Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Ana Carolina Sanches de Angelo

Ajuste de modelos aditivos generalizados a dados horários de precipitação de Morretes-PR assumindo a família Tweedie

Curitiba
2020

Ana Carolina Sanches de Angelo

Ajuste de modelos aditivos generalizados a dados horários de precipitação de Morretes-PR assumindo a família Tweedie

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Wagner Hugo Bonat

Curitiba
2020

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Ajuste de modelos aditivos generalizados a dados horários de precipitação de Morretes-PR assumindo a família Tweedie

Ana Carolina Sanches de Angelo[1]
Wagner Hugo Bonat[2]

Resumo

A chuva é um fenômeno relevante para a sociedade e sua modelagem é essencial para áreas como a agronomia e a gestão de risco de desastres. Contudo, as características particulares e não Gaussianas dos conjuntos de dados horários de precipitação - excesso de zeros e assimetria à direita - limitam a utilização de abordagens pautadas na distribuição normal. Sob estes aspectos, testa-se neste trabalho uma variação da família Tweedie (Compound-Poisson com $1 < p < 2$) na descrição probabilística da chuva considerando quatro limiares críticos. Isto é realizado por meio do ajuste de Modelos Aditivos Generalizados (GAMs), que vêm demonstrando aplicabilidades em estudos ambientais. Foram utilizados dados de uma estação meteorológica em Morretes - PR com dez anos de registros. A estes dados foram adicionadas novos atributos síntese com destaque para estações do ano, eventos de chuva agrupados e lags das variáveis originais. Utilizando o critério de Akaike foram realizados testes com diversos parâmetros de suavização e com variáveis identificadas previamente como mais correlacionadas à chuva. O melhor modelo ajustado descreve cerca de 45% do comportamento da precipitação. O fluxo de trabalho apresenta aplicabilidade nas áreas de interesse, porém os modelos necessitam de testes com mais variáveis, amostras e parametrizações para terem seu potencial confirmado, em especial no que diz respeito à predição. Espera-se que o prosseguimento de testes neste âmbito possa contribuir com o melhor aproveitamento dos dados horários de precipitação publicamente disponíveis no Brasil.

Palavras-chave: "precipitação horária, família Tweedie, Modelos Aditivos Generalizados, Compound-Poisson".

Abstract

Rainfall is an important phenomenon and its modeling is essential for sectors such as agriculture and disaster risk management. However, hourly precipitation data behaves in a non-Gaussian way, characterized by zero inflation and skewness to the right. In order to address this anomalies, in this article we analyse the adoption of Compound-Poisson ($1 < p < 2$) variation of the Tweedie family to describe the probability of exceedance of four rainfall thresholds. This was achieved through adjusting Generalized Additive Models (GAMs), which has successfully been used in several environmental studies. The analysed data set consists of a ten year sequence of hourly registers from a meteorological station in Morretes (State of Paraná - Brazil). Further attributes were added to the data set including: season descriptions, grouped rainfall events and lagged versions of the original variables. Akaike's criteria was applied to compare the effects of both the variables more correlated with rainfall and of different smoothing parameters. The better adjusted model describes approximately 45% of rainfall variation. The workflow shows applicability to the interested sectors, but the models require further testing in order to confirm its potential, especially with regards to prediction. It is expected that the pursuance of tests in this field will contribute to expand the usage of publicly available hourly rainfall data in Brazil.

Keywords: *"hourly rainfall data, Tweedie family, Generalized Additive Models, Compound-Poisson".*

1 Introdução

A precipitação é relevante em diversos processos ambientais e varia espacial e temporalmente de forma dinâmica. Apesar de existirem métodos mais avançados para o monitoramento desta variável, os dados pontuais de postos pluviométricos seguem sendo a fonte mais popular de informações de pluviosidade no Brasil. Isto ocorre porque são disponibilizados abertamente pelo Instituto Nacional de Meteorologia, pela Agência Nacional de Águas e por agências regionais, o que torna viável a utilização desta informação por setores acadêmicos e de gestão pública menos especializados em relação à meteorologia.

Apesar de utilizados por cientistas e sociedade, os registros das estações meteorológicas ainda são subaproveitados, em especial aqueles de resolução horária. Ainda é muito popular a utilização de métodos como a desagregação de dados diários para estudos de fenômenos subdiários [1]. Isto ocorre também pela distribuição espacial das estações. Há um tradeoff de potencialida-

[1]Geógrafa, aluna do programa de Especialização em Data Science & Big Data, sanchesdeangelo@gmail.com.
[2]Professor do Departamento de Estatística - DEST/UFPR, wbonat@gmail.com.

des entre as estações automáticas e as convencionais. As primeiras dispõem de melhor resolução temporal com registros a cada hora ou alguns minutos, a despeito de contarem com menor série histórica. E as últimas são mais antigas - eventualmente remontando ao século XIX - e apesar de terem leitura manual apenas uma ou duas vezes ao dia, são melhor distribuídas pelo território. Mesmo com a expansão da rede brasileira de monitorameto hidrometeorológico nos últimos 20 anos [2, 3], as séries de dados oriundas de estações automáticas permanecem pouco exploradas do ponto de vista da modelagem probabilística, mesmo que já contemplem períodos maiores a partir dos quais se pode inferir mais seguramente acerca de comportamentos, tendências e correlações entre os dados observados.

Portanto, os dados horários têm um potencial informativo que precisa ser avaliado especialmente sob dois aspectos de interesse geral: as relações entre as variáveis na escala subdiária e a probabilidade de ocorrência de eventos de magnitudes específicas nesse recorte de tempo. A frequência horária de registro e análise é importante considerando que alguns tipos de eventos extremos de grande impacto para a sociedade tendem a ocorrer ou ter estopins que os dados diários não são capazes de discretizar. Isto é relevante, por exemplo para estudos que buscam a interpretação da recorrência eventos extremos de chuva e de outros processos hidrológicos influenciados por ela.

Além do debate acadêmico, atividades associadas à gestão de risco de desastres e obras de engenharia também podem se beneficiar de cenários criados a partir dos dados horários, âmbitos nos quais ainda é muito comum a utilização de dados diários, apesar da imprecisão associada. A modelagem pode contribuir neste sentido, porém ainda apresenta desafios. Os dados horários de precipitação não se enquadram na "normalidade"assumida pelas abordagens estatísticas básicas. Dessa forma, o objetivo deste trabalho é avaliar o uso de modelos aditivos generalizados sob a premissa de que a distribuição da chuva como variável resposta é uma distribuição da família tweedie. O presente trabalho analisa estes modelos em relação ao entendimento de relações entre as variáveis monitoradas e na previsão probabilística de eventos de chuva. Constam na literatura algumas abordagens próximas envolvendo dados de precipitação, com bons resultados [4, 5, 6, 7]. Porém, dados de resolução horária permanecem inexplorados sob esta perspectiva, o que foi buscado neste trabalho.

Este artigo está segmentado de maneira que esta introdução abordou o contexto do trabalho, suas motivações e objetivos, assim como os problemas de pesquisa. Na sequência, a revisão bibliografica ilustra aspectos teóricos das abordagens estatísticas na hidrologia e ciências correlatas, dos modelos lineares generalizados e da família tweedie de distribuições. A seção de métodos traz o conjunto de dados que foi utilizado, a linguagem de programação, o ambiente de desenvolvimento do trabalho e os pacotes utilizados desde a limpeza e organização dos datasets até a construção dos modelos. A sessão de resultados traz os produtos gerados neste trabalho, e a conclusão traz as reflexões feitas sobre a qualidade destes produtos, a aplicabilidade do método utilizado, as limitações e as demandas de continuidade da investigação.

2 Revisão Bibliográfica

As dificuldades e custos na obtenção de dados suficientes para a descrição dos fenômenos hidrometeorológicos fazem com que esta área da ciência utilize técnicas estatísticas em âmbitos diversos [8]. Parte fundamental do ciclo hidrológico, a chuva é um fenômeno em que interferem fatores físicos de circulação da atmosfera desde a escala local até a global. Instrumentos mais avançados de monitoramento da precipitação são os sensores remotos embarcados em satélites e modelos globais de circulação atmosférica, que fornecem dados quantitativos e espacializados. Entretanto, no Brasil os dados mais utilizados no estudo da precipitação em âmbito interdisciplinar são os dados pontuais oriundos de estações meteorológicas ou postos pluviométricos. A chuva é monitorada pela captação do volume precipitado em determinado período de tempo, informação que é armazenada em um datalogger e geralmente é apresentada em milímetros por hora. Enquanto fenômeno, a chuva é considerada como parcialmente determinística e parcialmente aleatória. Entretanto, diante da complexidade e da dificuldade do monitoramento de todos os fatores intervenientes, é recorrente tratá-la como uma variável totalmente aleatória, quando apresenta-se na forma de dado [9].

Os dados de precipitação consistem de valores reais iguais ou acima de 0.0; e de valores ausentes - no caso de falhas nas estações. As séries de precipitação tendem a ser infladas de zeros em virtude dos recorrentes passos de tempo em que a chuva não ocorre. Além disso, tendem a apresentar assimetria à direita, com a maior parte dos dados correspondendo a valores próximos de zero e um, e com a presença de eventuais valores extremos. No caso das séries de precipitação horária esta característica é ainda mais forte do que em dados registrados em frequência diária. Estes aspectos implicam na não-normalidade das séries pluviométricas e tornam inaplicáveis as distribuições de probabilidade mais comuns como a normal. Nesse sentido, utilizam-se diversas distribuições teóricas que auxiliam na identificação da probabilidade de eventos de determinada magnitude. Este tipo de estudo é utilizado amplamente em obras de engenharia, em sistemas de alertas, dentre outras aplicações na gestão de recursos hídricos [8]. Ainda assim, sua modelagem adequada é um tema de discussão recorrente, com constantes mudanças de paradigma no âmbito científico, nem sempre acompanhadas por suas aplicações práticas [10].

Devido ao desafio apresentado pelas particularidades dos dados horários de precipitação, a família Tweedie apresenta-se como uma possível alternativa para descrição de seu comportamento. Esta família de distribuições

pertence à classe dos modelos de dispersão exponencial sintetizada por Jorgensen [11] em menção ao trabalho de Tweedie [12]. Ela é caracterizada por um parâmetro de dispersão (ϕ) e um parâmetro de potência (p). Sendo Y uma variável aleatória com distribuição Tweedie tem-se que $E(Y) = \mu$ e variância $Var(Y) = \phi\mu^p$.

Cada valor do parâmetro de potência p nesta família corresponde a um modelo específico. Por exemplo, $p = 1$ corresponde à distribuição de Poisson; $p = 2$ corresponde à distribuição Gama e $p = 3$ corresponde a distribuição Normal Inversa. Neste contexto, o modelo *Compound Poisson* obtido quando p está entre 1 e 2 é o de melhor ajuste a dados com grande quantidade de zeros e valores contínuos positivos [4]. A função de densidade da família Tweedie é descrita pela seguinte equação:

$$f(y;\theta;\phi) = a(y,\phi)\exp\left\{\ \tfrac{1}{\phi}(y\theta - k(\theta))\ \right\}, \qquad (1)$$

em que $\mu = E(Y) = K'(\theta)$ é a esperança da distribuição, $(\phi) > 0$ é o parâmetro de dispersão, (θ) é o parâmetro canônico e $k(\theta)$ representa a função cumulante.

Para valores contínuos e positivos com excesso de zeros, a probabilidade de observar um evento igual a zero é dada por:

$$Pr(Y = 0) = \exp(-\lambda) = \exp\left\{-\frac{\mu^{2-p}}{\phi(2-p)}\right\}. \qquad (2)$$

Conforme Bonat e Kokonendji (2017) [7], a estimativa dos parâmetros da Tweedie a partir de métodos usuais como o da máxima verossimilhança é dificultada por uma série de restrições, inclusive computacionais. Contudo, apontam também para a adequação do método da quasi-verossimilhança nesta tarefa. Com base nisto, alguns pacotes computacionais recentes já dispõem de maneiras orimizadas de contornar este problema.

Modelos de regressão simples e modelos polinomiais não permitem o estabelecimento de distribuições alternativas para a variável resposta, o que impõe uma restrição em sua aplicação aos dados de chuva com o uso da Tweedie. Além disso, não contemplam as relações não lineares entre variáveis monitoradas em estações pluviométricas. Já modelos baseados em aprendizagem de máquina tem boas perspectivas de sucesso na modelagem de precipitação, contudo, não possibilitam explicações claras da análise de relações entre os parâmetros e tampouco permitem o cálculo de probabilidades. A incerteza é um tema muito caro à hidrologia porque modelos determinísticos dificilmente explicam as complexas relações do ciclo hidrológico sem alguma estimação ou extrapolação. Nas últimas décadas quaisquer fenômenos analisados e sobretudo preditos devem ser abordados na hidrologia com clareza acerca das incertezas envolvidas, para o que o cálculo de probabilidades é indispensável [13, 14].

Diante destes fatores, os Modelos Aditivos Generalizados (GAMs) apresentam-se como opção adaptável para ajuste de um modelo a estes dados, considerando a já estabelecida definição da família Tweedie como distribuição da variável resposta (precipitação). Conforme Wood (2020)[15], os modelos aditivos generalizados têm apresentado capacidade de representação da complexidade de algumas relações, sem perda de interpretabilidade. Definidos por Hastie e Tibshirani em 1986[16], os GAMs oferecem uma maneira flexível para a identificação de efeitos não lineares substituindo o preditor linear comum presente nos GLMs por uma soma de funções suaves. A partir de uma analogia, Wood[17] define um GAM como "um glm em que o preditor depende linearmente de funções de suavização das variáveis preditoras". A equação dos preditores lineares nos GAMs é então alterada para a seguinte forma:

$$\eta = s0 + \sum_{j=1}^{p} s_j(X_j), \qquad (3)$$

em que η é o preditor linear; X_j são os valores de j covariáveis e $s_j(X_j)$ representam funções suaves das covariáveis.

As funções suavizadoras são compostas por somas de funções bases que se assemelham à forma do ajuste polinomial. A complexidade do modelo é então amenizada através de termos adicionais de penalização para porções da série de dados que são definidas pelo usuário ou por otimização ("knots"). A forma mais comum de aplicação desta suavização é por meio de splines. Isto torna os GAMs flexíveis, além de incrementarem seu poder de generalização/predição.Esta flexibilidade é um fator que vem induzindo a popularidade destes modelos em diversos âmbitos da ciência, com destaque para análises das complexas relações entre variáveis ambientais [17, 18, 19].

3 Materiais e Métodos

Estão resumidas neste parágrafo as tarefas envolvidas na tentativa de uso dos GAMs para a identificação de relação entre as variáveis monitoradas e a chuva, bem como para a estimativa de probabilidades de eventos. Primeiramente os dados foram adquiridos e formatados para viabilizar o uso dos modelos. Depois, foram criadas variáveis adicionais a partir dos dados existentes. Foram estudadas as importâncias dos atributos originais e dos atributos criados do dataset. Então ajustou-se modelos com as variáveis identificadas como mais relevantes. Dois modelos com melhor desempenho foram selecionados. Foram definidos limiares de precipitação para estudo das suas probabilidades de excedência a cada passo de tempo. Estas probabilidades foram estimadas a partir das predições realizadas pelos modelos. Estas probabilidades foram comparadas às observações efetivas na base de treino. Todas as operações envolvidas neste trabalho foram realizadas por meio da linguagem R e de pacotes associados. O fluxograma na Figura 1 descreve resumidamente estas tarefas.

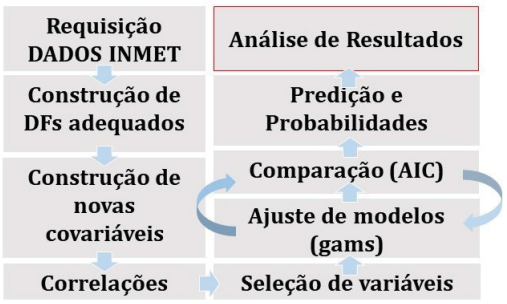

Figura 1: Fluxograma metodológico, com destaque em vermelho para o estado atual do trabalho

Variável	Unidade
Precipitação	mm/h
Pressão Atmosférica	hPa
Temperatura	ºC
Radiação Global	KJ/m^2
Umidade Relativa do Ar	%
Direção do Vento	º
Velocidade do Vento	m/s

Tabela 1: Variáveis contidas no conjunto de dados original (INMET, 2019)

3.1 O conjunto de dados

Os dados analisados pertencem a uma série horária de registros meteorológicos da estação Morretes A873 localizada no município homônimo no Paraná e operada pelo INMET - Instituto Nacional de Meteorologia. A série original era composta por dois arquivos de formato ".xls", com dias apresentados verticalmente e horas horizontalmente, com blocos de intervalos de 24 horas para cada variável. Cada bloco foi disposto lado a lado nas colunas. Os registros da série abrangem o período entre 12 de março de 2008 às 21 horas e 31/12/2019 às 20 horas. As variáveis registradas e as unidades de medida correspondetes estão dispostas na Tabela 1.

A Figura 2 demonstra os dados da série para a variável precipitação como forma de exemplo do conjunto disponível.

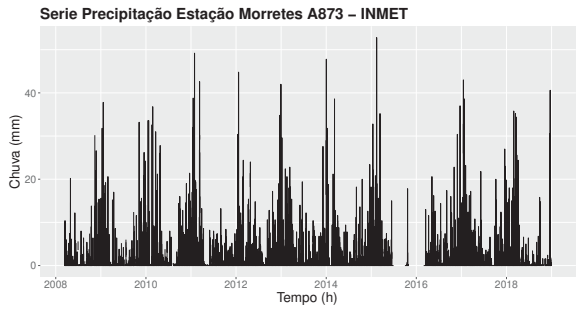

Figura 2: Serie Completa de Precipitação Horária

O manejo inicial dos dados implicou na transformação para o formato "tripa", voltada à análise de variações no tempo. Isto demandou a pivotação, o ajuste de datas do formato UTC para GMT-3 e a união de conjuntos de

Variáveis criadas	Unidade
Acumulado de Precipitação	mm
Data de início e fim	datetime
Intensidade máxima no evento	mm/h
Duração	horas
Eventos singularizados	-
Estação do Ano	-
Lags Pressão 1-5, 10, 15 e 30 horas	hPa
Lags Chuva 1-5, 10, 15 e 30 horas	mm
Lags Radiação 1-5, 10, 15 e 30 horas	kJ/m^2
Lags Umidade 1-5, 10, 15 e 30 horas	%
Lags Temperatura 1-5, 10, 15 e 30 horas	ºC
Lags Velocidade do Vento 1-5, 10, 15 e 30 horas	m/s

Tabela 2: Variáveis criadas adicionalmente no conjunto de dados original (INMET, 2019)

variáveis originalmente separados. Foram encontradas falhas de monitoramento (dados ausentes), porém, sem a presença de outliers. Na sequência foram programadas funções para criação de novas variáveis como: o valor acumulado desde o primeiro passo de tempo em que a chuva superou 0,2 mm - o que permitiu a singularização dos eventos de precipitação; o período entre o início e o fim da chuva; a codificação dos eventos de chuva, as estações do ano conforme as datas e os lags para cada variável.

Dessa forma, o primeiro conjunto de dados passou a contar adicionalmente com as variáveis descritas na Tabela 2. Ao final, o conjunto de dados dispunha de 103340 observações de registros horários e 64 covariáveis.

3.2 Análise de correlações

Os resultados das análises exploratórias mais relevantes estão dispostos nos gráficos a seguir. Eles indicam sobretudo as tendências sazonais da variável central. Para a estação meteorológica analisada, o verão apresenta-se como um período de maior pluviosidade acumulada e intensidade e o inverno como um período de precipitações menos intensas porém mais duradouras, como é característico da região.

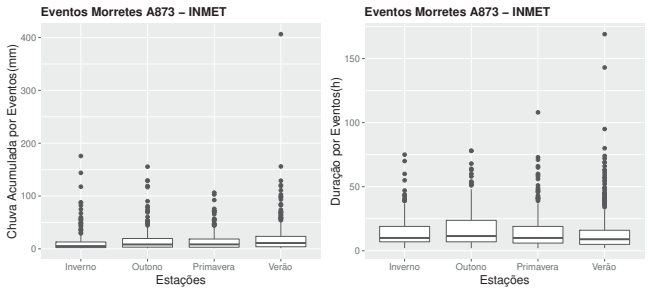

Figura 3: Boxplots de ilustração das frequências da variáveis Precipitação acumulada (mm) e Duração (mm)

Já a hora de ocorrência apresenta uma tendência de chuvas mais intensas a partir da tarde sobretudo no verão, conforme indica a Figura 4). Estas informações

são relevantes para a posterior definição de variáveis a serem consideradas no modelo.

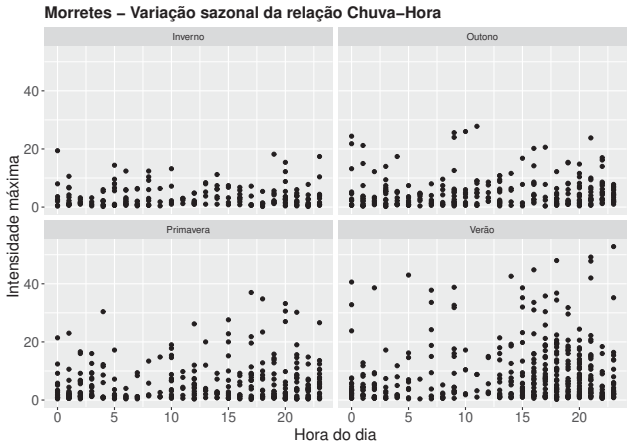

Figura 4: Distribuição da variável precipitação horária nas 4 estações. (Chuva (mm) vs. Hora do dia de 0 a 23.)

