11º Congresso Interinstitucional de Iniciação Científica – CIIC 2017
02 a 04 de agosto de 2017 – Campinas, São Paulo
ISBN 978-85-7029-141-7

GERAÇÃO DE SÉRIES TEMPORAIS DE DADOS METEOROLÓGICOS UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Henrique Lima Cará de Oliveira[1]; Stanley Robson de Medeiros Oliveira[2];
José Eduardo Boffino de Almeida Monteiro[2]

Nº 17602

RESUMO – *Este trabalho objetivou desenvolver uma metodologia baseada em algoritmos de Aprendizado de Máquina para gerar séries espaço-temporais de precipitação e temperatura. Foi definida uma região de estudo de formato retangular,* entre as latitudes de -18º e -22º e as longitudes de -52º e -39º, *incluindo a metade norte do Estado de São Paulo e parte do sul de Goiás, a metade sul de Minas Gerais e o Estado do Espírito Santo. A região foi escolhida por ser climaticamente bastante heterogênea e por conter muitas estações meteorológicas de diversas instituições, principalmente ANA e INMET. Foram utilizadas as séries temporais de precipitação e de temperatura máxima e mínima disponíveis na região, compreendendo o período de 01/01/1999 a 31/12/2013. Também foram utilizadas as bases externas TRMM e Nasa Power, cujos dados estão espacialmente dispostos em grades que cobrem a região de estudo. A região de estudo foi subdividida em formato de grade regular com resolução de 0,5º (latitude e longitude), resultando em 280 quadrículas, sendo 28 na horizontal e 10 na vertical. Para cada quadrícula foram ajustados modelos preditivos de precipitação diária, acumulada de 10 dias, temperatura máxima e temperatura mínima. Os resultados revelaram um bom ajuste dos modelos em relação aos valores previstos e observados, indicando um grande potencial da metodologia proposta tanto para a imputação de dados ausentes quanto para a geração de séries espaço-temporais em regiões sem a presença de dados medidos.*

Palavras-chave: Random Forest, Aprendizado com classes desbalanceadas, Modelos preditivos, Agrometeorologia, Imputação de dados.

1 Bolsista CNPq (PIBIC): Graduação em Engenharia da Computação, UNICAMP, Campinas-SP; henriquelimacoliveira@gmail.com
2 Embrapa Informática Agropecuária, Campinas, São Paulo, Brasil, {stanley.oliveira, eduardo.monteiro}@embrapa.br

11º Congresso Interinstitucional de Iniciação Científica – CIIC 2017
02 a 04 de agosto de 2017 – Campinas, São Paulo
ISBN 978-85-7029-141-7

ABSTRACT – *This work aimed at developing a methodology based on machine learning algorithms to generate space-time series of precipitation and temperature. A rectangular region of study was defined between the latitudes of -18º and -22º and longitudes of -52º and -39º, including the northern half of the State of São Paulo and part of the south of Goiás, the southern half of Minas Gerais and the State of Espírito Santo. This region was chosen due to its climate heterogeneity and also the high number of meteorological stations it contains, each of which belonging to one of several institutions, mainly ANA and INMET. There were used space-time series of precipitation, maximum and minimum temperature available in the region, covering the period ranging from 01/01/1999 to 12/31/2013. Also, there were used data extracted from the external bases, such as TRMM and Nasa Power, whose grids cover the study region. The study area was then subdivided into a regular grid format with a resolution of 0.5º (latitude and longitude), resulting in 280 squares, with 28 horizontally and 10 vertically. For each one of those squares, predictive models of daily precipitation, accumulated of 10 days, maximum temperature and minimum temperature were adjusted. The results showed a good fit of the models in relation to expected and observed values, which indicates a great potential of the proposed methodology both for imputation of absent data and for the generation of space-time series in regions with absence of measured data.*

Keywords: Random Forest, unbalanced class learning, Predictive modeling, Agrometeorology, Data imputation.

1 INTRODUÇÃO

Interrupções e erros em séries de dados gerados por estações meteorológicas são relativamente comuns devido a várias razões (DUMEDAH; COULIBALY, 2011; MWALE et al., 2012). Em estações automáticas, essas razões incluem desde o simples esgotamento de bateria até falha permanente de sensores ou da estação, que podem permanecer inativos até seu eventual reparo ou substituição, por dias ou meses. Em estações convencionais, as falhas estão mais associadas à ausência do observador meteorológico ou à quebra ou à falha dos instrumentos de medição. Por outro lado, mesmo quando registrados e armazenados em um banco de dados, os valores medidos estão sujeitos a imprecisões e erros de medida, de procedimentos de cálculo, de anotação pelo observador, de cópia ou transmissão de dados, de armazenamento ou transposição para tabelas. Em estudos agrometeorológicos, erros deste tipo implicam, quase sempre, na perda dos resultados

ISBN 978-85-7029-141-7

e da análise de todo um ciclo de cultura, safra ou ano, no local da ocorrência.

Técnicas de mineração de dados são uma alternativa promissora para estimar dados em séries temporais de precipitação e temperatura, já que alguns algoritmos utilizados no processo de descoberta de conhecimento são capazes de capturar relações não-lineares entre esses dados.

