simulation. *Journal of Simulation*, 5(4):279–291.

Bertrand, J. W. M. e Fransoo, J. C. (2002). Operations management research methodologies using quantitative modeling. *International Journal of Operations & Production Management*, 22(2): 241–264.

Bickel, P. J. e Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics*, 1(6):1071–1095.

Crawley, M. J. (2007). *The R book*. John Wiley & Sons.

Dao, N. A. e Genton, M. G. (2014). A monte carlo-adjusted goodness-of-fit test for parametric models describing spatial point patterns. *Journal of Computational and Graphical Statistics*, 23 (2):497–517.

Delignette-Muller, M. L. e Dutang, C. (2015). fitdistrplus: An r package for fitting distributions. *Journal of Statistical Software*, 64(4):1–34.

Delignette-Muller, M. L., Pouillot, R., Denis, J. B., e Dutang, C. (2014). fitdistrplus: help to fit of a parametric distribution to non-censored or censored data, 2010. *R package version 0.1–3*.

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics*, 1(2):279–290.

Ferrari, G. T. e Ozaki, V. (2014). Missing data imputation of climate datasets: Implications to modeling extreme drought events. *Revista Brasileira de Meteorologia*, 29(1):21–28.

González-Manteiga, W. e Crujeiras, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *Test*, 22(3):361–411.

Henseler, J. e Sarstedt, M. (2013). Goodness-of-fit indices for partial least squares path modeling. *Computational Statistics*, 28(2):1–16.

INMET (2017). Banco de dados meteorológicos para ensino e pesquisa. URL `http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep`. Instituto Nacional de Meteorologia.

Law, A. M. (2007). *Simulation modeling and analysis*. McGraw-Hill New York, 4 edition.

Micheaux, P. L. e Tran, V. A. (2016). Power: A reproducible research tool to ease monte carlo power simulation studies for goodness-of-fit tests in r. *Journal of Statistical Software*, 69(3): 1–42.

NOAA (2017). Data access - NOAA. URL https://www.ncdc.noaa.gov/data-access. National Oceanic and Atmospheric Administration.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Piccirilli, M. *R package to download, transform, analyze, and plot NOAA ISD weather data*, 2015. URL https://github.com/mpiccirilli/weatheR. R package version 0.0.1.

R-Team, C. (2017). R: A language and environment for statistical computing. vienna, austria: R foundation for statistical computing; 2014. r foundation for statistical computing.

Razali, N. M. e Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

Tatsch, J. *inmetr: A Package to Import Historical Data from Brazilian Meteorological Stations*. Santa Maria-RS, Brazil, 2016. URL https://github.com/jdtatsch/inmetr. R package version 0.0.1.

Venables, W. N. e Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 4 edition.

Vicente-Serrano, S. M., Beguería, S., López-Moreno, J. I., García-Vera, M. A., e Stepanek, P. (2010). A complete daily precipitation database for northeast spain: reconstruction, quality control, and homogeneity. *International Journal of Climatology*, 30(8):1146–1163.

Walpole, R. E., Myers, R. H., Myers, S. L., e Ye, K. (2014). *Probability and statistics for engineers and scientists*. Pearson Education, 9 edition.

Wichmann, B. A. e Hill, I. D. (1982). Algorithm as 183: An efficient and portable pseudo-random number generator. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(2): 188–190.