

## Previsão de riscos de alagamentos e inundações com uso de inteligência artificial

Sabrina Sambati<sup>1</sup>, Renata Genova Martins<sup>2</sup>, Ricardo Batista Vilela<sup>3</sup>, Moshe Cotacallapa<sup>4</sup>, Alex Sandro Aguiar Pessoa<sup>5</sup>, Jussara Dias<sup>6</sup>, Danielle de Almeida Bressiani<sup>7</sup> Gilca Palma Fernandes<sup>8</sup>.

### Resumo

A prevenção e mitigação dos danos provocados por eventos hidrometeorológicos severos em áreas urbanas, como a Região Metropolitana de São Paulo (RMSP), são estratégias necessárias para que as cidades se tornem mais inteligentes e resilientes. No entanto, a previsão de riscos de alagamentos e inundações em áreas populosas é desafiador. Usando o k vizinhos mais próximos (*k Nearest Neighbor – kNN*), um clássico algoritmo de aprendizado de máquina, foi possível mostrar a viabilidade do uso de técnicas de inteligência artificial para a previsão de riscos de alagamentos e inundações. A RMSP foi escolhida como área de estudo devido à sua importância social e econômica e também por representar a região de maior aglomeração urbana do Brasil e por sofrer com os impactos dos eventos hidrometeorológicos severos.

**Palavras-chave:** cidades inteligentes, inteligência artificial, aprendizado de máquina, eventos hidrometeorológicos, alagamentos e inundações, previsão.

### Abstract

The prevention and mitigation of damages related to severe hydro-meteorological events in urban areas, like the São Paulo Metropolitan Area (RMSP), is a necessary strategy to make smarter and more resilient cities. However, predicting floods in populated area is a challenging task. By using the k Nearest Neighbor (*kNN*), a classic machine learning algorithm, it is possible to demonstrate the feasibility of utilizing machine learning techniques to predict flooding. The RMSP was chosen as the study area due to its social and economic importance as well as because it represents the largest urban area in Brazil influenced by severe meteorological events and their impacts.

**Keywords:** smart city, artificial intelligence, machine learning, hydro-meteorological events, flooding, prediction.

---

<sup>1</sup>Climatempo - Agência Brasileira de Meteorologia Ltda., São José dos Campos, SP, E-mail: sabri-nabms@gmail.com

<sup>2</sup>E-mail: renata.martins@climatempo.com.br

<sup>3</sup>E-mail: ricardo.vilela@climatempo.com.br

<sup>4</sup>E-mail: frank.moshe@climatempo.com.br

<sup>5</sup>E-mail: alex.pessoa@climatempo.com.br

<sup>6</sup>E-mail: jussara.dias@climatempo.com.br

<sup>7</sup>E-mail: daniebressiani@gmail.com

<sup>8</sup>E-mail: gilca@climatempo.com.br

## 1 Introdução

Desastres ambientais como alagamentos, inundações e movimentos de massa têm ocupado posição de destaque nas discussões públicas. Estes riscos, decorrentes de eventos meteorológicos severos de curto prazo como tempestades convectivas, descargas elétricas, rajadas e chuvas intensas, são alarmantes principalmente em áreas habitadas (BROLLO; TOMINAGA, 2012; DIAS et al., 2013). A Região Metropolitana de São Paulo (RMSP), com praticamente 21 milhões de habitantes, constitui uma das maiores aglomerações urbanas do mundo, sua população constantemente sofre com transtornos resultantes da ocorrência destes tipos de eventos. Somente em 2008 as inundações na cidade de São Paulo refletiram em perdas de até R\$ 248 milhões à cidade e R\$ 560 milhões ao Brasil (HADDAD; TEIXEIRA, 2015). Caso o ritmo de expansão continue inalterado, pesquisas indicam que a mancha urbana será aproximadamente 38% maior que a atual no ano de 2030, como consequência, os impactos causados pelos alagamentos e inundações tendem a aumentar, acarretando danos sócio-econômicos à região ainda maiores (NOBRE et al., 2010).

O conceito de cidades inteligentes foi elaborado para alcançar os objetivos estabelecidos pelo Protocolo de Kyoto (COCCHIA, 2014) e sua adoção deve ser uma estratégia para a redução da vulnerabilidade urbana atuando nas necessidades da comunidade com o emprego de novas tecnologias de informação e comunicação. O processo de suporte à decisão necessita da ativa participação dos cidadãos por meio da disponibilização de profícuas informações em tempo real sobre o ambiente urbano. A previsão de eventos ambientais adversos possibilita instruir a população com intuito de minimizar o impacto no cotidiano. Portanto, conceber um modelo que forneça previsão de eventos severos de curtíssimo prazo torna-se imprescindível.

Temas que necessitam de soluções robustas em tempo hábil são excelentes aplicações para as técnicas de inteligência artificial (IA), uma

área que sempre despertou grande fascínio. A atual era digital propiciou a existência de um grande volume diversificado de informações provenientes de fontes e serviços distintos, gerados e atualizados em alta velocidade (KITCHIN, 2013). Conhecidos por *Big Data*, estes conjuntos de dados são responsáveis pelo êxito recente da inteligência artificial, cujas técnicas e métodos, que antes eram restritos a área científica, passaram a ter grande importância por permitir a realização de previsões mais precisas e intervenções mais assertivas numa escala aparentemente ilimitada (MCAFEE et al., 2012).

Dentre as técnicas de IA mais proeminentes está o aprendizado de máquina (AM) – uma ferramenta computacional capaz de melhorar o desempenho na realização de tarefas por meio da experiência, permitindo aos computadores adquirirem conhecimento através de exemplos e analogias (NEGNEVITSKY, 2005; MITCHELL et al., 1997).

No âmbito do projeto de pesquisa “*Gestão colaborativa de riscos a desastres: aplicativo e sistema de gestão de riscos por cidades, cidadãos e empresas mais resilientes*”, desenvolvido pela Climatempo, com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e da Financiadora de Estudos e Projetos (Finep), foi desenvolvido um sistema para auxiliar cidadãos e empresas a lidar com riscos ambientais na RMSP. A interface principal desse sistema com o público é o aplicativo e site Pédagua (VILELA et al., 2018). Informações de precipitação em tempo quase real, previsão de curtíssimo prazo (*nowcasting*) e mapeamento de suscetibilidade à alagamentos e inundações combinados através de técnicas de inteligência artificial compõem o núcleo do Pédagua.

O aprendizado de máquina destaca-se entre as técnicas de IA pois possibilita que os computadores aprimorem o desempenho na realização de tarefas por meio da experiência, exemplo e analogia (NEGNEVITSKY, 2005; MITCHELL et al., 1997). O kNN – *k Nearest Neighbor*, proposto por Cover e Hart (1967), é um método de aprendizado de máquina não

paramétrico e um dos algoritmos mais antigos e simples usados para classificação de padrões. Diante deste contexto foi feita uma análise da aplicação do *kNN* na identificação de riscos de alagamentos e inundações na Região Metropolitana de São Paulo em tempo real através do uso de diferentes combinações de dados ambientais.

Cabe lembrar que este artigo é uma extensão do artigo apresentado no II Workshop @NUVEM. Nesta versão, optou-se por descrever de forma mais detalhada a metodologia utilizada, bem como os resultados encontrados.

## 2 Previsão com inteligência artificial

Os sistemas de previsão de modelos físicos foram beneficiados com os avanços dos algoritmos de aprendizado de máquina que ofereceram soluções mais simples, robustas e com menor custo computacional. A capacidade de lidar com a não-linearidade de determinados modelos baseando-se somente em dados históricos é uma das características responsável pela popularização destes métodos (ARAGHINEJAD; AZMI; KHOLGHI, 2011).

Existem inúmeras aplicações que utilizam métodos de aprendizado de máquina, porém muitas delas, como é o caso da previsão de alagamento e inundação, encontram-se na fase inicial de estudo. Mosavi, Ozturk e Chau (2018) descrevem o estado da arte dos modelos de aprendizado de máquina na previsão de inundações e apresentam uma relação dos métodos mais promissores para previsão de longo e curto prazo. Algumas estratégias como metodologias híbridas, decomposição de dados, conjunto de algoritmos e otimização de modelos são relatados como as mais eficazes para a melhoria desses métodos.

No trabalho de Choubin et al. (2019) foi utilizado a previsão por conjunto para suscetibilidade a inundações. Essa abordagem produz uma combinação de previsões individuais de cada modelo, fornecendo geralmente melhores resultados. Neste estudo foram empregados os algoritmos SVM (Máquinas de Suporte Vetoriais) em conjunto com Análise Discrimi-

nante Multivariada (MDA) e Classificação e Regressão por Árvores (CART). Ao contrário do SVM e CART, classificados como métodos não-paramétricos, a técnica MDA necessita que a base de dados satisfaça as premissas de normalidade, característica de métodos paramétricos. Os autores enfatizaram que a principal limitação do estudo foi a escassez de informações sobre a localização da ocorrência de eventos hidro-meteorológicos, deixando claro a necessidade de soluções mais interessantes para preencher essa lacuna na base de dados.

O avanço dos métodos não paramétricos na hidrologia foi estimulado pelo reconhecimento da não linearidade da dinâmica dos processos hidrológicos. Utilizando médias locais ponderadas da variável dependente é realizada a estimação das densidades de probabilidade e funções de regressão (ARAGHINEJAD; AZMI; KHOLGHI, 2011). Esta é a base para os métodos dos vizinhos mais próximos. A primeira formulação de uma regra para esta metodologia foi esboçada por Cover e Hart (1967) e ficou conhecida por *k*-vizinhos mais próximos (*kNN*). Um dos métodos de classificação mais básicos, usa a semelhança entre as observações e os dados históricos para obter a melhor estimativa.

Um modelo de previsão eficiente baseado nos algoritmos de aprendizado de máquina é capaz de lidar com informações imprecisas e ruidosas sem efeito negativo perceptível na qualidade da resposta, conseguindo identificar padrões existentes entre as amostras desconhecidas. A definição de quais características (atributos) farão parte da base de dados é uma etapa crucial no processo de previsão. A base de dados, conhecida também por conjunto de treinamento, é composta por dados históricos e é usada como entrada na fase de aprendizado, em que o conhecimento é adquirido.

Com relação ao método de aprendizado, os algoritmos de AM podem ser classificados em supervisionado e não supervisionado. No aprendizado supervisionado existe um mapeamento entre cada padrão de entrada e a saída, com base na analogia o algoritmo aprende a reproduzir as saídas desejadas. Quando a saída é

representada por valores discretos ou categóricos, o problema é conhecido por classificação, se a saída consiste de valores contínuos, então é um problema de regressão (BISHOP, 2006). No caso do aprendizado não-supervisionado a saída não é mais necessária, o algoritmo é deixado livre para convergir na direção de qualquer estado sem ajuda externa

A não linearidade embutida na previsão de ocorrência de alagamento e inundação foi crucial para a definição do método de inteligência artificial adotado neste trabalho. O *kNN*, considerado um método não paramétrico, fornece estrutura flexível. Basicamente, o *kNN* atribui a um ponto não classificado a classe mais fortemente representada entre os *k* vizinhos mais próximos. Apenas dois fatores precisam ser definidos para aplicação desta metodologia: o parâmetro *k*, que indica o número de vizinhos que serão considerados durante o processo de classificação de novas amostras; e a métrica para o cálculo da similaridade entre um ponto e todos os outros pertencentes ao conjunto de treinamento.

A determinação do valor de *k* é feita empiricamente mas sempre priorizando valores ímpares, sendo que o *k* “ótimo” varia de acordo com o problema tratado. A escolha da medida de similaridade que será adotada é feita pela combinação de experiência, habilidade e conhecimento (GAN; MA; WU, 2007). Para cálculo da similaridade as principais métricas de distância usadas são: Minkowski, Manhattan e Euclidiana. A distância de Minkowski, dada pela Equação 1, é uma generalização das distância Euclidiana e Manhattan.

$$\sqrt[p]{\sum_{i=1}^n (|x_i - y_i|^p)} \quad (1)$$

onde *p* é um número real  $\geq 1$  e a escolha de seu valor depende do peso dado as distâncias maiores. Para *p* = 1 tem-se a distância Manhattan, e no caso mais específico, *p* = 2, corresponde a distância Euclidiana.

Com intuito de analisar a influência dos diferentes parâmetros e atributos na previsão de alagamentos e inundações aplicando o *kNN*, alguns cenários foram criados variando o valor

de *k*, a medida de similaridade e o conjunto de atributos (sub-conjuntos da base de referência) definido pelo WEKA – *Waikato Environment for Knowledge Analysis*. Para selecionar os melhores cenários a avaliação foi feita visualmente e qualitativamente através do número de pontos de alagamentos registrados.

## 2.1 Descrição dos dados

Para a análise dos fenômenos hidrometeorológicos mencionados neste trabalho, a base de referência, usada como conjunto de treinamento e alicerce para o sistema de previsão, foi formada por dados extraídos de diferentes fontes, sendo:

- (i) Precipitação estimada com informações de dois radares meteorológicos banda X do projeto Chuva Online da Universidade de São Paulo (USP) em parceria com a Climatempo, com resolução espacial de 100 *m* e temporal de 5 minutos;
- (ii) Precipitação estimada com dados do satélite GOES-16, resolução espacial de 2*km* e temporal de 15 minutos;
- (iii) Dados de descargas atmosféricas medidos pelas redes *STARNET – Sferics Timing And Ranging NETWORK* (RODRIGUEZ; NEVES; ANSELMO, 2011) e *ENTLN – Earth Networks Total Lightning Network* (LIU; HECKMAN, 2012);
- (iv) Mapa de suscetibilidade física aos alagamentos e inundações na RMSP;
- (v) Dados colaborativos usados como rótulos de ocorrência de eventos hidrometeorológicos, empregados na classificação dos elementos do conjunto de treinamento.

O radar (*Radio Detection And Ranging – Detecção e Telemetria por Rádio*), extremamente útil na representação das condições meteorológicas em tempo real, é um sensor que emite um feixe de ondas eletromagnéticas na faixa do micro-ondas, parte dessa energia é

retro-espalhada devido a interação com hidrometeoros na atmosfera e apenas parte retorna para o radar (BATTAN, 1973). A energia medida pelo radar é chamada de fator de refletividade ( $Z$ ). Para obter a taxa de precipitação quantitativa ( $R$ ) em  $mm/h$  é aplicada uma relação  $Z/R$ , possibilitando avaliar a chuva em tempo real e fornecer apoio para avisos e gestão de enchentes e inundações. Com os dados dos radares instalados nos campi da USP Leste e Butantã foi possível calcular as seguintes variáveis: precipitação acumulada em 1 hora por pixel ( $mm$ ), taxa de precipitação por pixel ( $mm/hora$ ) e precipitação média na bacia hidrográfica acumulada em 1, 2, 3 e 24 horas.

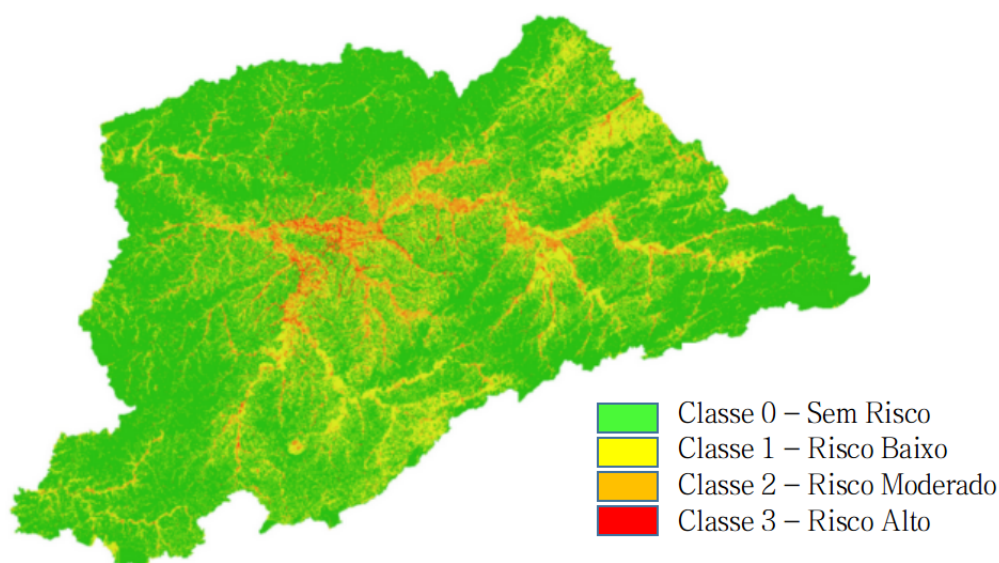
No caso do satélite, a estimativa de precipitação foi calculada utilizando o método Hidroestimador (VICENTE; SCOFIELD; MENZEL, 1998), em que as taxas de precipitação são obtidas por meio de uma relação empírica, expressa por uma curva exponencial, resultante da regressão entre as medidas instantâneas de precipitação derivadas de radares meteorológicos e a temperatura de brilho do canal  $10.8 \mu m$  do satélite GOES-16. Este método considera fatores adicionais como crescimento e estrutura do topo de nuvens e também o regime de umidade proveniente do WRF (*Weather Research and Forecasting Model*), um modelo numérico de previsão de tempo. Tendo como base a relação direta entre as maiores taxas de precipitação e as temperaturas de topos de nuvens mais frias, o algoritmo do Hidroestimador também considera a disponibilidade de umidade e a morfologia da célula convectiva, com a finalidade de evitar nuvens de topo frio que não causem precipitação, como é o caso da cirrus – nuvem predominantemente composta por gelo, apresentando forma distinta de núcleos convectivos precipitantes. Neste trabalho foram utilizados as seguintes variáveis fornecidas pelo Hidroestimador: precipitação acumulada em 1 hora por pixel ( $mm$ ), precipitação acumulada na bacia hidrográfica a cada 1, 2, 3 e 24 horas e taxa de precipitação por pixel ( $mm/hora$ ).

Os dados de descargas atmosféricas foram

obtidos de duas diferentes fontes: rede *STAR-NET* e rede *ENTLN*. As informações dessas redes são provenientes de antenas de rádio distribuídas na superfície que recebem o sinal eletromagnético emitido pela descarga, sendo que a diferença do tempo de chegada do sinal entre as antenas que compõem a rede é o que possibilita definir, com uma acurácia de 2 a 5  $km$ , a localização do raio. A rede *ENTLN* consegue detectar tanto os raios nuvem-solo quanto os intra-nuvens, já que utiliza antenas de diferentes frequências de recepção, operando num intervalo entre  $1HZ$  e  $12MHZ$ . A rede *STAR-NET* dispõem de doze antenas receptoras na faixa de frequência de VLF variando entre 7 a 15  $kHz$ , oferecendo somente medidas de raios nuvem-solo (LEE, 1986).

A estratégia adotada para gerenciar os riscos de inundação e alagamento nas regiões urbanas foi fundamentada na definição das áreas fisicamente suscetíveis a ocorrência destes eventos. Com a integração de tecnologias de sensoriamento remoto, do Sistemas de Informação Geográfica (GIS) e levantamentos de campo (RIMBA et al., 2017) foi possível mapear as áreas de risco. Foi empregada uma metodologia adaptada de Bressiani (2016) que tem como base o Processo de Hierarquia Analítica (*Analytical Hierarchy Process – AHP*) desenvolvido por Saaty (1994). Para elaborar o mapa de suscetibilidade utilizou-se cinco parâmetros espaciais: (i) taxa de infiltração (combinação do mapa de uso e ocupação do solo e pedologia); (ii) declividade; (iii) área de contribuição relativa à drenagem; (iv) distância horizontal e vertical relativa ao curso d'água mais próximo e (v) pontos de alagamento cadastrados pelo CGE – Centro de Gerenciamento de Emergências Climáticas. Utilizando o programa QGIS (*Quantum GIS*) para realizar a composição o mapa foi classificado em 4 categorias: 0 - Sem risco, 1- Risco baixo; 2- Risco moderado e 3: Risco alto, representado pela Figura 1.

Por último, o conjunto de treinamento utilizou dados colaborativos, uma rica fonte de informações, como rótulo de ocorrências de alagamento. Habitualmente cidadãos compar-



**Figura 1:** Mapa de susceptibilidade física da região metropolitana.

tilham suas experiências pelos meios de comunicação, estas informações conhecidas por dados colaborativos costumam ser extremamente detalhadas e com alta qualidade, superando diversas vezes as disponibilizadas pelo órgãos oficiais. A aquisição destes dados percorreu três fases: i) rastreamento e armazenamento de dados ambientais reunindo as frases obtidas dos relatos de variadas fontes de mídias sociais, como por exemplo: *Twitter*, *Flickr*, *Instagram*; ii) filtragem utilizando processamento de linguagem natural e agrupamento das informações georreferenciadas com a aplicação de um estimador de densidade por *kernel* – KDE, contudo para que as informações seja agrupadas é necessário a ocorrência de pelo menos três relatos num intervalo de tempo de 30 minutos e com uma distância máxima de 200 m entre eles iii) verificação de possíveis falsas ocorrências relatadas através da comparação com dados reais de precipitação da bacia hidrográfica correspondente ao relato (HORITA et al., 2018a; HORITA et al., 2018b).

Como observado, a base de dados utilizada como conjunto de treinamento é heterogênea, tornando necessário a adoção de técnicas de pré-processamento para tratar situações como

presença de anomalias ou discrepâncias, problemas de redundância, inconsistência e existência de ruído. As etapas descritas na Tabela 1 foram realizadas com o propósito de tornar a base mais robusta e íntegra.

Questões usualmente presentes em base de dados devem ser abordadas com cautela, a existência de dados ausentes é um exemplo. O preenchimento destas lacunas com um determinado valor é uma forma de lidar com esta situação, sendo que a convenção mais global é o uso NaN (acrônimo para *Not a Number*), um valor especial que faz parte da especificação de ponto flutuante do IEEE (VANDERPLAS, 2016).

Outro ponto importante é a presença de atributos com diferentes escalas, situação normalmente encontrada em bases heterogêneas e responsável por problemas como desbalanceamento em algoritmos de classificação. Uma solução para esta questão é o uso da normalização, cujo objetivo é impedir que atributos com intervalos inicialmente grandes superem os atributos com intervalos inicialmente pequenos. Existem diversos métodos de normalização incluindo normalização por escala decimal, *z-score* e normalização mínimo-máximo.

**Tabela 1:** Visão geral das etapas de pré-processamento

Limpeza dos dados:	Remoção de ruído e de dados inconsistentes
Integração dos dados:	Junção de diversas fontes de dados
Seleção dos dados:	Seleção dados mais representativos
Transformação dos dados:	Conversão de dados para um determinado tipo e aplicação de normalização
Mineração de dados:	Extração de padrões com métodos inteligentes
Avaliação de padrões:	Identificação dos padrões mais representativos

A normalização por escala decimal transforma o valor dos atributo movendo o ponto decimal, a *z-score* utiliza informações de média e desvio padrão, e a normalização mínimo-máximo realiza uma transformação linear nos dados originais usando o menor e o maior valor dos atributos (SHALABI; SHAABAN; KASASBEH, 2006). Neste trabalho foi adotada a normalização mínimo-máximo por preservar o relacionamento entre os valores dos atributos originais (HAN; MICHELINE; PEI, 2011).

Para representar de forma consistente o conhecimento desejado, ou seja, os eventos hidrometeorológicos e suas consequências, foi extraída da base de dados uma série temporal composta por 18 atributos. A Tabela 2 ilustra os atributos distribuídos em quatro grupos: precipitação por radar, precipitação por satélite, rede de raios e dados colaborativos, presentes na primeira coluna. Na segunda coluna aparece o nome e a resolução temporal e na terceira coluna está a unidade dos atributos que serão usados pelo *kNN*.

### 3 Análise dos dados e resultados

Após o pré-processamento procedeu-se a análise exploratória da base de dados com intuito de avaliar a qualidade, extraíndo informações relevantes e identificando possíveis relações entre as variáveis meteorológicas selecionadas. Esta análise é um poderoso enredo a partir do qual originam-se muitas inspirações. A Figura 2(a) apresenta a distribuição dos dados de precipitação acumulada em 1, 2, 3 e 24 horas provenientes do Hidroestimador e Figura 2(b) do radar. Pode-se verificar que em ambas a variância da precipitação parece ser

positivamente correlacionada com o tempo. Já a Figura 3 representa as distribuições da precipitação acumulada do Hidroestimador em 4 intervalos, percebe-se que grande parte dos dados ficam concentrados nos valores nulos, um comportamento esperado para o tipo de informação representada – precipitação.

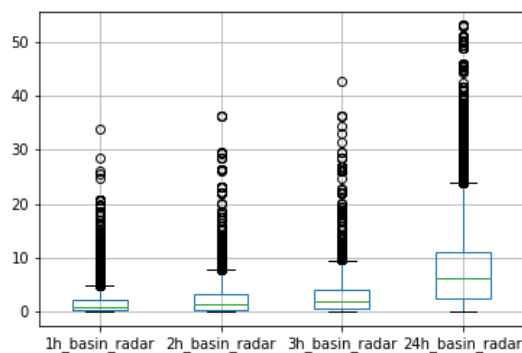
Outra etapa importante do trabalho foi identificar quais fatores estavam associados as ocorrências dos eventos hidrometeorológicos. Diversas combinações de dados ambientais foram feitas em busca do conjunto ou cenário que representasse a melhor estimativa. Assim como relatado por Han, Chan e Zhu (2007), a seleção ideal entre um grande número de várias combinações e parâmetros de entrada é um desafio real para qualquer modelo utilizando algoritmo de aprendizado de máquina. Com auxílio do WEKA – *Waikato Environment for Knowledge Analysis*, um reconhecido sistema de mineração de dados (WITTEN et al., 2016) alguns conjuntos foram definidos. E por meio de avaliação visual e qualitativa (número de pontos de alagamentos previstos) dos resultados obtidos, os cenários que obtiveram os melhores desempenhos na previsão de risco de alagamento e inundações na RMSF foram: cenário A, B e C, listados na Tabela 3.

O recorte temporal adotado para compor o conjunto de treinamento foi de 11 de outubro de 2017 a 19 de maio de 2018, com resolução espacial de  $0.001^\circ$  e frequência de 15 minutos. Este conjunto possui 91290 amostras, entre elas 268 representam pontos de alagamento identificados pelos dados colaborativos e órgãos governamentais.

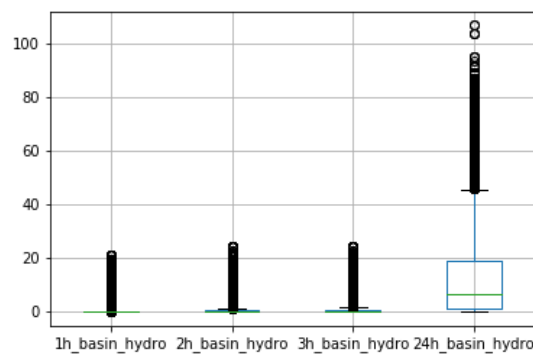
Para avaliar a qualidade e eficiência do método empregado como classificador foi esco-

**Tabela 2:** Conjunto de treinamento usado pelo *kNN*

	Atributos	Tipo
Localização:	Latitude Longitude	Real
	Mapa de suscetibilidade física	Inteiro
Precipitação por Radar:	Precipitação na bacia: 1h, 2h, 3h, 24h Precipitação no pixel 1 h Taxa precipitação	Real
Precipitação por satélite: (Hidroestimador)	Precipitação na bacia: 1h, 2h, 3h, 24h Precipitação no pixel 1h Taxa precipitação	
Rede de raios:	STARNET ENTLN	Binário
Dados colaborativos:	Alagamento e inundação	



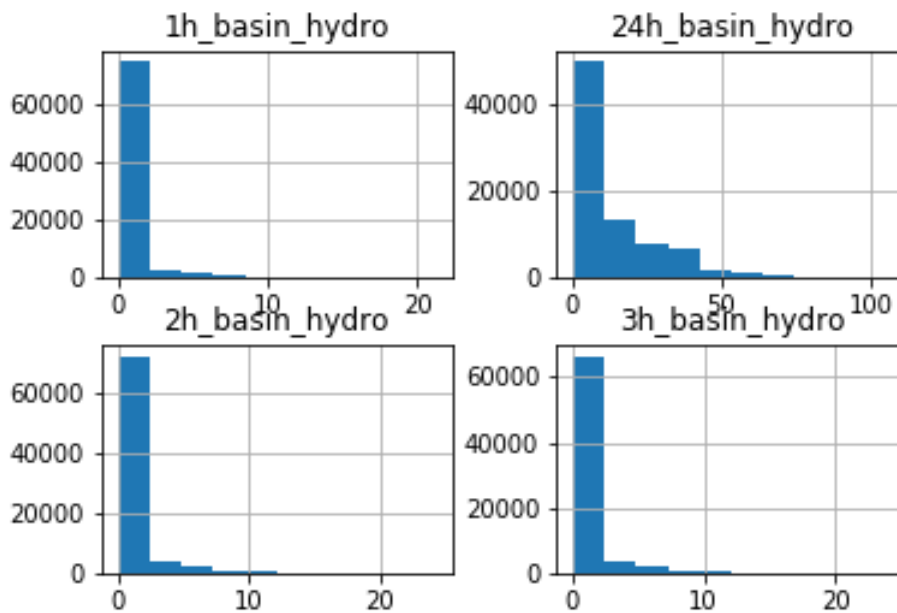
(a) Radar



(b) Hidroestimador

**Figura 2:** Precipitação na bacia: 1h, 2h, 3h e 24h





**Figura 3:** Hidroestimador - precipitação acumulada na bacia hidrográfica a cada 1, 2, 3 e 24 horas.

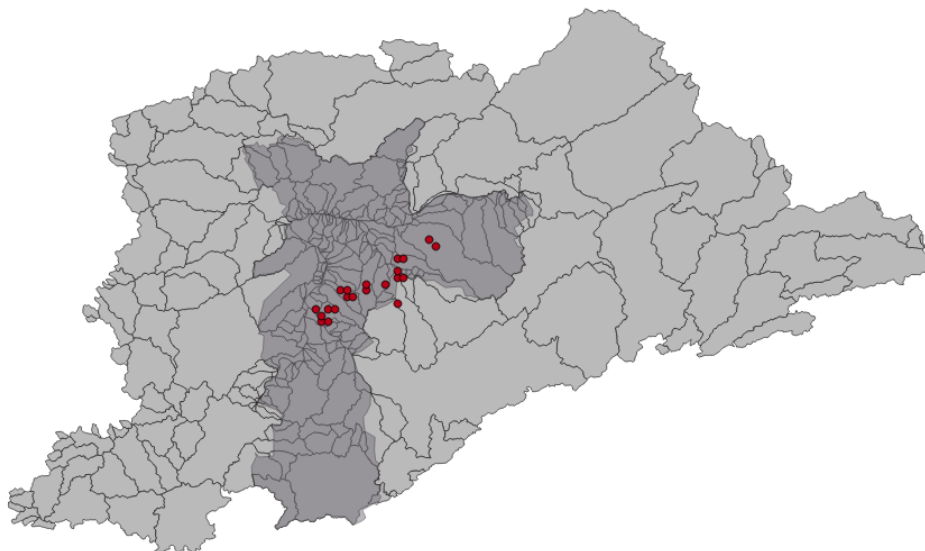
**Tabela 3:** Relação dos cenários selecionados pelo WEKA.

Cenário A	Rede de raios	<i>ENTLN</i>
	Precipitação por satélite	Taxa precipitação
		Precipitação na bacia 3h
		Precipitação na bacia 24h
	Suscetibilidade	
Cenário B	Precipitação por radar	Precipitação 1h
		Precipitação na bacia 1h/2h/3h/24h
	Precipitação por satélite	Precipitação 1h
		Precipitação na bacia 1h/2h/3h/24h
Cenário C	Rede de raios	<i>ENTLN</i>
	Precipitação por Radar	Precipitação 1h
	Precipitação por satélite	Precipitação na bacia 1h

lhido a data de 17 de outubro de 2018, pois (Figura 4) de acordo com os dados do CGE (Centro de Gerenciamento de Emergências Climáticas da Prefeitura de São Paulo) houve diversas ocorrências de alagamentos na região metropolitana de São Paulo neste dia. Este conjunto, usado para validação, é composto por 1775 amostras, sendo que 21 correspondem a pontos de alagamento. Este cenário pode ser

visto na Figura 4 em que as bacias hidrográficas da RMSP são representadas pela cor cinza clara, o município de São Paulo está representado pelo cinza mais escuro e os pontos vermelhos correspondem aos pontos de alagamentos observados informados pelo CGE.

Foi utilizado o parâmetro  $k = 1$  e distância euclidiana como métrica para o cálculo da similaridade entre os pontos. Os resultados obti-



**Figura 4:** Pontos de alagamentos relatados pelo CGE na região metropolitana de São Paulo.

dos com os cenários descritos na Tabela 3 podem ser vistos na Figura 5, em que as ocorrências de alagamento relatadas pelo CGE são identificadas pela cor vermelha, o resultado obtido com o cenário A é representado pela cor verde, a cor amarela corresponde ao cenário B e a cor azul indica os resultados obtidos com o cenário C. É possível verificar que os eventos identificados pelo *kNN* nos três diferentes cenários seguem um padrão similar aos observados. Os cenários A e C obtiveram os melhores resultados. porém analisando as Figuras 6-a e 6-b percebe-se que um ponto de alagamento foi identificado com precisão pelo cenário A.

#### 4 Conclusão

Um dos principais desafios da atualidade é transformar os dados processados em conhecimento de maneira inteligente e automática. Esta necessidade causou o avanço das técnicas de inteligência artificial, permitindo que diversas áreas da ciência explorem melhor a grande massa de dados disponível. Aplicações meteorológicas e ambientais que exigem computação intensiva e precisão tem se beneficiado com esse desenvolvimento.

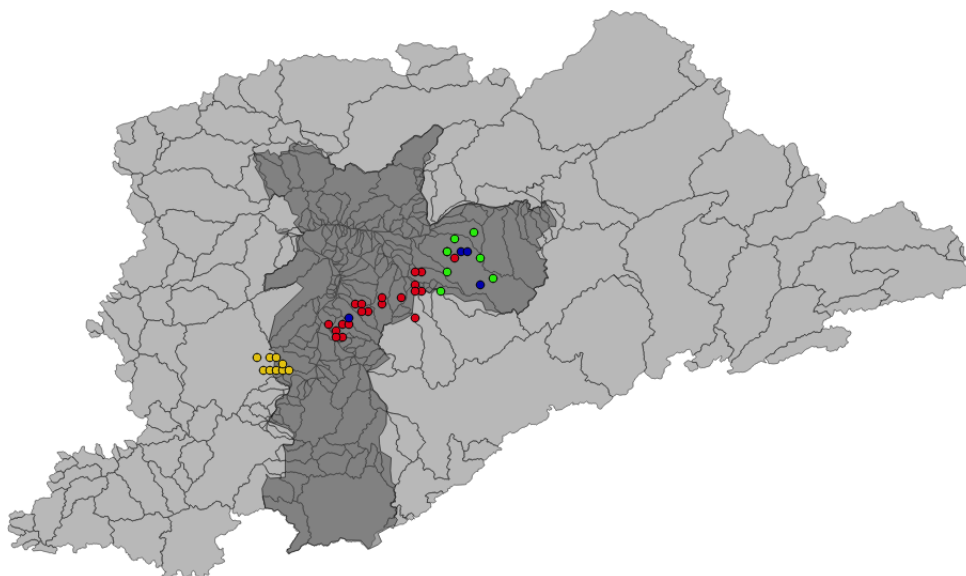
Neste trabalho, mostrou-se que a abordagem proposta usando o *kNN*, um método de aprendizado de máquina de simples implementação apresentou ótimas perspectivas na identifica-

ção de riscos de eventos hidrometeorológicos severos, como alagamento e inundação. Contudo, a fase de seleção dos atributos que definem os cenários exige extrema dedicação por estar diretamente ligado a qualidade dos resultados. Um ponto interesse a ser observado é que variáveis importantes para identificação de alagamento, como por exemplo a suscetibilidade física do terreno, nem sempre foram selecionadas no conjunto de atributos definidos pelo WEKA.

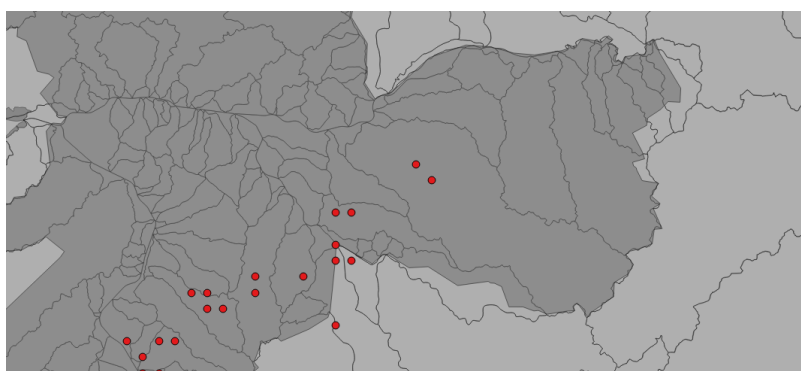
O *kNN* é um método computacionalmente dispendioso quando aplicado a grande base de dados, já que a cada nova amostra o algoritmo calcula a distância entre ela e todos os elementos pertencentes ao conjunto de treinamento, a fase de treinamento é implícita. Porém esta característica não foi considerada neste trabalho, pois o objetivo preliminar era verificar a viabilidade de realizar previsões dos eventos hidrometeorológicos severos utilizando técnicas de inteligência artificial.

#### Agradecimentos

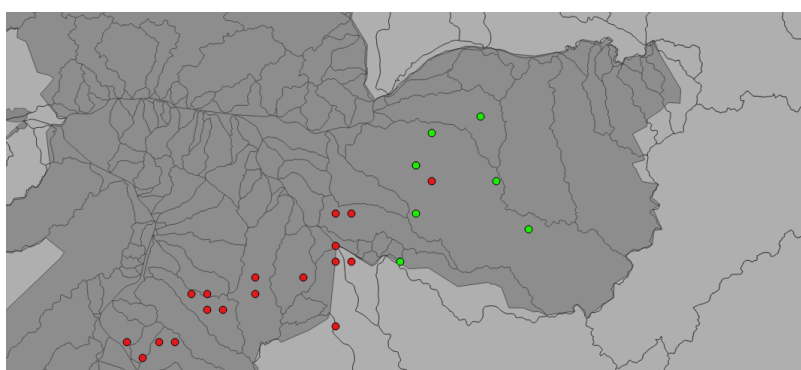
Os autores agradecem à Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) (números de processos: 2016/10229-3, 2018/05737-5, 2017/17007-9, 2018/18053-7, 2017/18783-2 e 2017/10394-7) e à Financiadora de Estudos e Projetos (Finep) (processo



**Figura 5:** Resultado do *kNN*: cenário A (verde), cenário B (amarelo), cenário C (azul).



(a) Observado (CGE)



(b) Cenário A

**Figura 6:** Pontos de alagamento

número 2017/50057-0) pelo apoio financeiro.

## Referências

- ARAGHINEJAD, S.; AZMI, M.; KHOLGHI, M. Application of artificial neural network ensembles in probabilistic hydrological forecasting. *Journal of Hydrology*, Elsevier, v. 407, n. 1-4, p. 94–104, 2011.
- BATTAN, L. J. Radar observation of the atmosphere. 1973.
- BISHOP, C. M. *Pattern recognition and machine learning*. [S.l.]: Springer, 2006.
- BRESSIANI, D. d. A. *Coping with hydrological risks through flooding risk index, complex watershed modeling, different calibration techniques, and ensemble streamflow forecasting*. Tese (Doutorado) — Universidade de São Paulo, 2016.
- BROLLO, M. J.; TOMINAGA, L. K. Desastres naturais e riscos geológicos no estado de são paulo: cenário de referência–2012. *Boletim*, 2012.
- CHOUBIN, B. et al. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Science of the Total Environment*, Elsevier, v. 651, p. 2087–2096, 2019.
- COCCHIA, A. Smart and digital city: A systematic literature review. In: *Smart city*. [S.l.]: Springer, 2014. p. 13–43.
- COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967.
- DIAS, M. A. F. S. et al. Changes in extreme daily rainfall for são paulo, brazil. *Climatic Change*, Springer, v. 116, n. 3-4, p. 705–722, 2013.
- GAN, G.; MA, C.; WU, J. *Data clustering: theory, algorithms, and applications*. [S.l.]: Siam, 2007. v. 20.
- HADDAD, E. A.; TEIXEIRA, E. Economic impacts of natural disasters in megacities: The case of floods in são paulo, brazil. *Habitat International*, Elsevier, v. 45, p. 106–113, 2015.
- HAN, D.; CHAN, L.; ZHU, N. Flood forecasting using support vector machines. *Journal of hydroinformatics*, IWA Publishing, v. 9, n. 4, p. 267–276, 2007.
- HAN, J.; MICHELINE, K.; PEI, J. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011.
- HORITA, F. et al. Determining flooded areas using crowd sensing data and weather radar precipitation: a case study in brazil. In: 15TH INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS FOR CRISIS RESPONSE AND MANAGEMENT (ISCRAM). Rochester, USA, 2018.
- HORITA, F. et al. Uso de dados colaborativos na determinação de alagamentos. In: XVII ENCONTRO DOS ALUNOS DE POSGRADUACAO EM METEOROLOGIA DO CPTEC/INPE. Cachoeira Paulista, SP, 2018.
- KITCHIN, R. Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, v. 3, n. 3, p. 262–267, 2013.
- LEE, A. An experimental study of the remote location of lightning flashes using a vlf arrival time difference technique. *Quarterly Journal of the Royal Meteorological Society*, Wiley Online Library, v. 112, n. 471, p. 203–229, 1986.
- LIU, C.; HECKMAN, S. Total lightning data and real-time severe storm prediction. In: *Proc. Conf. on Meteorological Environmental Instrument and Methods of Observation Management Froup (10th session)*. [S.l.: s.n.], 2012.
- MCAFEE, A. et al. Big data: the management revolution. *Harvard business review*, v. 90, n. 10, p. 60–68, 2012.

- MITCHELL, T. M. et al. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, v. 45, n. 37, p. 870–877, 1997.
- MOSAVI, A.; OZTURK, P.; CHAU, K.-w. Flood prediction using machine learning models: Literature review. *Water*, Multidisciplinary Digital Publishing Institute, v. 10, n. 11, p. 1536, 2018.
- NEGNEVITSKY, M. *Artificial intelligence: a guide to intelligent systems*. [S.l.]: Pearson Education, 2005.
- NOBRE, C. A. et al. Vulnerabilidades das megacidades brasileiras às mudanças climáticas: Região metropolitana de são paulo. *Embaxada Reino Unido, Rede Clima e Programa FAPESP em Mudanças Climáticas*, 2010.
- RIMBA, A. B. et al. Physical flood vulnerability mapping applying geospatial techniques in okazaki city, aichi prefecture, japan. *Urban Science*, Multidisciplinary Digital Publishing Institute, v. 1, n. 1, p. 7, 2017.
- RODRIGUEZ, C. A. M.; NEVES, J.; ANSELMO, E. Sferics timing and ranging network-starnet: evaluation over south america. *ICAE 2011: Abstracts and Program*, 2011.
- SAATY, T. L. How to make a decision: the analytic hierarchy process. *Interfaces*, *Interfaces*, Inform, v. 24, n. 6, p. 19–43, 1994.
- SHALABI, L. A.; SHAABAN, Z.; KASASBEH, B. Data mining: A preprocessing engine. *Journal of Computer Science*, v. 2, n. 9, p. 735–739, 2006.
- VANDERPLAS, J. *Python data science handbook: essential tools for working with data*. [S.l.]: O’Reilly Media, 2016.
- VICENTE, G. A.; SCOFIELD, R. A.; MENZEL, W. P. The operational goes infrared rainfall estimation technique. *Bulletin of the American Meteorological Society*, v. 79, n. 9, p. 1883–1898, 1998.
- VILELA, R. et al. Sistema para previsão de riscos de alagamentos e inundações: um estudo de caso para a região metropolitana de são paulo (rmsp). In: XVII ENCONTRO DOS ALUNOS DE PÓS-GRADUAÇÃO EM METEOROLOGIA DO CPTEC/INPE. Cachoeira Paulista, SP, 2018.
- WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016.