Trabalho apresentado no CNMAC, Gramado - RS, 2016.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Um procedimento de tratamento de *Missing Data* em dados agrometeorológicos

Lee Yun Sheng[1]
Universidade Federal de Mato Grosso, UFMT, Sinop, MT
Laurimar Gonçalves Vendrusculo[2]
Empresa Brasileira de Pesquisa Agropecuária, EMBRAPA, Sinop, MT

Resumo. Dados falhos ou omissos ("*missing data*") num banco de dados são prejudiciais para as análises e tomadas de decisão. Este trabalho tem como objetivo apresentar um procedimento baseado em análises estatísticas, modelos de regressão polinomial e logística, para tratamento dos dados falhos. Utilizou-se um banco de dados climáticos, cuja a estação se localiza no município de Água Boa, MT. As variáveis de estudo foram: temperatura mínima do ar, temperatura máxima do ar e precipitação acumulada no dia. Nos modelos apresentados neste estudo verificou-se uma alta correlação (maior que 0,62) e uma reprodução comportamental fidedigna as mesmas características dos dados reais. Portanto a metodologia se mostrou factível para os dados deste estudo bem como para serem incorporados em outros estudos agroclimatológicos.

Palavras-chave: Dados falhos ou omissos ("*missing data*"), regressão polinomial, regressão logística, dados climatológicos.

1 Introdução

A previsão do tempo, que é uma ciência milenar, ganhou novas dimensões com o advento computacional, principalmente para o entendimento das mudanças climáticas. Acurácia na previsibilidade climática tem grande influência no desenvolvimento econômico do país, principalmente no setor agrícola e de geração de energia, dentre outros [1, 5]. Todavia, desafios para a modelagem estatística baseada em séries temporais existem, pois condições climatológicas locais não podem ser simplesmente descritos por temperaturas globais. A necessidade de desenvolvimento destes modelos é crucial para a consequente tomada de decisão.

Devido a isto, a coleta de dados meteorológicos de qualidade para estudos científicos ganhou uma importância cada vez maior. Com o auxílio destes dados, pode-se por exemplo, realizar simulações de estudos numéricos em crescimentos de culturas, e estas usualmente requerem dados de precipitação acumulada diária, temperatura mínima do ar no dia, temperatura máxima do ar no dia, etc. Porém, estes possuem muitos campos vazios/falhos

[1]leeufmt@yahoo.com.br
[2]laurimar.vendrusculo@embrapa.br

DOI: 10.5540/03.2017.005.01.0404

("*missing data*"), o que se torna uma das maiores barreiras para estudos de maior exatidão gerando assim resultados de análises que diferem dos casos reais [7].

Assim, os estudos sobre o tratamento destes campos vazios se tornaram cada vez mais significativo, pois a acurácia desta informação agrega valor e alavanca as pesquisas na área de modelos hidrológicos e ambientais. Uma das formas de preenchimento para estes campos vazios é a utilização de médias de dados observados (normal climatológicas) ou dados sintéticos obtidos por geradores de dados. Vários autores apresentaram geradores de dados para simularem dados meteorológicos diários, os quais são amplamente utilizados [2–4, 6].

O objetivo deste trabalho é apresentar um procedimento baseado em análises estatísticas, modelos de regressão polinomial e logística, para tratamento dos dados falhos, minimizando assim a imprecisão das análises.

2 Materiais e métodos

Para compor a base de dados no desenvolvimento deste estudo, escolheu-se aleatoriamente uma estação meteorológica do estado de Mato Grosso disponível na base de dados fornecidos pelo site $< http://www.inmet.gov.br/ >$. Esta estação se localiza no município de Água Boa, cujos os dados compreendem o período de 7 de novembro de 2006 a 28 de outubro de 2014. Os atributos com dados diários,incluindo os faltantes ou vazios, tratados foram: temperatura mínima do ar, temperatura máxima do ar e precipitação acumulada no dia, denominadas aqui de variáveis de estudo. Aliado a cada um deste registro a variável data. Como a data é a única variável de dependência, sendo este um dado qualitativo, houve a necessidade de converte-la num dado numérico, isto é, utilizou-se o dia juliano respectivo da data e esta divide-se pela quantidade de dias do respectivo ano, exemplos: 1 de janeiro de 2010 corresponderia a 1/365 (isto é, o valor atribuído é 0,00274), e 12 de junho de 2012 corresponderia a 164/366, pois 2012 é um ano bissexto (logo o valor atribuído é 0,448087). Esta nova variável, a data convertida, será denominada de data transformada (DT).

Para viabilizar os métodos estatísticos optou-se por trabalhar somente com os campos preenchidos, isto é, todas as observações com campos vazios foram eliminados da amostra. Após esta etapa, as análises comportamentais dos dados restantes foram realizas por meio de: gráficos de dispersão, para avaliar a distribuição de pontos entre a data e as variáveis de estudo; e verificação da correlação de Pearson entre a data transformada e as variáveis de estudo. Utilizou-se nesta fase o software Excel 2010, para tanto se verificou a sazonalidade das variáveis de estudo.

Após estas análises, a amostra restante foi separada aleatoriamente 70% (setenta por cento) em massa de treinamento (MT), isto é a qual serão feito os modelos estatísticos e aplicações lógicos matemáticos, e o restante, 30% (trinta por cento), será a massa de validação (MV), isto é, após determinar os modelos calculados usando a massa de treinamento estas serão aplicadas nesta massa de validação para verificar a sua aderência aos dados reais.

Devido ao fato que a única variável dependente é a data transformada optou-se que os

DOI: 10.5540/03.2017.005.01.0404

modelos de inferências estatísticas escolhidas a serem aplicados nas temperaturas mínimas e máximas do ar são a regressão polinomial de grau dez, o que utiliza o método dos mínimos quadrados para a determinação da equação $y = \sum_{i=0}^{10} a_i \cdot x^i$, a qual y representa a variável dependente (a variável a ser estudada), a_i os coeficientes determinados pelo método e x a variável independente (a data transformada). Estes cálculos foram implementados utilizando scripts no software SAS software versão 9 em ambiente Windows.

Ao transformarmos a precipitação acumulada do dia em ocorrência de precipitação pode-se então inferir através de uma regressão logística em probabilidade de se incidir a chuva ou não, e através desta controlar uma função cosseinodal a qual os pontos máximos são o percentil 99 e o mínimo é 0. Consideramos que esta função cosseinoidal é distribuída de maneira aleatória a obedecer faixas de percentis para simular o efeito da precipitação, uma vez que esta é discreta e dispersa.

3 Resultados e Discussões

Observou-se claramente a sazonalidade (possibilitando assim aplicação de modelos de regressão) nas análises visuais, conforme os gráficos de dispersão da temperatura mínima e máxima do ar, e a precipitação ilustrados pelas figuras 1 a 3.

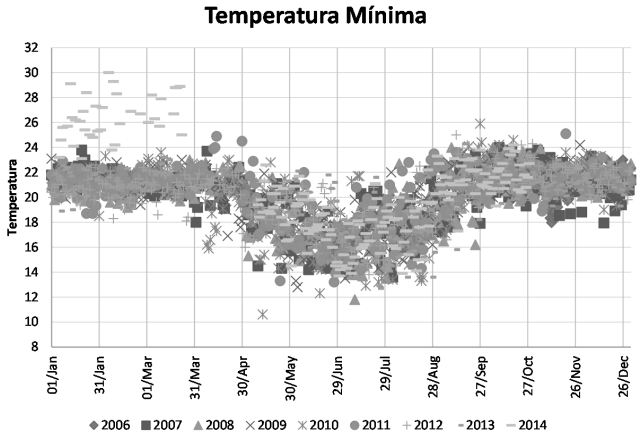

Figura 1: Temperatura mínima do ar no dia de cada ano em seus respectivos dias e mês do ano, sobrepostos.

