



MODELAGEM PREDITIVA DE EXTRAVASAMENTO EM BACIA DE CONTENÇÃO DE EFLUENTES

Brenner Biasi Souza Silva¹

Karla Patrícia Oliveira Esquerre²

Robson Wilson Silva Pessoa³

Tarssio Brito Barreto⁴

Resumo

Apesar da complexidade de criar um modelo para o fenômeno de extravasamento em sistemas de contenção de efluentes, a modelagem preditiva é de extrema importância devido a sua capacidade funcional com ferramenta de suporte a decisão. Neste artigo foi desenvolvida e aplicada uma metodologia para modelagem preditiva em sistema de drenagem pluvial em área industrial utilizando aprendizado de máquina com o objetivo de indicar se a bacia de contenção extravasará no intervalo de 24 horas. As técnicas utilizadas foram *k-nearest neighbors* (KNN) e *Random Forest* (RF). Cenários e variações de técnicas de amostragem para o treinamento foram utilizados. Os melhores resultados foram obtidos a partir do algoritmo RF com emprego da técnica de reamostragem *oversampling*, *undersampling* e ROSE.

Palavras-chave: aprendizado de máquina, *random forest*, *knn* e extravasamento de efluentes.

Abstract

Although it is complex to create a model about the phenomenon of overflow at wastewater retention systems, predictive modeling is extremely important due to its functional capacity as a decision support tool. In this paper, a methodology was developed and applied for predictive modeling in a rainwater drainage system in an industrial area using machine learning with the objective of indicating whether the retention tank will overflow within 24 hours. The techniques used were *k-nearest neighbors* (KNN) and *Random Forest* (RF). