

Artigo

Avaliação e Comparação de Dois Métodos de Qualificação de Dados Diários de Precipitação no Estado de Santa Catarina

Renzo Angelo Viloche Morales¹, Carlos Eduardo Salles de Araujo²¹*Conselho Nacional de Desenvolvimento Científico e Tecnológico, Florianópolis, SC, Brasil.*²*Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina, Florianópolis, SC, Brasil.*

Recebido: 7 de Dezembro de 2015 – Aceito: 14 de Maio de 2016

Resumo

Dados de chuva provenientes de uma rede meteorológica do Sul do Brasil são usados para avaliar o desempenho de dois algoritmos de detecção de dados espúrios. Ambos os métodos empregam uma abordagem estatística e consistência espacial baseada nas distâncias e diferenças de altitude entre duas medidas de pluviômetros. Uma variação do método de múltiplas Gamas de You, Hubbard, Nadarajah e Kunkel (2007) é considerada neste estudo. Neste método a distribuição de precipitação média de estações vizinhas é particionada supondo-se que cada intervalo pode ser modelado por uma distribuição Gama. O segundo método não assume nenhuma distribuição a priori, usando informação pontual espacial e acumulada temporal de medidas pluviométricas vizinhas na consistência dos dados de chuva diários. Para avaliar a confiabilidade e a precisão em detectar dados espúrios por ambos algoritmos, são introduzidos erros semeados na série histórica de chuvas. Um modelo probabilístico bidimensional de erros introduzidos/detectados (sim-não) é empregado para calcular métricas referentes a probabilidades de correta detecção e alarme falso cometido pelo algoritmo. Verifica-se que o novo algoritmo proposto supera o algoritmo das múltiplas distribuições Gama.

Palavras-chave: dados de chuva, pluviômetro, detecção de dados espúrios.

Evaluation of Two Different Methods of Data Quality Analysis Using Local Daily Precipitation Measurements at Santa Catarina State

Abstract

Daily rainfall data from a meteorological network in Southern Brazil is used to assess the performance of two different outliers detection algorithms. Both methods use a statistical and spatial consistency approach based in distance and elevation difference between two rain gauge measurements. A variation of the Multiple Interval Gamma Distribution method of You, Hubbard, Nadarajah e Kunkel (2007) is considered in this study. Neighboring stations data is gathered to get the local average rainfall distribution. The precipitation range of values is partitioned so one makes the assumption that every interval can be modeled by a Gamma distribution. The second method assumes no prior distribution characteristic, and instead uses point spatial and cumulated temporal information from neighboring rain gauge stations to consist daily rainfall data. In order to assess the reliability of the detected outliers, as well the accuracy, seeded errors are introduced in the historical rainfall series. A two dimensional probability model of introduced/detected error (yes-no) is used to compute metrics related to the correct detection and false alarm probabilities made by the algorithm. We verify that the new proposed method overcomes the Multiple Interval Gamma Distribution method.

Keywords: rainfall data, pluviometer, outlier detection.

1. Introdução

O monitoramento de chuvas possibilita compreender o ciclo hidrológico que influencia a dinâmica ecológica e ambiental, afetando também atividades de caráter econô-

mico e social. A precipitação é comumente registrada utilizando-se uma rede de pluviômetros espacialmente distribuídos que medem volumes acumulados por períodos de tempo pré-determinados. Essas medidas estão sujeitas a diferentes fontes de erros (Habib, Krajewski e Kruger,

2001): falhas eletromecânicas do dispositivo (interrupção de transmissão, falha temporária de alimentação, entupimento do coletor); operação e leitura manual incorreta; desconformidade do tamanho do receptor em relação ao padrão de volume utilizado; fontes de perturbação externa (vento ou fauna). Além disso, devido a crescente tendência de grandes volumes de dados disponíveis, há uma maior necessidade do uso de procedimentos automatizados para uma qualificação de dados eficiente (Durre *et al.*, 2010). Desta forma, torna-se necessário desenvolver métodos para avaliar a qualidade dos dados de chuva. A literatura relativa a este tema é ainda relativamente escassa, sendo que a qualificação é tratada principalmente a nível diário (Hubbard, *et al.*, 2005; You, Hubbard, Nadarajah e Kunkel, 2007; Hamada, Arakawa e Yatagai, 2011; Nie *et al.*, 2012), e mensal (Eischeid, Baker, Karl e Diaz, 1995; González-Rouco, Jiménez, Quesada e Valero, 2001). A principal dificuldade encontrada ao abordar este tema reside no fato de que, dependendo do tipo de escala considerada, a precipitação tende a se comportar como uma variável descontínua no tempo e no espaço.

Neste trabalho são estudados dois algoritmos com abordagens estatísticas diferentes para a qualificação de medidas de precipitação diária. O primeiro é uma versão adaptada do trabalho de You, Hubbard, Nadarajah e Kunkel (2007) que utiliza distribuições Gamas para diferentes faixas de valores pluviométricos. O segundo é um novo método desenvolvido pelos autores e que leva em consideração o contexto espacial e temporal de precipitação. Uma metodologia de sementeira de erros aleatórios baseada nos trabalhos de Hubbard, *et al.* (2005) e You, Hubbard, Nadarajah e Kunkel (2007) é utilizada para avaliar a capacidade de detecção de dados espúrios para cada um dos algoritmos, utilizando-se dados de chuva de estações pluviométricas do estado Santa Catarina.

2. Material e Métodos

Neste estudo foram utilizados dados acumulados diários de chuva das 12 h UTC, de janeiro de 2000 até dezembro de 2015, provenientes do banco de dados do Centro de Informações de Recursos Ambientais e Hidrometeorologia da Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina (EPAGRI/Ciram).

Um tratamento preliminar dos dados foi feito de forma a excluir valores extremos de precipitação horária que excederam uma probabilidade acumulada de 99,9% para distribuições Gama ajustadas para as séries históricas de precipitação de todas as estações. Foram também excluídos todos os registros (exceto o primeiro) com medidas repetidas em sequência para um período de 5 horas ou mais.

Na avaliação e teste dos métodos de detecção de dados espúrios selecionou-se um subconjunto de cinco estações com no mínimo 5 anos de medidas de chuva. A Tabela 1 apresenta a lista das escolhidas, denominadas de estações alvo ou de análise, suas respectivas coordenadas geográficas, acumulados anuais médios de chuva e o número mediano de estações vizinhas. Foram selecionadas para constituir o grupo de análise apenas estações com densidades semelhantes (define-se como densidade o número de estações vizinhas existentes dividido pela área de ocupação da vizinhança - A área de ocupação da vizinhança é a região ocupada pela circunferência centrada na estação alvo e de raio igual a 150 km utilizado para localizar estações vizinhas). Ademais, foram priorizadas as estações onde sua distribuição de vizinhos se mostra visualmente homogênea. Isto foi realizado para evitar a influência da densidade e distribuição de estações vizinhas na interpretação de resultado garantindo uma maior confiabilidade das estatísticas.

Estações vizinhas são definidas em função da distância e da variação de altitude de acordo com os seguintes critérios: as estações próximas ao nível do mar, até 200 m de altitude, são vizinhas àquelas com distância máxima de 150 km e dentro desta mesma faixa altitudinal; as estações localizadas acima de 200 m, além da distância máxima de 150 km, consideram também uma variação de altitude máxima de 250 m entre estações.

A Fig. 1 apresenta a distribuição espacial das cinco estações alvo (em vermelho) e das 240 estações vizinhas consideradas (em azul) para aplicação dos algoritmos de qualificação da precipitação diária.

2.1. Método de múltiplas Gamas para detecção de dados espúrios

De acordo com You, Hubbard, Nadarajah e Kunkel (2007) é possível encontrar valores suspeitos para acumulados diários de precipitação usando a informação de precipitação diária média na vizinhança, particionada em um conjunto de n intervalos de intensidade de chuva.

Tabela 1 - Estações escolhidas para análise. O número de vizinhos representa um valor mediano para o período de estudo.

Estação	Latitude (°S)	Longitude (°W)	Altitude (m)	Média anual (mm)	Número de vizinhos
Blumenau	26,906	49,079	40	1.716,4	15
Urussanga	28,532	49,315	48	1.751,9	13
Ponte Serrada	26,921	51,928	1.025	2.174,6	17
Rio do Campo	26,937	50,146	592	2.036,1	21
Joinville	26,253	48,857	20	2.257,9	14

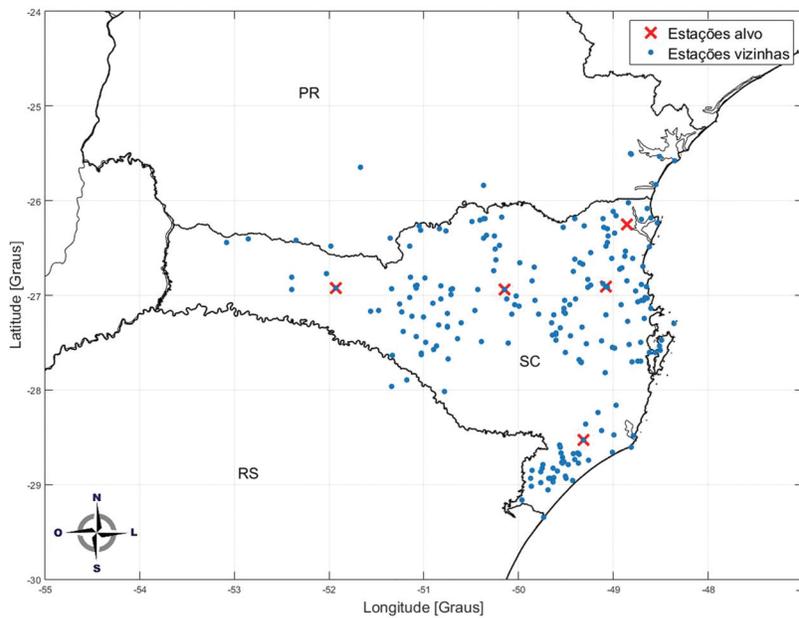


Figura 1 - Localização geográfica das estações em estudo do estado de Santa Catarina. São apresentadas as estações de análise (alvo) em vermelho e o conjunto total disponível de estações vizinhas em azul.

O algoritmo inicia considerando uma estação alvo, localizando sua vizinhança e obtendo o valor médio diário de chuvas para esta vizinhança. Os limites de cada regime são definidos de forma a preservar o número de amostras por intervalo (densidade de pontos por intervalo aproximadamente constante). A distribuição acumulada de probabilidade de precipitação diária média é, portanto, dividida em n intervalos, onde cada intervalo representa um regime de precipitação média na vizinhança. Os limites de cada intervalo são associados a cada uma das probabilidades acumuladas do particionamento: $p = 0, 100 \times (1/n), 100 \times (2/n), \dots, 100$. Assume-se que é sempre possível encontrar uma distribuição Gama que represente bem o comportamento de cada regime.

O procedimento consiste em escolher um intervalo de precipitação média e aplicar um teste bilateral a partir de probabilidades de corte p e $1-p$ da distribuição Gama correspondente ao intervalo selecionado, verificando se o valor da medida da estação alvo encontra-se dentro destes limites de corte. O critério para escolher uma dentre as n distribuições é feito de acordo com o valor médio μ de chuva na vizinhança. O intervalo que contém μ define a distribuição escolhida para aplicar o teste bilateral. Isto ajuda a garantir que o teste verifique o valor analisado sempre com o regime de chuvas mais apropriado. Devido ao fato de ser usado um modelo composto por uma família de distribuições Gama para representar regimes de chuva por intervalos, valores nulos de precipitação não podem ser avaliados pelo método.

2.2. Novo método para detecção de dados espúrios

Foi desenvolvido um método híbrido que utiliza a informação do contexto espacial e temporal das chuvas. O contexto espacial é considerado ao comparar a medida da estação de análise com medidas das estações vizinhas para uma mesma data. O contexto temporal é considerado de maneira semelhante à Nie *et al.* (2012), porém neste caso são usados os acumulados de chuva de d dias para as estações alvo e vizinhas.

A técnica empregada se aproxima do conceito de dupla-massa comumente aplicado na área de hidrologia (Gordon, Finlayson e McMahon, 2004). Desta maneira, cada acumulado é o somatório da precipitação para uma janela de d dias, com a janela centrada na data em análise e sendo d um número ímpar.

Seja a medida para uma dada estação alvo $x(t,j)$, onde os índices t e j são, respectivamente, a data e a estação alvo. A vizinhança é composta por medidas para a mesma data t , para estações j_1, j_2, j_3, \dots , etc. O algoritmo é então composto pelo conjunto de três testes descritos a seguir. Teste 1: Para um dado dia t , é verificado se o valor de precipitação $x(t,j)$ se encontra dentro dos limites definidos por $\mu - f_1\sigma < x(t,j) < \mu + f_1\sigma$, ou seja, pelo valor médio mais f_1 desvios padrões da distribuição de precipitação de estações vizinhas para esta data. Teste 2: Verifica se a mediana dos desvios (em valor absoluto) é inferior a um limite pré-estabelecido m . Teste 3: Similarmente ao teste 1, verifica se a precipitação acumulada para um intervalo temporal de d dias encontra-se dentro de limites definidos pelo valor médio dos acumulados da vizinhança mais f_2 desvios padrões desta distribuição. Cada um destes testes é aplicado na sequência, caso o dado analisado $x(t,j)$ não tenha passado no teste prece-

dente. Caso o dado seja aprovado em algum destes testes (condição é verificada) então o mesmo é considerado válido. Caso tenha sido reprovado por todos os testes, então o dado é marcado como suspeito devendo ser submetido a uma análise manual posterior.

2.3. Semeadura de erros

Para avaliar e comparar a eficiência dos métodos em identificar dados espúrios é necessário semear em uma base de dados consistida um número fixo de erros conhecidos. O procedimento de semeadura consiste em duas etapas: (i) escolher aleatoriamente registros relativos a eventos de chuva; (ii) adicionar um erro de natureza também aleatória a cada uma das medidas. Um gerador de números pseudo-aleatórios com distribuição uniforme é utilizado para garantir que o processo seja aleatório. O erro introduzido possui magnitude aleatória, dada por $E_x = r \cdot \sigma_j$, onde r é o erro normalizado de distribuição uniforme, e σ_j representa o desvio padrão do regime de precipitação diário da estação j . O valor da medida com o erro semeado é dado pela expressão $x(t,j)+E_x$. Para evitar medidas não físicas, novos valores de erro normalizado são calculados sempre que o erro semeado resultante produz medidas negativas. Desta maneira, a distribuição dos erros normalizados possui uma forma assimétrica. A Fig. 2 apresenta um exemplo típico para esta distribuição.

2.4. Avaliação de desempenho da detecção de erros semeados

Na literatura, alguns trabalhos fizeram uso de um critério subjetivo para avaliar o método empregado na detecção de dados espúrios (Kunkel *et al.*, 2005 e Durre *et al.*, 2010), onde é considerado apenas o número de “falsos-positivos” para uma amostra de dados selecionados aleatoriamente. Por outro lado, trabalhos com critérios objetivos de avaliação consideraram apenas frações de erros do tipo I e do tipo II (Hubbard *et al.*, 2005; You, Hubbard, Nadarajah e Kunkel, 2007). Neste trabalho, é adotado um critério descrito por duas variáveis binárias discretas: observação (verdadeira/falsa) de erro e detecção (verdadeira/falsa) de erro. Os eventos para este espaço amostral são descritos por uma matriz 2×2 conhecida como tabela de contingência (Tabela 2). As frequências absolutas de cada evento são dadas por q_1, q_2, q_3 , e q_4 . A quantidade q_1

Tabela 2 - Tabela de contingência.

		Erros observados	
		Sim	Não
Erros detectados	Sim	q_1	q_2
	Não	q_3	q_4

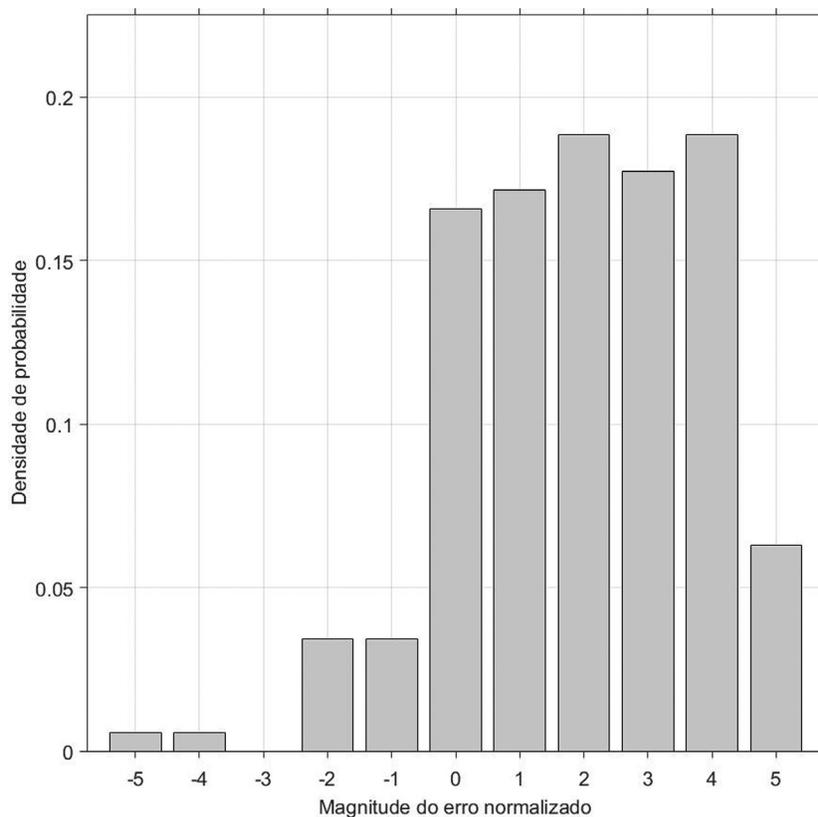


Figura 2 - Distribuição de probabilidade típica da magnitude de erros aleatórios utilizados no processo de semeadura. Este conjunto de erros foi gerado durante o processo de semeadura da estação de Blumenau.

representa o número de acertos do método, ou seja, medidas que foram semeadas com erro e marcadas corretamente como suspeitas. Os números q_2 e q_3 , representam, respectivamente, o número de alarmes falsos e de falhas para o método. O número da correta rejeição de erros (medidas livres da presença de erro) é dado por q_4 .

As métricas consideradas são: erros α do tipo I, erros β do tipo II, taxa de alarme falso (TAF), taxa de acerto crítico (TAC), e índice de sucesso crítico (ISC) definidas, respectivamente, pelas Eqs. (1) a (5):

$$\alpha = \frac{q_2}{q_1 + q_2 + q_3 + q_4} \quad (1)$$

$$\beta = \frac{q_3}{q_1 + q_2 + q_3 + q_4} \quad (2)$$

$$TAF = \frac{q_2}{q_2 + q_4} \quad (3)$$

$$TAC = \frac{q_1}{q_1 + q_3} \quad (4)$$

$$ISC = \frac{q_1}{q_1 + q_2 + q_3} \quad (5)$$

Como o processo de semeadura é de natureza aleatória, as equações apresentadas devem ser vistas como estimativas referentes a cada métrica. Desta forma, o valor médio de cada expressão representa um valor experimental para um processo de Monte Carlo de amostragem aleatória simples usando um conjunto de n_s amostras. A abordagem adotada apresenta duas vantagens: (i) o grau de precisão de uma métrica é tanto maior quanto maior for o número de amostras utilizadas; (ii) o experimento adotado permite compreender o processo de detecção de erros semeados a partir das probabilidades condicionais associadas a cada evento. Observa-se que esta abordagem probabilística bidimensional é comumente empregada no diagnóstico de doenças usando imagens de radiologia (Langlotz, 2003), e avaliação da qualidade das previsões meteorológicas (Thornes and Stephenson, 2001).

3. Resultados e Discussões

3.1. Avaliação do método de múltiplas Gamas

Considerando as estações descritas pela Tabela 3, foram utilizados, respectivamente, os seguintes parâmetros de probabilidade de corte e número de intervalos para o método de múltiplas Gamas: $p = 0.99$ e $n = 3$. Para cálculo das métricas, foi utilizado um número n_s , igual a 30 amostras e foram semeados cerca de 10% de erros em relação ao total de medidas de chuva para cada estação. Os erros normalizados introduzidos na série pertencem ao intervalo $[-5,5]$, sendo que a Tabela 3 apresenta os valores de desvio

Tabela 3 - Desvios padrões das séries históricas de cada estação e o erro máximo semeado.

Estação	Desvio padr. (mm/dia)	Erro máx. (mm/dia)
Blumenau	14,45	72,26
Urussanga	13,41	67,03
Ponte Serrada	17,71	88,56
Rio Campo	14,82	74,13
Joinville	14,76	73,79

Tabela 4 - Métricas obtidas para o algoritmo de múltiplas Gamas usando os parâmetros: $p = 0.99$ e $n = 3$.

Estação	α	β	TAF	TAC	ISC	p_{erro}
Blumenau	0,4038	0,0232	0,4464	0,7570	0,1447	0,0954
Urussanga	0,4139	0,0251	0,4575	0,7374	0,1382	0,0954
Ponte Serrada	0,3847	0,0237	0,4251	0,7506	0,1488	0,0951
Rio do Campo	0,4355	0,0238	0,4819	0,7524	0,1362	0,0962
Joinville	0,3858	0,0233	0,4265	0,7558	0,1498	0,0954
Mediana	0,4038	0,0237	0,4464	0,7524	0,1447	0,0954

padrão da distribuição de precipitação diária (considerando apenas dias chuvosos) e o erro máximo introduzido ($\pm 5\sigma_j$).

A Tabela 4 apresenta os resultados da aplicação dos métodos das múltiplas Gamas para o conjunto de parâmetros descritos. Observa-se que apesar do percentual de falhas β (erros do tipo II) ser baixo, a fração de medidas avaliadas inválidas, porém livres de erro, tende a ser considerável (valor mediano de erros α do tipo I igual a 40%). Isto sugere que o método é muito restritivo, pois considera um número excessivo de dados inválidos (mediana da taxa de alarme falso próximo a 44%). Estas observações são consistentes com os valores apresentados pelas métricas da taxa de acerto crítico e do índice de sucesso crítico, sendo este último mais sensível, pois leva em consideração a quantidade absoluta de falsos alarmes para cálculo do índice. O alto valor do TAC, próximo a 75%, e baixos valores de ISC, próximo a 14% estão relacionados com o fato de o método apresentar um elevado número de alarmes falsos. A última coluna apresenta a fração de erros p_{erro} efetivamente inseridos pelo processo de semeadura. Estes valores diferem ligeiramente do percentual previsto de 10% pois, como o processo de semeadura é aleatório, alguns erros foram introduzidos em datas onde as estações não possuíam o número mínimo de vizinhos para a análise do método.

As Figs. 3 e 4 apresentam uma análise do comportamento de TAF e TAC em torno da variação dos parâmetros do método de múltiplas Gamas. Os pontos traçados para estes gráficos foram obtidos calculando o valor mediano de TAF e TAC para as cinco estações. Foram utilizados valores de probabilidade de corte de 90, 95, 99, 99,5 e 99,9%, e

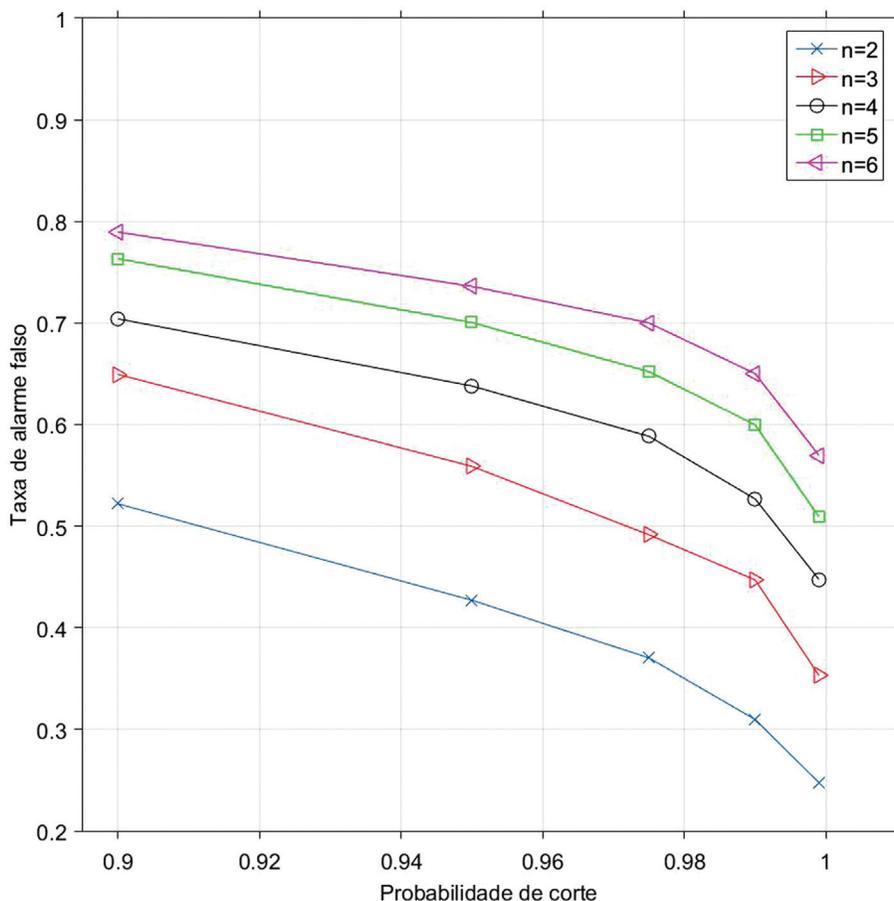


Figura 3 - Mediana da taxa de alarme falso (TAF) para todas as estações alvo, aplicando-se o método de múltiplas Gammas e diferentes combinações de parâmetros. Foram utilizados os seguintes valores de parâmetros para probabilidade de corte (p_{corre}): 90; 95; 97,5 99; 99,9%, e números de intervalos $n = 2, 3, 4, 5$ e 6 .

número de intervalos variando de 2 até 6. Observa-se que, de maneira geral, é possível reduzir as taxas de alarme falso aumentando-se as probabilidades de corte, porém isto implica também em uma redução das taxas de acerto crítico. Utilizando-se como exemplo o menor número de intervalos possível ($n = 2$) e uma probabilidade de corte de 99%, reduz-se as taxas de alarme falso para pouco mais de 30%, mas a taxa de acerto crítico fica acima de 65%. Para o conjunto de parâmetros utilizados (Tabela 4), os valores da taxa de acertos são sempre superiores às taxas de alarme falso, no entanto, os valores de TAF ainda podem ser considerados excessivos.

Adicionalmente, foi realizada uma análise mais detalhada da fração de acertos (TAC) para diferentes faixas de magnitude de erro introduzido. As seguintes classes foram consideradas para a análise: $[-5, -0,5)$, $[-0,5, 0,5)$, $[0,5, 1,5)$, $[1,5, 2,5)$, $[2,5, 3,5)$ e $[3,5, 5)$. A Fig. 5 apresenta o resultado da avaliação desta métrica de forma individual para cada uma das 5 estações pluviométricas em estudo. É possível ver que a fração de erros detectados cresce de forma aproximadamente linear a partir do valor de $-0,5$ de

erro normalizado. Para erros normalizados maiores ou iguais a 3,5 este índice fica próximo a 100% de acerto.

3.2. Novo método em modo sensível à detecção de erros

Utilizando os seguintes parâmetros: $f_1 = 0,25$, $m = 5$, e $f_2 = 0,25$ e tamanho de janela $d = 3$, número de amostras $n_s = 30$ e percentual de erros semeados de 10%, o método desenvolvido opera com uma sensibilidade maior à presença de erros semeados. Isto significa que é possível alcançar elevados valores da taxa de acerto (TAC mediano próximo a 87%), acompanhado de um valor considerável do número de alarmes falsos (TAF mediano próximo a 41%). Quando comparado ao método de múltiplas Gammas, o novo algoritmo apresenta um valor superior de TAC para uma valor semelhante de TAF. Os valores obtidos para o ISC são baixos com mediana próxima de 17% (influenciado principalmente pelo valor de TAF).

Para erros do tipo I e tipo II, são obtidos valores medianos próximos a 37% e 1% respectivamente. A Tabela 5 apresenta os resultados das métricas citadas.

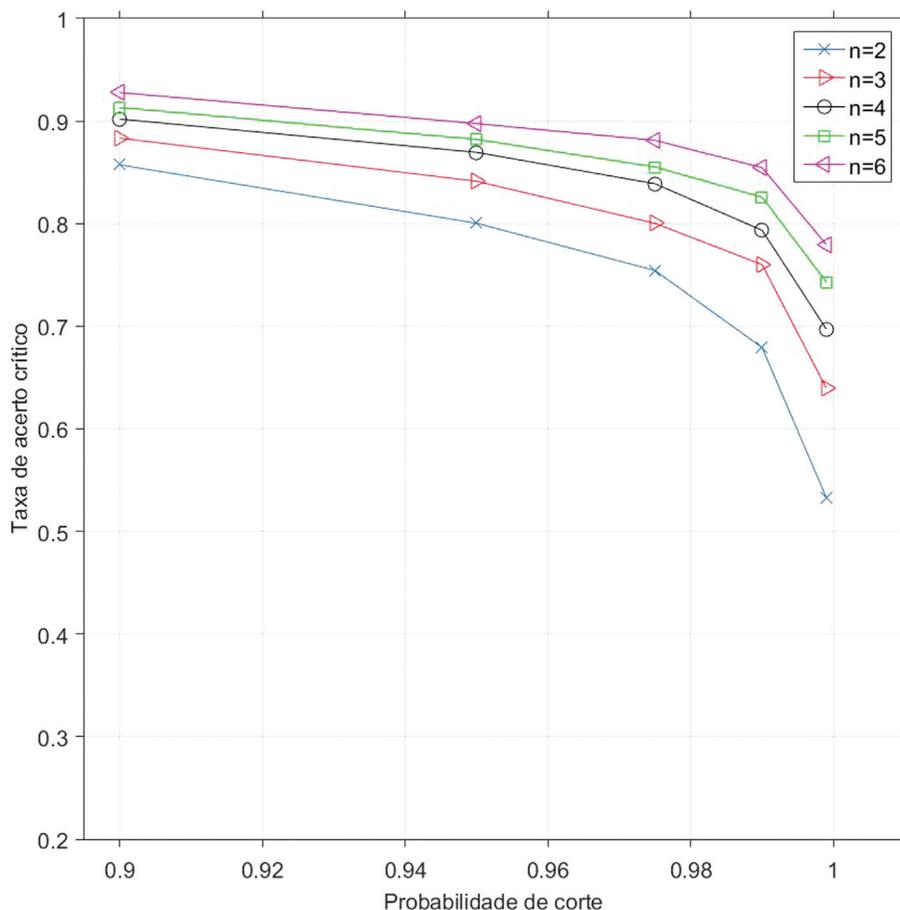


Figura 4 - Mediana da taxa de acerto crítico (TAC) para todas as estações alvo, aplicando-se o método de múltiplas Gamas e diferentes combinações de parâmetros. Foram utilizados os seguintes valores de parâmetros para probabilidade de corte (p_{corte}): 95; 97,5; 99; e 99,9%, e números de intervalos $n = 2$,

A Fig. 6 mostra a análise de TAC por faixas de magnitude de erro normalizado. O algoritmo neste modo de operação é capaz de detectar erros semeados em todas as faixas de magnitudes de erro, sendo que em todos os casos foram detectados mais de 50% do total de erros semeados.

3.3. Novo método em modo conservador (restritivo)

Demonstra-se em seguida, que o novo método permite operar também em um modo mais restritivo usando o seguinte conjunto de parâmetros: $f_1 = 2$, $m = 10$, e $f_2 = 1,5$, com o mesmo número de amostras n_s e tamanho de janela d . Essencialmente, neste modo prioriza-se cometer poucos alarmes falsos (TAF mediano próximo a 4%) mesmo que isso implique em taxas de acerto menores (TAC mediano próximo a 60%). Nesta situação, o valor mediano de ISC obtido é próximo a 44%.

A fração de erros do tipo I e II produziu valores medianos de cerca de 4% para ambos os casos. A Tabela 6 apresenta os resultados destas métricas.

A Fig. 7 apresenta a taxa de acerto por faixa de magnitude de erro introduzido para cada estação analisada. Verifica-se que o algoritmo foi muito mais seletivo em relação à

magnitude do erro: apenas erros semeados de magnitude superior a 0,5 foram detectados. Isto confere uma característica conservadora ao algoritmo, onde apenas erros de magnitude mais notável são detectados.

3.4. Curva de operação característica dos métodos

Além de medir as frações de erros/alarmes falsos para situações específicas, as métricas de TAC e TAF também podem ser utilizadas para representar de forma analítica o comportamento da detecção de erros semeados para cada um dos métodos. Segundo Metz (1978) esta informação pode ser extraída traçando o gráfico de dispersão das métricas TAF x TAC, para diferentes parâmetros, obtendo assim, a curva característica de operação do método.

Avaliou-se as medianas para TAF e TAC do grupo de estações em estudo, executando-se uma varredura de parâmetros para três algoritmos distintos: Gama simples, múltiplas Gamas, e o novo método desenvolvido.

Os dois primeiros algoritmos consideram 1 (gama simples), 2 e 3 intervalos de regimes de chuva distintos, empregando-se como probabilidades de corte valores variando de 0% até 100% (marcadores de azul para Gama

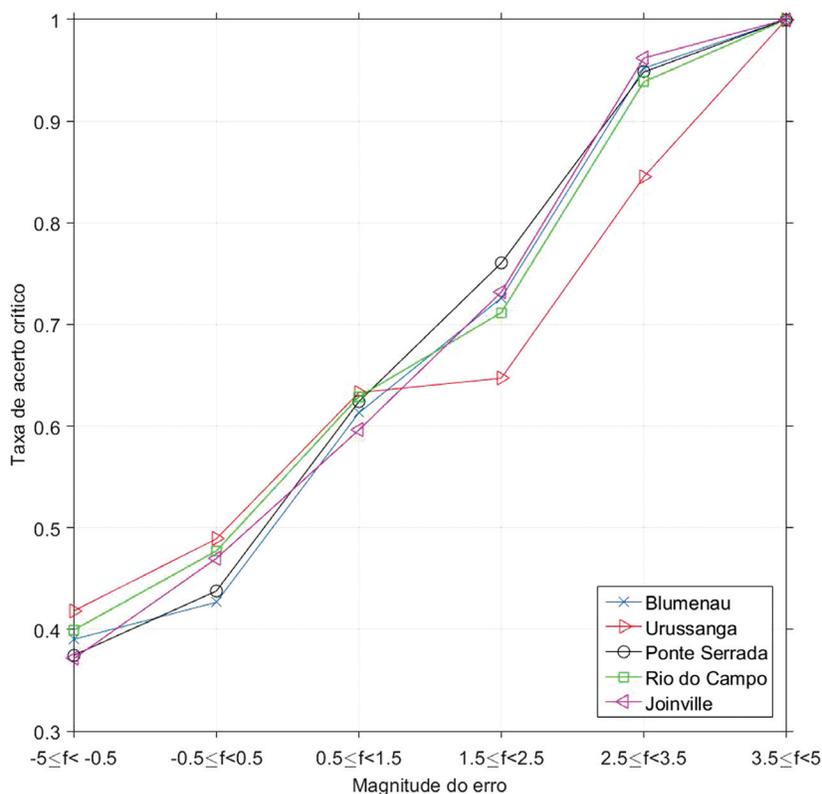


Figura 5 - Taxa de acerto crítico para diferentes estações em função da magnitude de erro introduzido (método de múltiplas Gamas). Para cálculo desta estatística, o seguinte conjunto de parâmetros foi utilizado: $p = 0,99$ e $n = 3$.

Tabela 5 - Métricas obtidas para o novo algoritmo em modo sensível à detecção de dados espúrios considerando os parâmetros: $f_1 = 0,25$, $f_2 = 0,25$, $m = 5$ e $d = 3$.

Estação	α	β	TAF	TAC	ISC	p_{erro}
Blumenau	0,3650	0,0123	0,4034	0,8715	0,1804	0,0953
Urussanga	0,3744	0,0130	0,4137	0,8628	0,1744	0,0949
Ponte Serrada	0,4534	0,0094	0,5010	0,9014	0,1561	0,0949
Rio do Campo	0,4180	0,0123	0,4622	0,8708	0,1621	0,0956
Joinville	0,3748	0,0128	0,4143	0,8662	0,1756	0,0953
Mediana	0,3748	0,0123	0,4143	0,8708	0,1744	0,0953

simples e preto e rosa para demais casos representados na Fig. 8). Na avaliação do novo método considerou-se variações de 0 até 2,4 desvios padrão, com parâmetro m variando de 5 até 15 (marcadores vermelhos exibidos na Fig. 8) e parâmetro d igual a 3. Em todas as situações, as simulações do processo aleatório de semeadura utilizaram 30 amostras.

O objetivo de construir gráficos usando as métricas descritas é de avaliar e encontrar pontos de operação dos algoritmos (Metz, 1978). O método é justificável quando suas taxas de acertos são superiores às de alarme falso. A reta de 45° (representada por linha tracejada na Fig. 8) indica, portanto, o limite inferior de operação dos algoritmos. Sobre essa reta, as taxas de acerto e de alarme falso se

Tabela 6 - Métricas obtidas para o novo algoritmo em modo conservador à detecção de dados espúrios considerando os parâmetros: $f_1 = 2$, $f_2 = 1,5$, $m = 10$ e $d = 3$.

Estação	α	β	TAF	TAC	ISC	p_{erro}
Blumenau	0,0333	0,0376	0,0365	0,6044	0,4486	0,0951
Urussanga	0,0300	0,0400	0,0331	0,5784	0,4398	0,0950
Ponte Serrada	0,0438	0,0328	0,0484	0,6567	0,4505	0,0957
Rio do Campo	0,0419	0,0367	0,0463	0,6152	0,4278	0,0955
Joinville	0,0347	0,0379	0,0383	0,6023	0,4416	0,0952
Mediana	0,0347	0,0376	0,0383	0,6044	0,4416	0,0952

equilibram e o algoritmo decide a presença ou não de erro de forma aleatória.

De acordo com essa condição, o método de múltiplas Gamas apresenta pontos de operação apenas para valores de TAF superiores a 20% (o uso do método para baixas taxas de alarme é impraticável, pois não foram registrados muitos pontos de operação abaixo da linha tracejada).

Outra característica das múltiplas Gamas é que o desempenho de detecção de erros semeados é melhor quanto menor for o número de intervalos adotados. Considerando o caso de uma única distribuição Gama, esta abordagem se mostrou melhor que o algoritmo de múltiplas

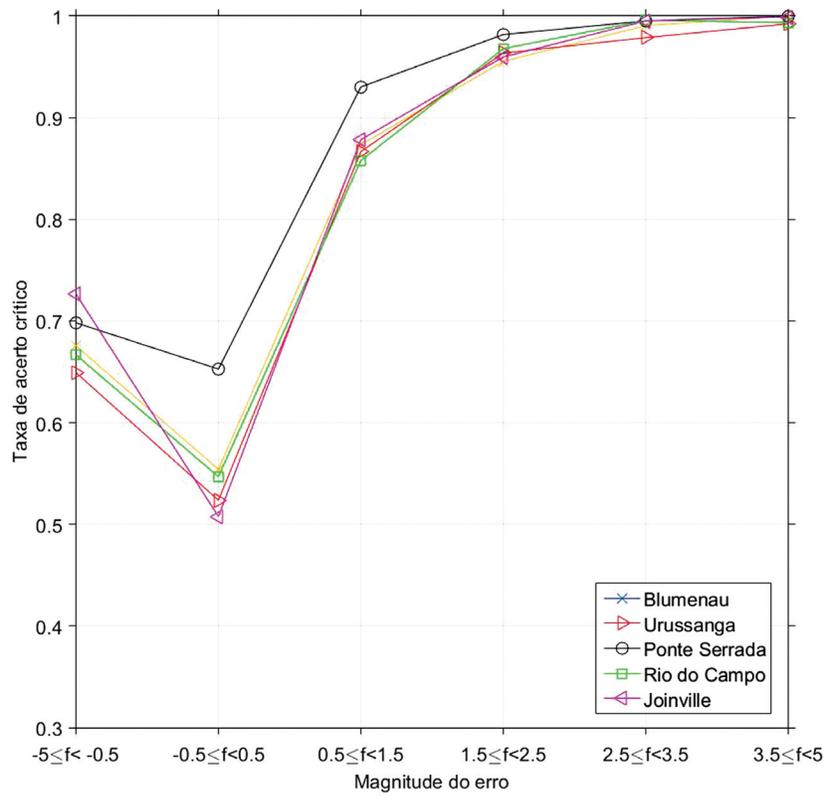


Figura 6 - Taxa de acerto crítico em função da magnitude de erro introduzido (novo método operando em modo sensível). Para cálculo desta estatística, o seguinte conjunto de parâmetros foi utilizado: $f_1 = 0,25$, $m = 5$, e $f_2 = 0,25$. O tamanho da janela para acumulados de precipitação é de $d = 3$.

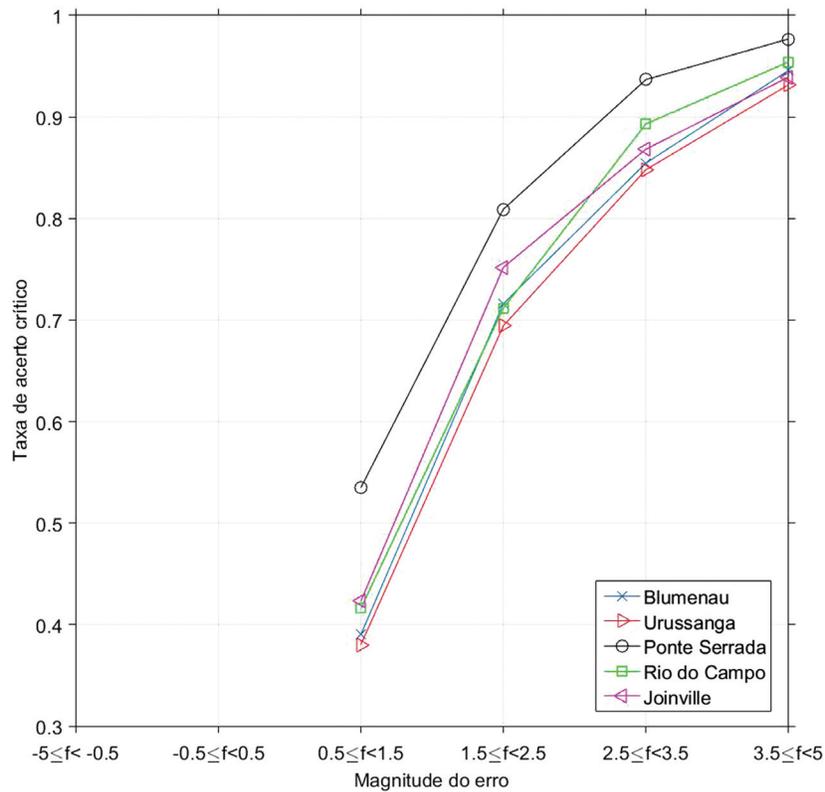


Figura 7 - Taxa de acerto crítico em função da magnitude de erro introduzido (novo método operando em modo conservador). Para cálculo desta estatística, o seguinte conjunto de parâmetros foi utilizado: $f_1 = 2$, $m = 15$, e $f_2 = 1,5$. O tamanho da janela para acumulados de precipitação é de $d = 3$.

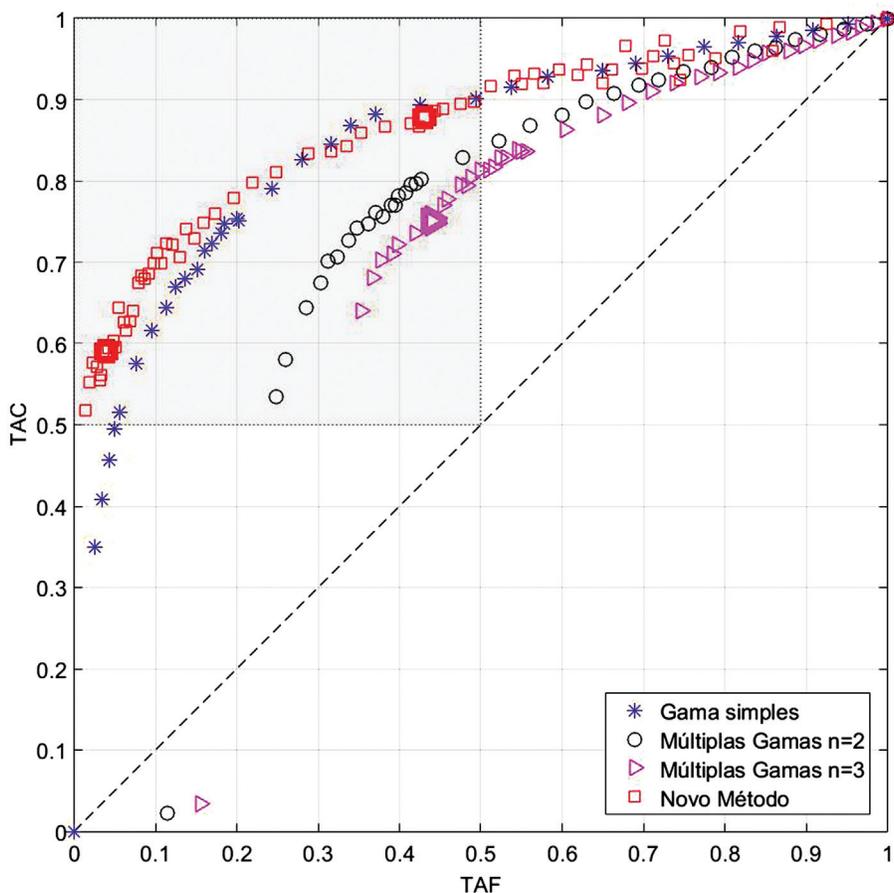


Figura 8 - Curva característica de operação para os métodos estudados. A linha tracejada indica um desempenho de detecção aleatório. Os valores das taxas de acerto crítico e de alarme falso representam valores medianos para o conjunto das cinco estações em estudo e obtidos por um processo de amostragem aleatória simples usando 30 amostras. Para o método de múltiplas Gamas variou-se de 0 a 100% a probabilidade de corte p . Para o novo método foram variados o número de desvios f_1 e f_2 de 0 a 2,4 e o parâmetro m de 5 a 15. O tamanho de janela considerado é de $d = 3$ dias. Os marcadores em destaque representam os pontos experimentais descritos anteriormente no trabalho. A área em destaque no quadrante superior esquerdo representa a região do gráfico para a escolha dos diferentes modos de operação.

distribuições Gama (permite operar para valores de TAF inferiores a 20%).

Por outro lado, o método desenvolvido é capaz de reproduzir o comportamento anterior e oferecer ainda melhores opções quando se deseja operar para baixos valores de TAF (inferiores a 25%). A variedade de pontos de operação (em toda a área destacada do primeiro quadrante na Fig. 8) da curva característica do novo método é particularmente importante quando se possui objetivos diferentes de qualificação de dados. Apesar de apresentar comportamento próximo a Gama simples em algumas situações, o método concebido garante um desempenho melhor para baixos valores de TAF.

Por exemplo, considere duas aplicações diferentes: (a) qualificação de um banco de dados e (b), consistência e comparação de dados com estimativas de satélite. No primeiro caso, é necessário um método que detecte dados espúrios cometendo poucos alarmes falsos, ou seja, detectando apenas dados espúrios notáveis. No segundo caso, o método deve priorizar poucas falhas, o que implica em

descartar uma quantidade maior de dados mantendo apenas a parte mais confiável.

4. Conclusão

Neste trabalho foram avaliadas duas técnicas para detectar dados espúrios de chuva usando acumulados diários obtidos a partir de estações localizadas no estado de Santa Catarina.

No método das múltiplas Gamas assume-se uma longa série histórica de chuvas (no trabalho original os autores consideram séries com até 30 anos de dados), homogeneidade espacial de chuvas diárias nas estações vizinhas, e que os diferentes regimes de chuva possam ser modelados por distribuições Gamas.

A análise para diferentes parâmetros (Fig. 8) usando o método das múltiplas Gamas sugere que o mesmo não é adequado ao objetivo proposto, possivelmente devido a quantidade limitada de dados utilizados no estudo (disponíveis de séries com até 15 anos de dados). Apesar de esta abordagem detectar grande parte de erros introduzidos de

forma aleatória, a taxa de alarmes falsos (métrica TAF) é excessiva (acima de 20%) e este comportamento independe da escolha de parâmetros.

A avaliação do método das múltiplas Gamas revela ainda a importância de serem analisadas diferentes métricas: a fração total de acertos (métrica TAC) não pode ser utilizada como único indicador conclusivo. Outra deficiência apontada por este método é que ele só pode ser aplicado para eventos de chuva, uma vez que a distribuição Gama desconsidera valores nulos de precipitação.

O novo método desenvolvido analisa os valores diários pontuais no espaço e também valores acumulados para um determinado período de dias. Esta abordagem apresentou melhores resultados: valores superiores de TAC quando comparados aos valores obtidos pelo método das múltiplas Gamas (pontos experimentais do método desenvolvido tendem a se localizar no eixo cartesiano acima do conjunto de pontos de múltiplas Gamas). Adicionalmente, o novo método permite operar em uma ampla faixa de valores de TAF, inclusive para percentuais abaixo de 10%.

Observa-se que as cinco estações analisadas tendem a apresentar um comportamento similar em relação ao percentual de erros detectados. A análise da métrica TAC, em função da magnitude dos erros aleatórios introduzidos, evidencia pelo menos duas formas distintas de detectar dados espúrios: frações significativas de erros detectados para todas as faixas de magnitude (modo sensível), e frações significativas de erros detectados somente para erros mais notáveis, ou seja, de magnitude normalizada acima de 1,5 (modo conservador).

Diferentemente de múltiplas Gamas, o novo método pode realizar também a qualificação de dados para situações em que não ocorreram eventos de chuva (medidas nulas de precipitação).

Finalmente é importante ressaltar que as métricas analisadas para a avaliação de desempenho dos métodos de consistência estão sujeitas a flutuações nos valores obtidos. A principal fonte destas flutuações são erros preexistentes ao processo de sementeira aleatória, que não foram detectados pela consistência básica inicial. Este problema é minimizado ao analisar diferentes estações e realizar um processo de amostragem de Monte Carlo simples.

Agradecimentos

Os autores gostariam de agradecer o suporte e apoio financeiro concedido pelo CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico) sem o qual o trabalho não poderia ter sido realizado. De mesma forma, um agradecimento é dirigido a EPAGRI/Ciram de Santa Catarina por fornecer acesso à base de dados e infraestrutura para desenvolver o presente estudo.

Referências

- DURRE, I.; MENNE, M.; GLEASON, B.; HOUSTON, T. and VOSE, R. Comprehensive automated quality assurance of daily surface observations. **Journal of Applied Meteorology and Climatology**, v. 49, n. 8, p. 1615-1633, 2010.
- EISCHEID, J.K.; BAKER, C.B.; KARL, T. and DIAZ, H. The quality control of long-term climatological data using objective data analysis. **Journal of Applied Meteorology**, v. 34, n. 12, p. 2787-2795, 1995.
- GONZÁLEZ-ROUCO J.F.; JIMÉNEZ J.L.; QUESADA V.; VALERO F. Quality Control and Homogeneity of Precipitation Data in the Southwest of Europe. **Journal of Climate**, Boston v. 14, p. 964-978, 2001.
- GORDON, N.D.; FINLAYSON, B.L.; MCMAHON, T.A. **Stream hydrology: an introduction for ecologists**. John Wiley and Sons, 2004.
- HABIB, E.; KRAJEWSKI, W.F.; KRUGER, A. Sampling errors of tipping-bucket rain gauge measurements. **Journal of Hydrologic Engineering**, v. 6, n. 2, p. 159-166, 2001.
- HAMADA, A.; ARAKAWA, O.; YATAGAI, A. An automated quality control method for daily rain-gauge data. **Global Environ. Res**, v. 15, n. 2, p. 183-192, 2011.
- HUBBARD, K.; GODDARD, S.; SORENSEN, W.; WELLS, N. and OSUGI, T. Performance of quality assurance procedures for an applied climate information system. **Journal of Atmospheric and Oceanic Technology**, v. 22, n. 1, p. 105-112, 2005.
- KUNKEL, K.; EASTERLING, D; HUBBARD, K.; REDMOND, K.; ANDSAGER, K; KRUK, M. and SPINAR, M. Quality control of pre-1948 cooperative observer network data. **Journal of Atmospheric and Oceanic Technology**, v. 22, n. 11, p. 1691-1705, 2005.
- LANGLOTZ, C.P. Fundamental Measures of Diagnostic Examination Performance: Usefulness for Clinical Decision Making and Research 1. **Radiology**, v. 228, n. 1, p. 3-9, 2003.
- METZ, C. E. Basic principles of ROC analysis. In: **Seminars in nuclear medicine**. WB Saunders, 1978. p. 283-298.
- NIE, S.-P.; LUO, Y.; LI, W.-P.; WU, T.-W.; SHI, X.-L. and WANG, Z.-Z. Quality Control and Analysis of Global Gauge-Based Daily Precipitation Dataset from 1980 to 2009. **Advances in Climate Change Research**, v. 3, n. 1, p. 45-53, 2012.
- THORNES, J.E. and STEPHENSON, D.B. How to judge the quality and value of weather forecast products. **Meteorological Applications**, v. 8, n. 03, p. 307-314, 2001.
- YOU, J.; HUBBARD, K.; NADARAJAH, S. and KUNKEL, K. Performance of quality assurance procedures on daily precipitation. **Journal of Atmospheric and Oceanic Technology**, v. 24, n. 5, p. 821-834, 2007.