Os modelos de regressão polinomial de grau dez dos dados, conforme supracitado, são:

$$\begin{aligned} TMIN \quad &= 21,35019 + DT \cdot 20,72923 + DT^2 \cdot (-343,36477) + DT^3 \cdot 2137,25779 + \\ &+ DT^4 \cdot (-4442,15762) + DT^5 \cdot (-6725,34896) + DT^6 \cdot 40753 + \\ &+ DT^7 \cdot (-59849) + DT^8 \cdot 32222 + DT^{10} \cdot (-3773,29278) \end{aligned} \tag{1}$$

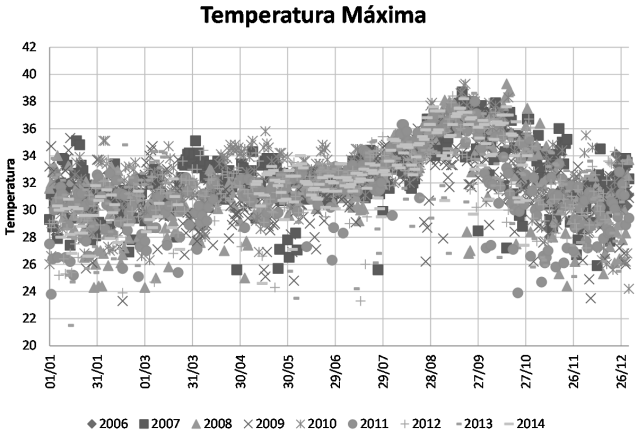

Figura 2: Temperatura máxima do ar no dia de cada ano em seus respectivos dias e mês do ano, sobrepostos.

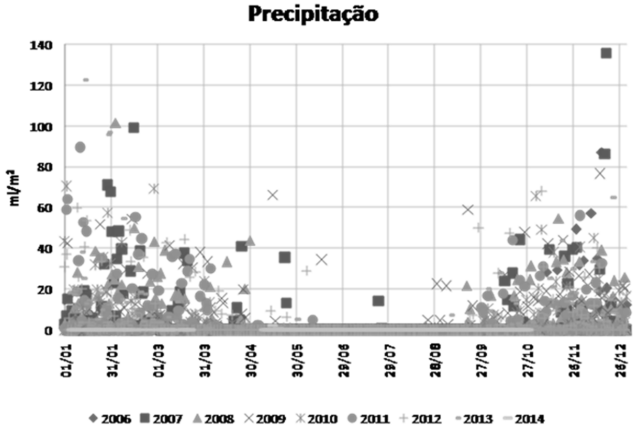

Figura 3: Precipitação no dia de cada ano em seus respectivos dias e mês do ano, sobrepostos.

$$\begin{aligned} TMAX \quad &= 30,64948 + DT \cdot (-12,66167) + DT^2 \cdot 67,31004 + DT^3 \cdot (-52,92461) + \\ &+ DT^4 \cdot 2233,22780 + DT^5 \cdot (-17715) + DT^6 \cdot 50614 \\ &+ DT^7 \cdot (-65046) + DT^8 \cdot 33913 + DT^{10} \cdot (-4000,48409) \end{aligned} \tag{2}$$

Onde $TMIN$ e $TMAX$ representam os modelos de regressão das temperaturas mínimas e máximas do ar, cuja única variável dependente é a data transformada (DT). A correlação de Pearson entre o modelo obtido e os dados reais é de no mínimo 0,62, conforme mostra na Tabela 1. O desvio padrão dos dados reais é maior que os obtidos pelo modelo, isto ocorre pelo fato que os modelos se baseiam no método dos mínimos quadrados. Como se trata de modelos de regressão, ao analisar a massa de validação, a confiabilidade desta

DOI: 10.5540/03.2017.005.01.0404

é alta comparada ao erro absoluto médio. Por exemplo: na temperatura mínima é de $1,25^oC$ (6,19% em relação ao real) e na temperatura máxima é de $1,59^oC$ (5,00% em relação ao real).

Tabela 1: Estatística da massa de treinamento (MT) e validação (MV) com alguns comparativos entre os dados reais e os obtidos pela modelagem.

	TMIN (MT)	TMAX (MT)	TMIN (MV)	TMAX (MV)
Correlação	0,76	0,62	0,72	0,64
R^2	0,58	0,38	0,52	0,41
Desvio Padrão Real	2,45	2,69	2,44	2,73
Desvio Padrão Modelo	1,85	1,64	1,87	1,64
Erro Absoluto Médio	1,17	1,58	1,25	1,59
% Erro Absoluto Médio	5,80%	4,97%	6,19%	5,00%
Média Real	20,16	31,86	20,13	31,92
Média Modelo	20,15	31,94	20,12	31,98

As medidas de precipitação apresentaram alto grau de aleatoriedade, devido a este fato, foi utilizado um modelo de regressão logística, como mostra na seguinte equação

$$PRECIPITACAO = \quad 1 - \frac{e^{\xi}}{1+e^{\xi}} \tag{3}$$

onde

$$\begin{aligned} \xi \quad &= -0,4661 + DT \cdot 0,6453 + DT^2 \cdot (-230,4) + DT^3 \cdot 3205,5 + \\ &+ DT^4 \cdot (-18728,4) + DT^5 \cdot 59811,3 + DT^6 \cdot (-107455) + \\ &+ DT^7 \cdot 103990 + DT^8 \cdot (-44742,5) + DT^{10} \cdot 4149,1 \end{aligned} \tag{4}$$

para definir a probabilidade de se incidir a chuva ou não, e através desta controlar uma função cosseinodal a qual os pontos máximos são o percentil 99 e o mínimo é 0. Conforme na Figura 4, os dados reais são discretos com valores muito dispersos. Mas o modelo matemático conforme supracitado obteve-se a seguinte distribuição conforme mostram as Figuras 4 e 5, referente as massas de treinamento e validação respectivamente.

A Tabela 2 sumariza as estatísticas básicas entre a MT e a MV entre os dados reais e do modelo em relação a precipitação. Apesar da baixa correlação (0,14 e 0,18 na massa de treinamento e validação, respectivamente), ao analisarmos os testes de estatística, observou-se que a distribuição do modelo de precipitação obedece as mesmas características da precipitação real. Além disso, a diferença entre as médias da precipitação real e do modelo de precipitação são pequenas: na MT é de 0,79 e na MV é de 0,54. O que torna preenchimento de um período longo plausível, em vez de assumir simplesmente um valor médio.

4 Conclusões

Os dados falhos ou omissos num banco de dados é prejudicial para a análises e tomadas de decisão, devido a esta, buscou-se modelos estatísticos plausíveis e representativos para

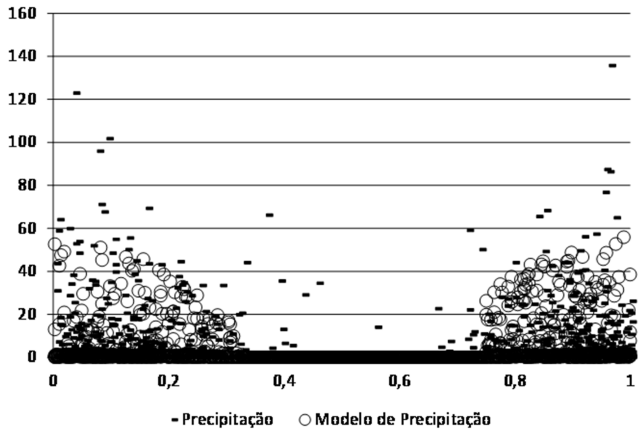

Figura 4: Dispersão entre os valores de precipitação real e do modelo, na massa de treinamento.

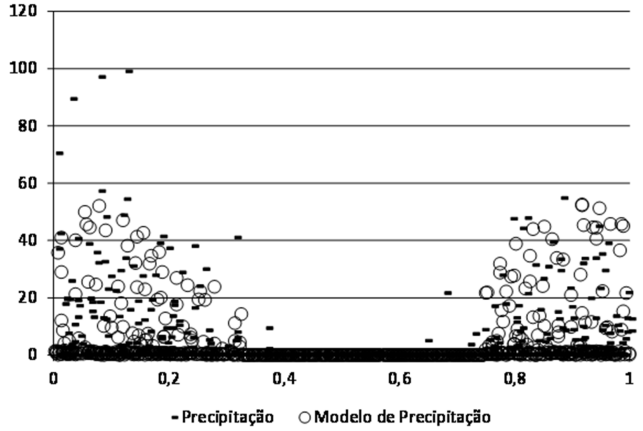

Figura 5: Precipitação no dia de cada ano em seus respectivos dias e mês do ano, sobrepostos.

este preenchimento.

Este trabalho apresentou modelos de regressão para as temperaturas máximas e mínimas do ar durante o dia escolhido. E um modelo matemático controlado por regressão logística para a precipitação de chuva no dia. Em todos os modelos apresentados verificou-se uma alta correlação (maior que 0,62) e uma reprodução comportamental acurada de um período com as mesmas características dos dados reais.

Este estudo propicia pesquisas futuras, pois há a necessidade de se validar em mais estações. Além do mais, a incorporação desta metodologia proposta em modelos hidrológicos possibilita estudos mais precisos a qual se quer uma aplicação.

DOI: 10.5540/03.2017.005.01.0404

Tabela 2: Análises através de estatísticas básicas da massa de treinamento e validação com alguns comparativos entre os dados reais e os obtidos pela modelagem de precipitação.

	MT	MV
Correlação	0,14	0,18
R^2	0,02	0,03
Desvio Padrão Real	11,68	11,31
Devio Padrão Modelo	8,91	9,86
Média Real	4,22	4,35
Média Modelo	3,43	3,81

Agradecimentos

Agradecemos as instituições EMBRAPA-Sinop e UFMT-Sinop pelo apoio.

Referências

[1] P. H. Abelson. *Agriculture and Climate Change, Science*, vol. p.9, 1992.

[2] K. L. Bristow, G. S. Campbell. *On the relationship between incoming solar radiation and daily maximum and minimum temperature*, Agricultural and Forest Meteorology, v.31, p.159-166, 1984.

[3] M. Donatelli, G. S. Campbell. *RadEst, a program to estimate global solar radiation.* In: 1st International Symposium Modelling Cropping Systems, 21-23 July, Lleida, Italia, p.289-290, 1999.

[4] S. Geng, F. W. Penning De Vries, I. Supit. *A simple method for generating daily rainfall data.* Agricultural and Forest Meteorology, v.36, p.363-376, 1986.

[5] J. P. Ometto, J. L. Stech, A. C.P. Cimbleris, M. A. Dos Santos, L. P. Rosa, D. Abe, J. G. Tundisi, N. Barros, F. Roland. *Carbon emission as a function of energy generation in hydroelectric reservoirs in Brazilian dry tropical biome.* Energy Policy, July 2013, Vol.58, pp.109-116

[6] C. W. Richardson, D. A. W. Wright. *A model for generating daily weather variables.* U.S. Department of Agriculture, Agriculture Research Service, Washington, D.C., ARS-8, p.88, 1984.

[7] J. T. Ritchie, D. C. Godwin, U. Singh. *Soil and weather inputs for the IBSNAT crop models.* In: Proceedings of IBSNAT Symposium of Decision Support System for Agrotechnology Transfer. Dept. of Agronomy and Soil Science, College of Tropical Agriculture and Human Resources, University of Hawaii, Honolulu, HI., p.31-45, 1990.