Um dos correlogramas obtidos pode ser observado na Figura 5. Ele demonstra não haver forte correlação entre as variáveis primárias registradas pela estação e os valores de precipitação observados.

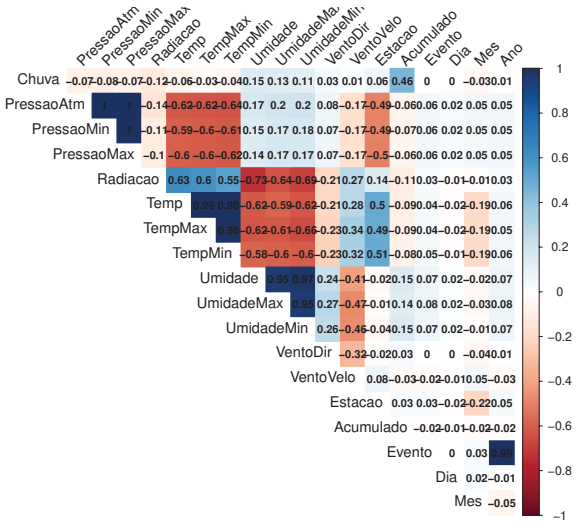

Figura 5: Correlações entre as variáveis do conjunto de dados observados

Para além das correlações entre variáveis intrínsecamente associadas como valores máximos, mínimos e médios de umidade, pressão ou temperatura, por exemplo, a única variável que apresentou correlações entorno de 0,50 foi a variável "estação do ano"em relação a aspectos como umidade e pressão. Para a variável de maior interesse - a chuva - correlações expressivas não foram observadas, exceto no caso dos valores de chuva acumulada e de defasagem de chuva.

Também foi criado um correlograma com todas as defasagens de tempo criadas (lags), conforme exibe-se na Tabela 3. Por limitação de espaço aparecem na tabela apenas as desafasagens das 3 primeiras horas.

	Acum	Estacao	LagP1	LagP2
Chuva	0.46	-0.1	-0.05	-0.05
Acum	1	-0.11	-0.06	-0.06
	LagTemp2	LagTemp3	LagVelo1	LagVelo2
Chuva	-0.04	-0.03	-0.04	-0.04
Acum	-0.1	-0.1	-0.09	-0.09
	LagC2	LagC3	LagC5	LagC15
Chuva	0.63	0.43	0.18	0.05
Acum	0.55	0.56	0.53	0.36
	LagU2	LagU3	LagRad3	LagC1
Chuva	0.17	0.15	-0.07	0.82
Acum	0.25	0.25	-0.13	0.51
	LagVelo3	LagC1	LagP3	LagTemp1
Chuva	-0.03	0.82	-0.05	-0.05
Acum	-0.09	0.51	-0.06	-0.1
	LagC30	LagU1	LagRad1	LagRad2
Chuva	0.03	0.19	-0.09	-0.08
Acum	0.11	0.25	-0.11	-0.12

Tabela 3: Correlações entre as variáveis do conjunto de dados de lag (deslocamento no tempo)

Os correlogramas subsidiaram a eleição de algumas variáveis mais importantes, dentre as originais e do novo dataset, para o ajuste e o teste dos primeiros modelos aditivos generalizados em caráter de experimentação.

3.3 Modelos empregados

Para a construção dos modelos utilizou-se o pacote `mgcv` [20]. Este pacote dispõe de diversas possibilidades de aplicações dos GAMs, com destaque para a implementação de algoritmos capazes de estimar alguns parâmetros por otimização, dentre eles o grau de suavização (lambda). Esta possibilidade trazida representou grande avanço para o uso dos GAMs, e é mais comumente realizada por meio de métodos como a validação cruzada generalizada (GCV), e as máximas verossimilhanças comum (ML) e restrita (REML). Outra potencialidade do pacote é a variedade de distribuições possíveis de serem assinaladas à variável resposta, dentre as quais constam versões da família Tweedie. Ainda, ele viabiliza também a escolha do parâmetro p por otimização para $1 < p < 2$, condição adequada para os dados de chuva. Dessa forma, foram testados ajustes dos GAMs para várias combinações entre as covariáveis originais e criadas e a variável resposta - a chuva horária. A função de ligação selecionada foi a logaritmica. O parâmetro p é estimado simultâneamente com a definição da restrição entre 1.01 e 1.99.

As funções de suavização foram variadas também para fins de teste, eventualmente considerando parâmetros de ciclicidade ou repetição como agrupamentos dos nós de cálculo (parâmetro "by"da suavização).

Inicialmente foram criados 50 modelos com apenas uma variável explicativa para o estabelecimento de um critério de exclusão das variáveis. O método de estimação escolhido foi o de máxima verossimilhança restrita (REML), que apresenta-se na literatura como menos propenso a estimações errôneas induzidas por mínimos locais [20]. A comparação entre os modelos foi realizada por meio do critério de Akaike (AIC) [21] para cada grupo de variáveis, começando pelas categóricas,

para as quais não foram assinaladas suavizações. A Figura 6 demonstra os principais resultados desta etapa para as variáveis não contínuas e menos associadas a parâmetros físicos.

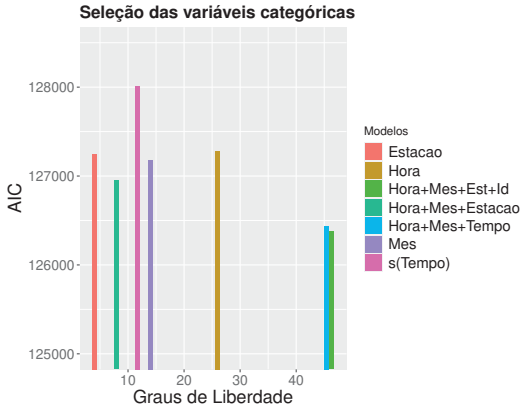

Figura 6: Análise da importância das variáveis categóricas pelo critério de Akaike

Posteriormente foram criados modelos com os parâmetros físicos de maneira individualizada, e conforme se demonstraram importantes na explicação da chuva, foram considerados candidatos de presença no modelo final. As comparações realizadas pelo método AIC com os parâmetros físicos podem ser conferidas na Figura 7.

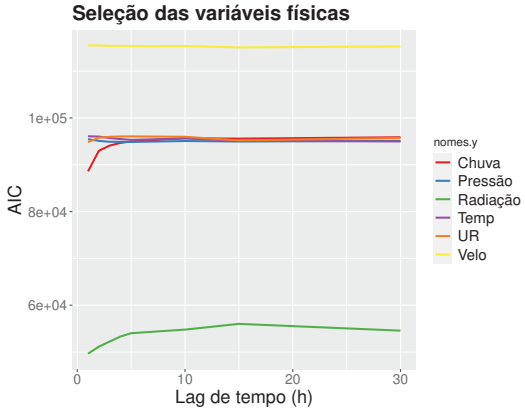

Figura 7: Análise da importância das variáveis físicas pelo critério de Akaike

Então foram testados modelos com mescla entre as variáveis por adição gradativa das mesmas. Dentre estes últimos modelos, mais elaborados, foram selecionados dois mediante o critério AIC.

A base de dados foi separada em período pré e pós 2019. Os modelos foram ajustados com base no período entre 2008 e 2018. A base de testes (ano de 2019) foi separada por meses e dias, para acompanhamento da capacidade preditiva em um período de tempo relevante para a gestão pública e que possibilitasse melhor visualização dos resultados.

A partir do ajuste dos dois modelos de melhor desempenho foram feitas as predicões em relação à chuva para cada passo de tempo. Foram escolhidos quatro limiares de chuva para estimativa da probabilidade de excedência conforme a família tweedie. Os limiares escolhidos foram 5, 10, 20 e 40 milímetros. Para cada passo de tempo obteve-se um o valor da probabilidade de excedência desse limiar. Para o estabelecimento de cenários probabilisticos foi utilizado o pacote `tweedie` [22]. Foram criados então datasets com o resultado das predições, os valores reais obtidos por monitoramento em 2019 e as probabilidades de excedência para os 4 limiares estabelecidos conforme a tweedie.

4 Resultados e Discussões

As análises exploratórias não apontaram correlação expressiva entre as variáveis observadas ou derivadas na estação Morretes A873. A Tabela 3 referente aos lags também demonstra que as correlações das variáveis defasadas reduzem conforme se regride no tempo confirmando um aspecto esperado de que valores registrados em um momento tendem a estar mais relacionados aos registros ocorridos em passos de tempo mais próximos. Uma exceção é a radiação que está relacionada à incidência de sol, podendo ter maior correlação com dados do dia anterior do que de horas contíguas se as mesmas corresponderem à noite. Ainda que baixas, as maiores correlações indicaram que além da própria precipitação, a radiação, a umidade, a estação do ano e o evento de chuva individualizado consistem nos atributos com maior potencial de explicação do comportamento da precipitação nesta série. A partir disso, algumas variáveis foram priorizadas na composição dos modelos, assumindo também os aspectos físicos da precipitação e as interações não lineares entre elas, possívelmente captadas por meio da modelagem com GAMs. Dentre as variáveis que se revelaram muito pouco importantes na variação da resposta nos modelos suavizados, constam a pressão e a velocidade do vento, que foram pouco ou não consideradas nos modelos finais. Fatores como a hora do dia e a estação do ano apresentaram relevância em relação aos volumes precipitados e a duração das chuvas, conforme os boxplots apresentados anteriormente 73. Além disso, estes fatores não físicos e discretos apresentaram importância conjunta quando modelados e analisados pelo AIC 6, por isso foram mantidos nos modelos como preditores lineares. Outras variáveis demonstraram ter influência expressiva sobre os resultados dos modelos testados, como é o caso da radiação e de seus lags, e dos legs de precipitação. Estas constatações nortearam as escolhas das variáveis para os modelos finais. A Tabela 4 apresenta os valores do AIC para os melhores modelos de cada variável individualmente segundo seu Lag (quando aplicável).

A Figura 8 ilustra a relação entre desempenho e graus de liberdade para alguns dos modelos analisados, e permite a comparação entre modelos simples e compostos. Os modelos nomeados com nomes de atributos (e.g. Chuva, Radiação, Pressão) representam modelos que utilizaram apenas uma variável para predição. Os modelos com nomes de siglas, são composições. Os mode-

Modelo	AIC
Radiacao1	49649.14
Chuva Lag 1	88592,78
Temperatura Lag 1	9497565
Radiação Lag 1	49649,14
Umidade Lag 1	94849,14
Pressão Lag 5	94862,06
Velocidade do Vento	115273,5

Tabela 4: Lags com os GAMS de melhores AICs dentre as variáveis analisadas em relação a seus grupos

Modelo	DF	AIC
R42Evento	106.24	25304.18
R42CC	94.21	36138.39

Tabela 5: Dois melhores modelos e suas composições de variáveis

los que consideraram somente uma variável tem maior valor de AIC e menos graus de liberdade, concetrando-se no canto superior esquerdo do gráfico. Os modelos com mais variáveis têm mais graus de liberdade, acompanhado em geral de melhor desempenho - no canto inferior direito. Aqueles que utilizaram o custoso mecanismo de "By"como fator juntamente na suavização, apresentam também bons AIC, mas aumentam exageramente os graus de liberdade, assim como o tempo de processamento, que excedeu 2 horas em alguns casos.

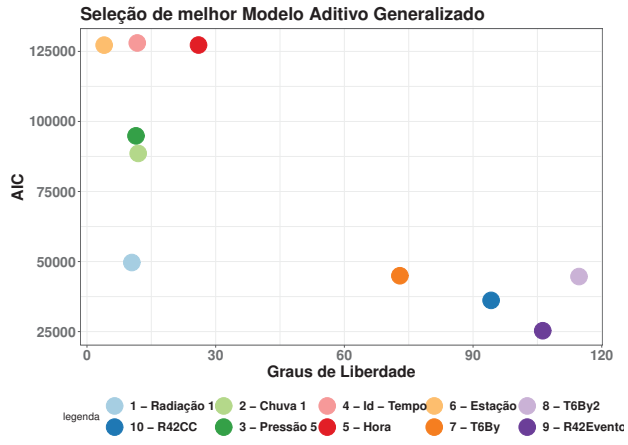

Figura 8: Modelos mesclados - critério de Akaike

Observa-se que modelos de melhor desempenho apresentam resultados parecidos para o AIC. Dentre estes modelos, foram escolhidos dois para a predição e para estimativa das probabilidades conforme os limiares escolhidos. A Tabela **??** traz informações para os dois modelos - R42CC, mais focado nos lags de Chuva e na Radiação; e R42Evento que é similar, porém considera maior heterogeneidade de variáveis com a inclusão de temperatura e pressão e, principalmente, da categoria Evento, que separa os eventos de precipitação. A coluna "DF"corresponde aos valores de graus de liberdade.

Os modelos finais escolhidos correspondem aos modelos 9 e 10 do gráfico 8. Os quadros a seguir constém informações resumo sobre os modelos escolhidos, com a descrição dos parâmetros utilizados.

R42Evento

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value	
s(Id)	7.865	8.611	6.994	1.02e-09	***
s(LagChuva1)	7.286	8.227	18.242	< 2e-16	***
s(LagRadiacao1)	7.254	8.263	16.453	< 2e-16	***
s(LagRadiacao2)	7.485	8.436	6.072	3.04e-07	***
s(LagRadiacao3)	5.429	6.614	5.049	1.90e-05	***
s(LagRadiacao4)	6.660	7.805	3.268	0.00101	**
s(LagRadiacao5)	7.257	8.281	3.006	0.00346	**
s(LagRadiacao30)	6.980	8.070	4.181	4.57e-05	***
s(LagUmidade1)	7.905	8.411	53.630	< 2e-16	***
s(LagPressao5)	4.398	5.422	33.771	< 2e-16	***
s(LagTemp30)	3.206	4.102	1.957	0.09910	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = -0.309 **Deviance explained = 45.1%**

-REML = 10029 Scale est. = 4.758 n = 17896

Figura 9: Resumo do modelo R42Evento

R42CC

Approximate significance of smooth terms:

	edf	Ref.df	F	p-value	
s(Id)	2.325	2.925	1.668	0.191983	
s(LagChuva1)	7.782	8.567	32.265	< 2e-16	***
s(LagChuva2)	1.002	1.004	5.863	0.015410	*
s(LagChuva3)	1.650	2.049	0.875	0.403724	
s(LagChuva4)	4.121	5.012	3.026	0.009612	**
s(LagChuva5)	2.072	2.598	1.057	0.488183	
s(LagChuva10)	1.827	2.278	8.053	0.000207	***
s(LagRadiacao1)	7.354	8.331	27.135	< 2e-16	***
s(LagRadiacao2)	7.549	8.471	7.614	1.96e-10	***
s(LagRadiacao3)	5.252	6.433	9.616	5.07e-11	***
s(LagRadiacao4)	6.473	7.613	4.866	8.41e-06	***
s(LagRadiacao5)	7.933	8.585	13.464	< 2e-16	***
s(LagUmidade1)	5.429	6.342	70.170	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = -0.453 **Deviance explained = 41.1%**

-REML = 14047 Scale est. = 5.2822 n = 25767

Figura 10: Resumo do modelo R42CC

Observa-se a partir das tabelas que o modelo R42Evento explica aproximadamente 45,1% da variação da precipitação horária neste conjunto de dados, e com relação ao mesmo quesito, o modelo R42CC explica aproximidamente 41,1% da variação da chuva.

Foram preditos os valores de precipitação conforme os dois modelos, R42CC e R42Evento e posteriormente foram obtidas as probabilidades de excedência dos 4 limiares propostos, para cada hora. O resíduo médio da predição em relação ao dados observado foi de 0.255 para o modelo R42Evento e de -0.262 para o modelo R42CC. Apesar disso, observou-se em relação aos valores extremos uma tendência de subestimação do modelo R42Evento com o máximo predito constando como 1.98 mm e uma tendência de superestimação pelo modelo R42CC, que considera mais variáveis associadas à chuva. O máximo valor de precipitação horária estimada por esse modelo é de 82 mm, o que não é factivel.

5 Conclusão

A utilização de modelos aditivos generalizados assumindo a distribuição Tweedie Compound Poisson para o entendimento do comportamento de chuvas horárias demonstrou-se útil para compreender algumas relações entre as variáveis. Contudo, necessita de mais testes e parametrizações para ter sua aplicabilidade confirmada, mesmo com relação o conjunto de dados utilizado. Devem ser testadas outras variáveis monitoradas e outros exemplos de estação em contextos geográficos distintos com a construção de datasets com mais parâmetos que suportem aspectos físicos do fenômeno. Da mesma forma, devem ser consideradas séries com menos falhas de registro.

Com relação à parametrização, também é pertinente buscar novas formas de avaliar os efeitos dos parâmetros, explorando as possibilidades dadas pelo GAMs da forma como são aplicados nos pacotes utilizados. Dentre as vantagens do tipo de modelo aplicado para a modelagem da chuva, destaca-se a de identificação de importantes relações não lineares entre múltiplos fatores quando considerados conjuntamente, com as respectivas suavizações, o que justificou ao menos parte da variação da precipitação. Como exemplo, observou-se especial importância da covariável radiação para a ocorrência e o volume das chuvas, o que guarda relação com a cobertura de nuvens que se forma no céu durante os episódios de precipitação. A dependência da radiação consiste também em uma das fragilidades dos modelos propostos, já que este dado não é observável à noite, fato que enviesou a construção destes modelos iniciais. Como alternativas para a solução desta questão figuram a utilização de outras variáveis e a via computacional, com a aplicação de distintos modelos a cada período.

Com relação aos cenários probabilísticos criados a partir da família tweedie, os mesmos se provaram factíveis do ponto de vista de sua construção e da implementação, pensando em sistemas de alerta e outros sistemas informativos úteis para a sociedade. Elas são consistentes ao representar a menor probabilidade dos eventos mais severos, em relação aos menos severos, tal como se distribui a variável chuva. Entretanto, dependem da qualidade da predição do GAM para ganharem consistência, e serem avaliados adequadamente. Dessa maneira, este estudo deve ter continuidade para assegurar a aplicabilidade dos métodos utilizados.

Agradecimentos

A.C.S. agradece à coordenação da Especialização em Data Science e Big Data da Universidade Federal do Paraná pela oportunidade. Ao Dr. Gavin Simpson e ao LEG-UFPR pelo trabalho de divulgação científica na forma de materiais de grande utilidade aos estudos, e que são disponibilizados abertamente. À Gabriela Branco de Souza. Ao INMET - Instituto Nacional de Meteorologia pela solicitude e compartilhamento de dados. Ao LHG-UFPR pela motivação e aos colegas da turma de 2019-2020 pelos aprendizados no tempo em que convivemos.

Referências

[1] CPRM-SERVIÇO GEOLÓGICO DO BRASIL. *Atlas Pluviométrico do Brasil*.

[2] Paulo Henrique Caramori, Pablo Ricardo Nitsche, Flavio Deppe, Eduardo Alvim Leite, and Rodrigo Yoti Tsukahara. Agrometeorologia operacional no estado do paraná. (1):65–70, 2016.

[3] ANA ANA-AGÊNCIA NACIONAL DE ÁGUAS. *Inventário Estações Pluviométricas*. 2009.

[4] Peter K. Dunn and Gordon K. Smyth. Series evaluation of tweedie exponential dispersion model densities. *Statistics and Computing*, 15(4):267–280, 2005.

[5] Md Masud Hasan and Peter K. Dunn. Understanding the effect of climatology on monthly rainfall amounts in australia using tweedie glms. *International Journal of Climatology*, 32(7):1006–1017, 2012.

[6] Rossita M. Yunus, Masud M. Hasan, Nuradhiathy A. Razak, Yong Z. Zubairi, and Peter K. Dunn. Modelling daily rainfall with climatological predictors: Poisson-gamma generalized linear modelling approach. *International Journal of Climatology*, 37(3):1391–1399, 2017.

[7] Wagner Hugo Bonat and Célestin C. Kokonendji. Flexible tweedie regression models for continuous data. *Journal of Statistical Computation and Simulation*, 87(11):2138–2152, 2017.

[8] David R Maidment et al. *Handbook of hydrology*, volume 9780070. McGraw-Hill New York, 1993.

[9] VT Chow, DR Maidment, and LW Mays. *Applied Hydrology*. 1988.

[10] Demetris Koutsoyiannis, Christian Onof, and Howard S. Wheater. Multivariate rainfall disaggregation at a fine timescale. *Water Resources Research*, 39(7), 2003.

[11] Bent Jorgensen. Exponential dispersion models. 49(2):127–162, 1987.

[12] M Tweedie. An index which distinguishes between some important exponential families. pages 579–604.

[13] Keith Beven. On hypothesis testing in hydrology. *Hydrological Processes*, 15(9):1655–1657, 2001.

[14] Yuqiong Liu and Hoshin V. Gupta. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, 43(7):1–18, 2007.

[15] Simon N. Wood. Inference and computation with generalized additive models and their extensions. 29(2):307–339, 2020.

[16] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):314–318, 1986.

[17] Simon N. Wood, Natalya Pya, and Benjamin Säfken. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563, 2016.

[18] Eric J. Pedersen, David L. Miller, Gavin L. Simpson, and Noam Ross. Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ*, (5), 2019.

[19] Gavin L. Simpson. Modelling palaeoecological time series using generalised additive models. *Frontiers in Ecology and Evolution*, 6(OCT):1–21, 2018.

[20] Simon Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

[21] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[22] Peter K. Dunn and Gordon K. Smyth. Evaluation of tweedie exponential dispersion model densities by fourier inversion. *Statistics and Computing*, 18(1):73–86, 2008.