Em particular, o algoritmo *Random Forest* (BREIMAN, 2001) é eficiente para estimar séries temporais em várias áreas do conhecimento (VILLANUEVA, 2006; KANE et al., 2014). Suas principais vantagens são: (a) tem natureza não-paramétrica; (b) é um dos algoritmos de aprendizado de máquina com alta taxa de acurácia; (c) pode lidar com milhares de variáveis de entrada sem a necessidade prévia de redução de dimensionalidade; (d) identifica quais variáveis são as mais importantes na classificação; (e) é eficaz para estimar os dados faltantes, mesmo quando boa parte dos dados não está disponível; (f) é robusto na presença de ruído e de variáveis sem importância; (g) pode ser altamente flexível para realizar vários tipos de análise de dados, incluindo regressão, classificação e aprendizagem não supervisionada.

Neste trabalho, o algoritmo *Random Forest* será utilizado na geração de séries espaço-temporais de precipitação, temperatura máxima e temperatura mínima. Além disso, como existem eventos raros, principalmente em séries temporais de precipitação (precipitação diária acima de 50 mm), com o objetivo de melhorar os resultados previstos será apresentado um novo algoritmo (*OversamplingR Interval*s), que é uma proposta de adaptação do algoritmo *SmoteR* (TORGO et al., 2013) e consiste em aplicar um método de amostragem *Over-sampling* com geração de casos sintéticos no conjunto de treinamento para auxiliar o algoritmo Random Forest no processo de aprendizado de eventos raros. A amostragem visa equilibrar a distribuição dos casos menos representados com as observações mais frequentes. Esse procedimento é feito somente no conjunto de treinamento; o conjunto de teste não pode ser alterado para não comprometer a qualidade do modelo gerado.

Nesse contexto, o objetivo deste estudo foi desenvolver uma metodologia baseada em algoritmos de Aprendizado e Máquina (*Random Forest* e *OversamplingR Interval*s) para gerar séries temporais de dados meteorológicos.

2 METODOLOGIA

Foi definida uma região de estudo de formato retangular, entre as latitudes de -18º e -22º e as longitudes de -52º e -39º, incluindo a metade norte do Estado de São Paulo e parte do sul de Goiás, a metade sul de Minas Gerais e o Estado do Espírito Santo. Essa região foi escolhida por

ISBN 978-85-7029-141-7

conter um grande número de estações meteorológicas de diversas instituições, como ANA, INMET, IAC, CEMIG, Embrapa, EPAMIG, Universidades e outras. Além disso, foi escolhida por ser climaticamente bastante heterogênea, tanto no tempo (estacionalidade) como no espaço, com oito classificações diferentes pelo sistema de Köppen (Af, Am, Aw, Cfb, Cwa, Cwb, Cfa), devido à atuação de diversos condicionantes climáticos como áreas com relevo e altitudes bastante distintas, áreas montanhosas e áreas de planícies, influência litorânea, maior ou menor continentalidade vs. maritimidade nos extremos, com influências estacionais das massas de ar tropical atlântica, tropical continental, equatorial continental e polar atlântica. A metodologia deste trabalho foi dividida em três principais etapas, a saber: aquisição e pré-processamento de dados, modelagem de dados e reamostragem de eventos raros e, por fim, reamostragem, geração e validação dos modelos preditivos.

2.1 AQUISIÇÃO E PRÉ-PROCESSAMENTO DE DADOS

Todos os dados climáticos analisados consistem de séries temporais compreendidas no período de 01/01/1999 a 31/12/2013. Foram extraídas séries temporais diárias incompletas de precipitação e de temperatura máxima e mínima das estações meteorológicas. As de precipitação estão presentes em todas as estações, já as de temperaturas máxima e mínima estão presentes em apenas 20,87% delas. Além disso, foram utilizadas as bases externas AgMERRA, AgCFSR, Modelo de radiação solar GL 1.2 CPTEC, TRMM e Nasa Power, cujos dados estão espacialmente dispostos em grades que cobrem a região de estudo, das quais foram extraídas séries temporais completas de precipitação (TRMM, Nasa Power, AgMERRA, AgCFSR), radiação solar (Nasa Power, GL 1.2), evapotranspiração potencial (Nasa Power) e temperaturas máxima, mínima e média (Nasa Power, AgMERRA, AgCFSR), para o mesmo período.

Com o objetivo de gerar os modelos preditivos de precipitação diária e precipitação acumulada de 10 dias, temperatura máxima e temperatura mínima, a região de estudo foi dividida numa grade com resolução de 0,5º (latitude e longitude), resultando em 280 quadrículas, sendo 28 na horizontal e 10 na vertical; as quais poderiam conter ou não estações meteorológicas com dados de precipitação (Figura 1) e de temperatura máxima e mínima (Figura 2). Em seguida, foram definidas oito subáreas distintas, cada uma contendo 6 quadrículas, totalizando 48 quadrículas. Os modelos preditivos serão gerados a partir das quadrículas que contêm uma ou mais estações meteorológicas.

11º Congresso Interinstitucional de Iniciação Científica – CIIC 2017
02 a 04 de agosto de 2017 – Campinas, São Paulo
ISBN 978-85-7029-141-7

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28
Y1	1	3	1	4	4	1	4	7	2	1	3	1	3	1	2	4	2	0	2	5	4	4	3	1	4	3	1	0
Y2	0	1	1	0	3	3	4	2	0	7	8	3	5	3	9	2	3	9	1	2	1	2	4	7	5	3	0	0
Y3	0	1	2	3	2	8	3	5	4	3	2	6	5	4	0	2	5	2	1	3	3	7	4	8	6	6	0	0
Y4	0	4	2	2	1	4	3	2	0	3	3	8	3	6	8	5	5	2	6	9	4	3	5	6	4	0	0	0
Y5	2	1	2	3	5	3	3	2	5	7	6	5	3	1	2	6	14	15	7	2	7	2	7	9	6	1	0	0
Y6	1	0	7	14	8	12	6	12	11	15	10	1	6	6	4	10	14	12	4	4	4	4	5	12	7	0	0	0
Y7	2	5	10	6	7	9	12	17	13	13	13	6	4	2	2	4	6	6	4	7	5	8	9	10	1	0	0	0
Y8	1	6	10	13	14	13	8	15	17	16	15	11	3	8	6	13	9	9	6	7	2	4	5	1	0	0	0	0
Y9	3	6	12	16	9	11	8	13	18	16	17	13	5	5	11	6	11	4	8	6	6	4	4	0	0	0	0	0
Y10	5	6	6	10	9	13	9	7	10	19	11	16	3	9	6	9	17	11	11	14	11	2	2	0	0	0	0	0

Figura 1. Número de estações com séries de precipitação por quadrícula num grid com resolução de 0,5º (latitude e longitude) subdividido em 8 áreas de interesse.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28
Y1	0	2	0	3	2	0	2	3	1	0	1	0	2	0	0	1	0	0	0	1	1	3	0	0	3	1	0	0
Y2	0	0	0	0	1	2	3	0	0	5	5	0	2	0	8	1	1	4	0	0	0	1	0	1	0	0	0	0
Y3	0	0	0	1	1	6	1	0	2	0	1	1	3	2	0	0	3	1	0	1	1	4	0	1	1	2	0	0
Y4	0	1	1	1	1	2	1	1	0	2	1	3	1	3	5	2	1	0	2	9	1	0	1	0	2	0	0	0
Y5	0	0	1	0	0	1	0	0	2	5	3	2	0	0	0	2	5	5	3	1	3	0	1	1	0	0	0	0
Y6	0	0	2	4	0	4	0	1	2	2	0	0	3	1	1	6	1	6	0	0	1	0	0	0	3	0	0	0
Y7	0	2	1	1	0	1	1	3	1	1	3	3	1	0	0	0	2	3	0	4	1	1	1	1	0	0	0	0
Y8	0	2	1	2	3	2	0	4	4	4	2	2	2	2	1	7	4	2	1	1	1	1	1	0	0	0	0	0
Y9	0	0	1	3	0	2	0	1	1	4	2	3	3	2	3	1	3	0	4	1	1	1	2	0	0	0	0	0
Y10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 2. Número de estações com séries de temperatura por quadrícula num grid com resolução de 0,5º (latitude e longitude) subdividido em 8 áreas de interesse.

Dada uma dessas quadrículas, *quad*, é preciso selecionar uma estação *est* em *quad* e um conjunto de estações vizinhas *S* que contenham dados da variável de interesse, a fim de que os mesmos sejam utilizados como covariáveis no modelo preditivo. A seleção dessas estações, porém, é um grande problema, pois existe um alto nível de incompletude no total das séries espaço-temporais, nas quais há aproximadamente 28% de ausências de observações, e, além disso, dado um conjunto de estações, a disponibilidade dessas observações em cada uma delas depende da data e da própria estação, não havendo um padrão.

Assim, a fim de selecionar o maior número de observações feitas em diferentes estações mas com datas coincidentes, propõe-se utilizar uma adaptação do algoritmo heurístico de resolução do problema da mochila, modificado com pesos nas variáveis para o problema em questão (PISINGER, 1995). O problema da mochila (em inglês, *Knapsack problem*) é um problema de optimização combinatória. O objetivo é que se preencha uma mochila com o maior valor possível, não ultrapassando o peso máximo. No caso em questão, os itens correspondem às séries espaço-temporais das estações vizinhas; dado um item *i*, o peso de *i* é igual a porcentagem de dados excluídos nas intersecções de *i* com os outros itens (séries espaço-temporais) já escolhidos, o valor

ISBN 978-85-7029-141-7

de *i* é igual ao inverso da distância entre as posições da estação a qual o item *i* pertence e a estação *est*, e, por fim, o peso máximo é igual a porcentagem máxima de dados faltantes.

O número de estações vizinhas escolhidas varia entre n_{min} e n_{max}, e as estações escolhidas dessa forma satisfazem a exigência de que a porcentagem de dados faltantes na intersecção das séries espaço-temporais não supere um valor *p*. Quanto ao modo de operação, o algoritmo tenta, a cada iteração, adicionar as estações vizinhas mais próximas, sujeitas à condição anterior, e de tal forma que o raio médio da distância r_{avg} entre todas as estações que já foram adicionadas não supere um valor limite *threshold*, mas, eventualmente, este valor pode ser incrementado sucessivamente por um valor *incr* até que os critérios anteriores sejam satisfeitos.

2.2 MODELAGEM DE DADOS E REAMOSTRAGEM DE EVENTOS RAROS

Escolhida uma variável de interesse, seja *quad* uma quadrícula para a qual pretende-se gerar um modelo preditivo. Aplica-se o algoritmo do problema da mochila em *quad*, em seguida, são extraídas as séries espaço-temporais de *est* e de *S* da variável de interesse. Feito isso, são extraídas as séries espaço-temporais completas de cada uma das variáveis meteorológicas das bases externas nos pontos que recaem sobre *quad*. Vale notar que essa extração é sempre possível porque as bases externas recobrem todo a grade. Todas essas séries são então reunidas compondo um único conjunto de dados, de forma que a série espaço-temporal advinda de *est* é a variável alvo de predição, e as demais séries são as covariáveis do modelo preditivo. Em seguida, desse conjunto de dados, são eliminadas todas as observações que contêm dados faltantes.

Prever dados diários de precipitação é uma tarefa árdua, pois a maioria dos valores são iguais a zero, e isso tende a polarizar o processo de aprendizado, o que dificulta inferir eventos raros (precipitação diária acima de 50 mm). A fim de contornar essa dificuldade, propôs-se uma alteração na etapa de modelagem dos dados, que, nessa nova abordagem, consiste em aplicar um método de amostragem *Over-sampling* com geração de casos sintéticos no conjunto de treinamento. O objetivo é reamostrar os eventos raros para que o algoritmo possa aprendê-lo. A alteração tem por objetivo equilibrar a distribuição dos casos menos representados (mas igualmente importantes) com as observações mais frequentes (TORGO et al., 2013). Esse procedimento é feito somente no conjunto de treinamento para que o algoritmo de aprendizado possa aprender os eventos raros; o conjunto de teste não pode ser alterado.

A realização do *Over-sampling* no conjunto de treinamento, é feita por meio de uma adaptação no algoritmo *SmoteR* (TORGO et al., 2013). As modificações consistem em: não realizar

ISBN 978-85-7029-141-7

Under-sampling, mas apenas Over-sampling; e o fato de, para cada intervalo de dados da variável alvo, gerar uma porcentagem bem definida de exemplos sintéticos, que são interpolações entre os dados reais existentes no conjunto de treinamento; assim, o algoritmo modificado é bem flexível, pois permite ao analista de dados gerar diferentes porcentagens de valores raros nos intervalos. Além dos casos de precipitação, propõe-se aplicar a mesma abordagem nos casos de temperatura máxima e mínima, a fim de avaliar se há ganho na qualidade dos dados previstos.

2.3 REAMOSTRAGEM, GERAÇÃO E VALIDAÇÃO DOS MODELOS PREDITIVO

Após obter o conjunto de dados para a quadrícula *quad* pela etapa anterior, esse mesmo conjunto foi dividido aleatoriamente em dois subconjuntos (com o mesmo número de variáveis), sendo que 2/3 dos dados deram origem ao conjunto de treinamento D_{train} e os outros 1/3 deram origem ao conjunto de teste D_{test}. Ao conjunto D_{train} foram aplicados os procedimentos de reamostragem, e, por fim, o conjunto obtido foi utilizado na geração dos modelos preditivos baseados no algoritmo *Random Forest* (BREIMAN, 2001).

O *Random Forest* é uma técnica de classificação e regressão desenvolvida por BREIMAN (2001), que consiste num conjunto de árvores de decisão combinadas para solucionar problemas de classificação. Cada árvore de decisão é construída utilizando uma amostra aleatória inicial dos dados e, a cada divisão desses dados, um subconjunto aleatório de *m* atributos é utilizado para a escolha dos atributos mais informativos. No final, *Random Forest* gera uma lista dos atributos mais importantes no desenvolvimento da floresta, que são determinados pela importância acumulada do atributo nas divisões dos nós de cada árvore da floresta. Por fim, os modelos preditivos foram validados com o uso do conjunto D_{test} (método *Hold Out*). Todo o processo foi feito tanto para os casos de temperatura máxima e temperatura mínima, quanto para os casos de precipitação diária e precipitação acumulada de 10 dias.

As implementações dos algoritmos do problema da mochila, *Random Forest*, *SmoteR* adaptado (que a partir desse trabalho recebe o nome *OversamplingR Intervals*), bem como os procedimentos da etapa de pré-processamento foram realizados no ambiente RStudio (RStudio, 2017), que é um software livre e ambiente de desenvolvimento integrado na linguagem de programação R.

3 RESULTADOS E DISCUSSÃO

Nas execuções do algoritmo heurístico de resolução do problema da mochila foram utilizados

ISBN 978-85-7029-141-7

os seguintes valores de parâmetros: n_{min} = 6, n_{max} = 12, p = 0,15, *threshold* = 1 e *incr* = 1.

Foram gerados modelos preditivos de precipitação diária, precipitação acumulada de 10 dias, temperatura máxima e temperatura mínima em todas as quadrículas que continham estações com dados da variável de interesse; mas, por uma questão de espaço, dado o alto número de modelos, estão expostas aqui as avaliações de desempenho dos modelos preditivos de precipitação obtidos em apenas duas dessas quadrículas: uma na qual os modelos tiveram bom desempenho, e a outra em que o desempenho foi inferior. Nos casos de temperatura máxima e mínima, o desempenho foi bom em quase todas as quadrículas.

Nas execuções em que houve amostragem *Over-sampling*, foram testadas algumas configurações do algoritmo *OversamplingR Intervals* de divisões dos dados em intervalos de valores distintos e diversas porcentagens de reamostragem em cada um desses subintervalos. A configuração que produziu os melhores resultados foi selecionada em cada um dos casos de precipitação.

Em relação aos modelos preditivos de precipitação diária obtidos na quadrícula (2, 1), o uso da técnica de amostragem (Figura 4, direita) resultou num aumento de aproximadamente 10% no valor da correlação, quando comparado ao valor obtido sem o uso de amostragem (Figura 4, esquerda); já para os modelos preditivos de precipitação diária obtidos na quadrícula (12, 9), o uso da técnica de amostragem (Figura 5, direita) resultou num aumento de aproximadamente 1% no valor da correlação, quando comparado ao valor obtido sem o uso de amostragem (Figura 5, esquerda). Essas diferenças ocorrem porque a amostragem faz um esforço no sentido de equilibrar a distribuição dos casos menos representados, sendo esse um problema muito presente na quadrícula (2, 1), e menos presente na quadrícula (12, 9).

Para os modelos preditivos de precipitação acumulada de 10 dias (Figuras 6 e 7), também houve ganhos utilizando reamostragem, conforme o esperado.

11º Congresso Interinstitucional de Iniciação Científica – CIIC 2017
02 a 04 de agosto de 2017 – Campinas, São Paulo
ISBN 978-85-7029-141-7

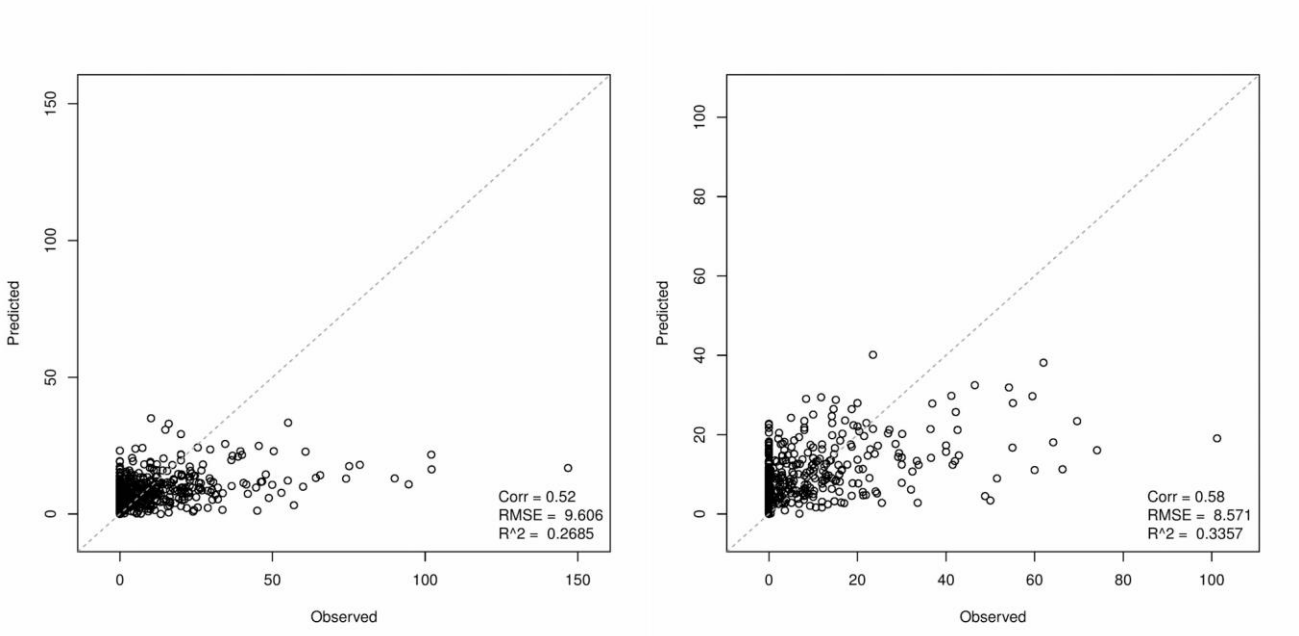

Figura 4. Gráficos de avaliação de desempenho dos modelos preditivos baseados no algoritmo *Random Forest* para precipitação diária obtidos na quadrícula (2, 1), sem e com amostragem *Over-sampling* (esquerda e direita, respectivamente).

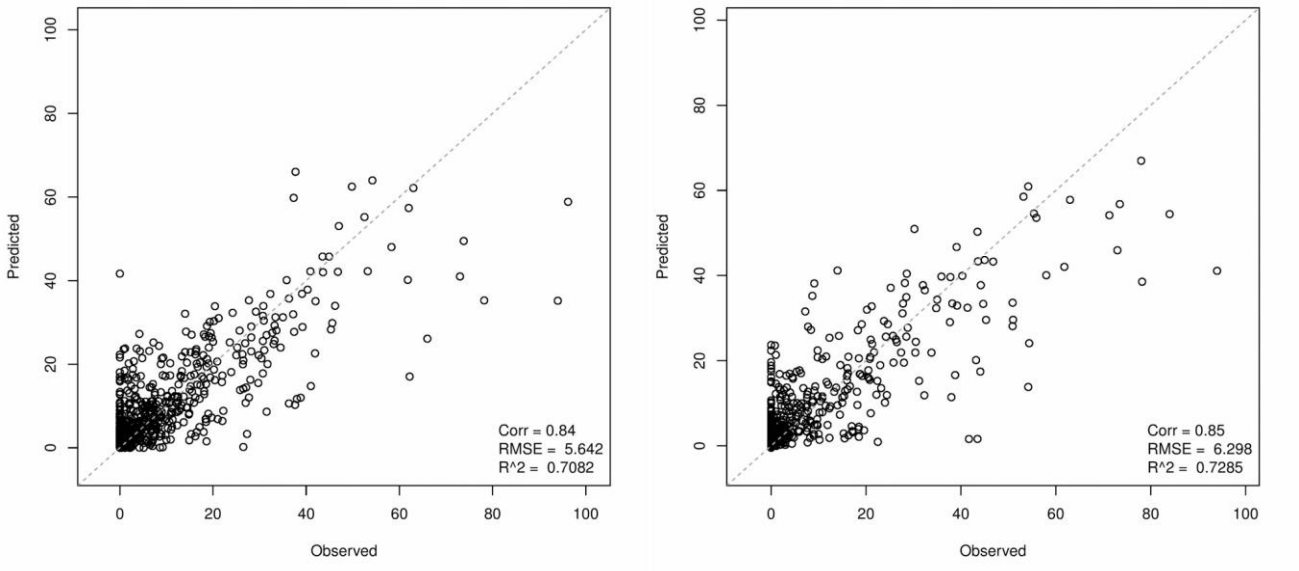

Figura 5. Gráficos de avaliação de desempenho dos modelos preditivos baseados no algoritmo *Random Forest* para precipitação diária obtidos na quadrícula (12, 9), sem e com amostragem *Over-sampling* (esquerda e direita, respectivamente).

11º Congresso Interinstitucional de Iniciação Científica – CIIC 2017
02 a 04 de agosto de 2017 – Campinas, São Paulo
ISBN 978-85-7029-141-7

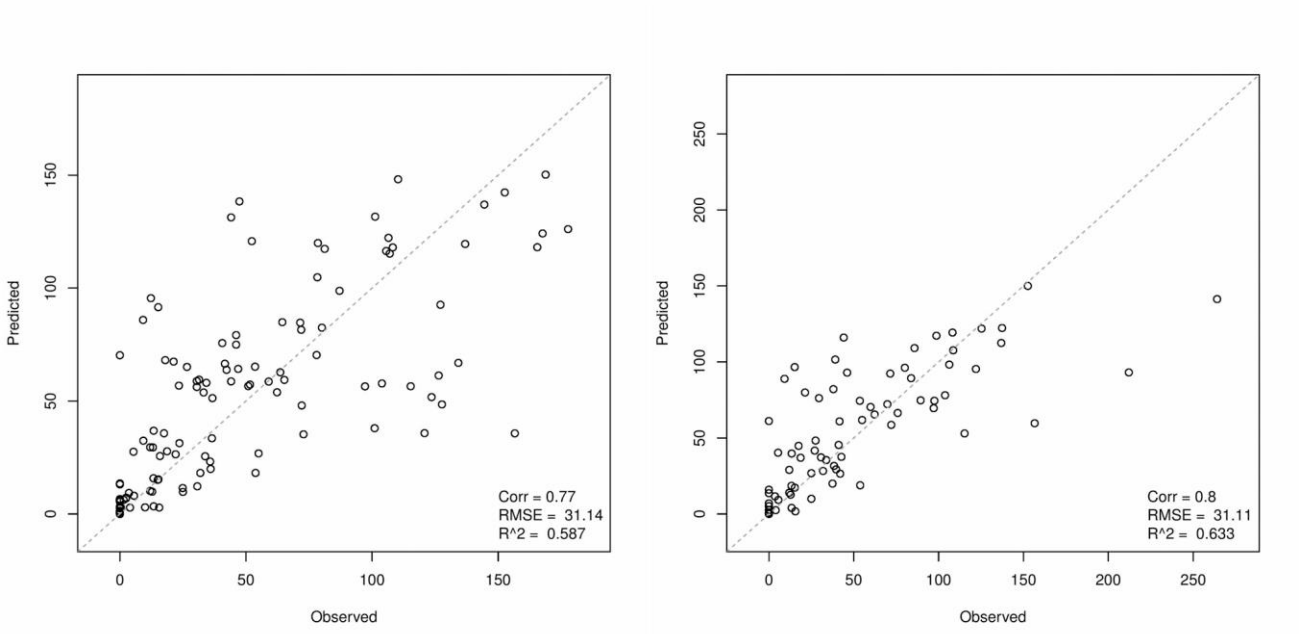

Figura 6. Gráficos de avaliação de desempenho dos modelos preditivos baseados no algoritmo *Random Forest* para precipitação acumulada de 10 dias obtidos na quadrícula (2, 1), sem e com amostragem *Over-sampling* (esquerda e direita, respectivamente).

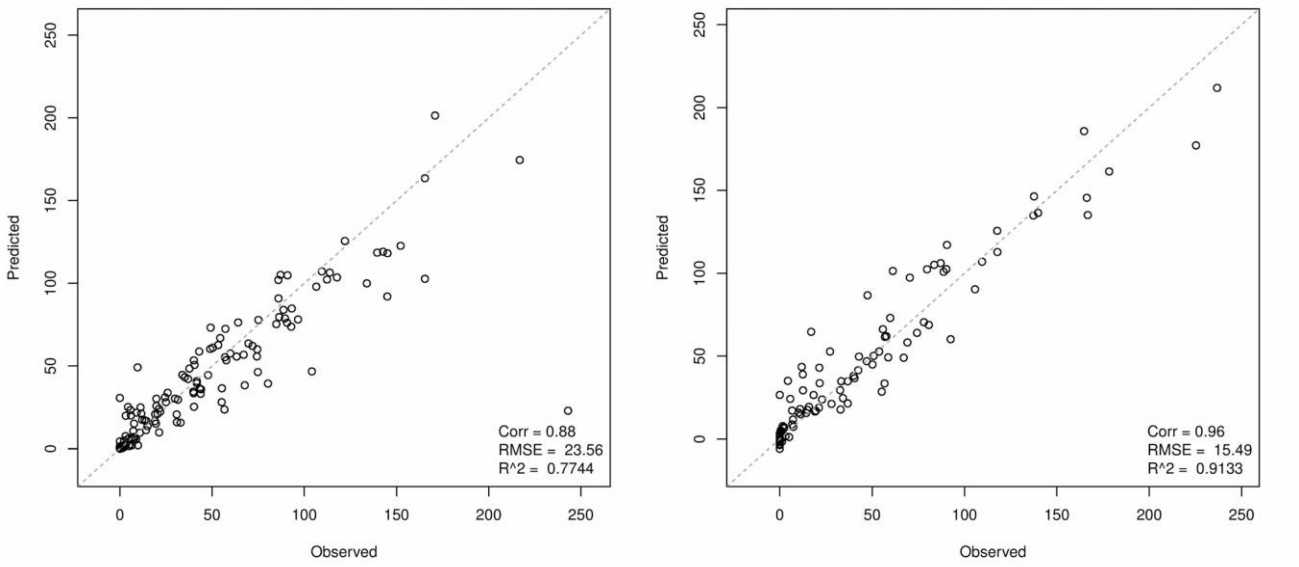

Figura 7. Gráficos de avaliação de desempenho dos modelos preditivos baseados no algoritmo *Random Forest* para precipitação acumulada de 10 dias obtidos na quadrícula (12, 9), sem e com amostragem *Over-sampling* (esquerda e direita, respectivamente).

Nos casos de temperatura (Figura 8), o uso de amostragem *Over-sampling* não melhorou os

ISBN 978-85-7029-141-7

resultados, uma vez que os dados já estavam bem distribuídos, sendo esse o comportamento esperado. Portanto, a amostragem *Over-sampling* apresenta resultados satisfatórios para os casos de precipitação diária e acumulada.

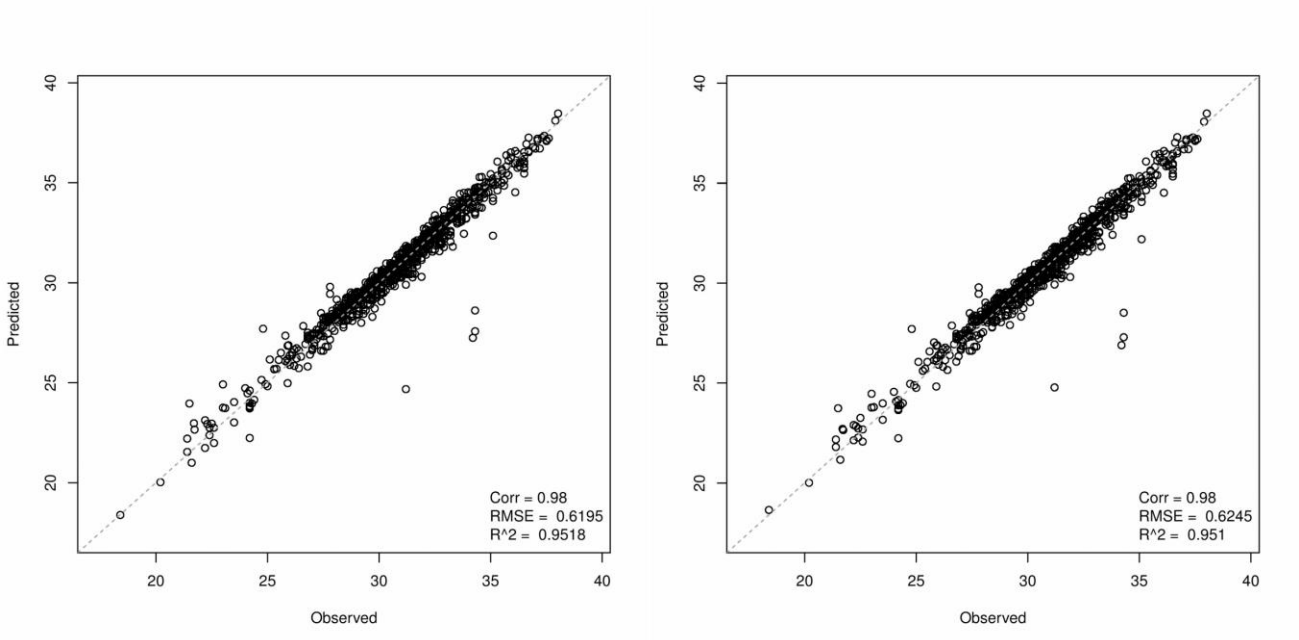

Figura 8. Gráficos de avaliação de desempenho dos modelos preditivos baseados no algoritmo *Random Forest* para temperatura máxima obtidos na quadrícula (2, 1), sem e com amostragem *Over-sampling* (esquerda e direita, respectivamente).

A avaliação de desempenho dos modelos foi realizada em todas as quadrículas que continham dados na região de estudo. A Tabela 1 apresenta os valores de correção dos modelos para precipitação diária. Foi observado que 63,66 das quadrículas exibiram correlação superior a 0,7. Isso indica a grande utilidade dos modelos para a geração de séries em quadrículas ou regiões sem a presença de dados medidos.

Tabela 1. Porcentagem de quadrículas totais cujos valores de correlação dos modelos preditivos de precipitação diária estão contidos nas respectivas faixas de valores

Faixa de valores (correlação)	Porcentagem de quadrículas
Abaixo de 0,5	4,67%
0,5 - 0,6	11,30%
0,6 - 0,7	20,36%
0,7 - 0,8	40,12%
0,8 - 0,9	22,62%
0,9 - 1.0	0,92%

ISBN 978-85-7029-141-7

4 CONCLUSÕES

A metodologia proposta, baseada no algoritmo de aprendizado de máquina *Random Forest* e no algoritmo *OversamplingR Interval*s permitiu a geração de séries espaço-temporais de precipitação diária e acumulada de 10 dias, que exibiram um bom ajuste em relação aos valores observados, mostrando grande potencial de uso para imputação de dados ausentes ou na geração de séries em quadrículas ou regiões sem a presença de dados medidos. Quanto às séries de temperatura, pelo fato dos dados já estarem bem distribuídos, o uso do *OversamplingR Interval*s não resultou em diferenças significativas.

5 AGRADECIMENTOS

Ao programa CNPq/PIBIC pela concessão da bolsa de Iniciação Científica, processo Nº 145155/2016-1.

6 REFERÊNCIAS

BREIMAN, L. Random Forests. In Machine Learning, v.45, p.5–32, 2001.

DUMEDAH, G.; COULIBALY, P. Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. Journal of Hydrology, v. 400, p. 95-102, 2001. doi:10.1016/j.jhydrol.2011.01.028.

KANE, M. J.; PRICE, N.; SCOTCH, M.; RABINOWITZ, P. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics, v.15, n.1, 2014.

MWALE, F.D.; ADELOYE, A.J.; RUSTUM, R. Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi – A self organizing map approach. Physics and Chemistry of the Earth, 50–52, p. 34–43, 2012.

PISINGER, D. Algorithms for Knapsack Problems. PhD thesis, Department of Computer Science, University of Copenhagen, February 1995.

RStudio – Open source and enterprise-ready professional software for R. Disponível em https://www.rstudio.com. Consultado em 3 de Maio de 2017.

TORGO, L.; RIBEIRO, R. P.; PFAHRINGER, B.; BRANCO, P. Smote for regression. In: PORTUGUESE CONFERENCE ON ARTIFICIAL INTELLIGENCE, 16., 2013, Angra do Heroismo. **Progress in artificial intelligence**: proceedings. New York: Springer, 2013. p. 378-389. (Lecture notes in artificial intelligence, 8154). EPIA 2013.

VILLANUEVA, W. J. P. **Comitê de Máquinas de Predição de Séries Temporais**. 2006. p.150. (Dissertação) - Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas.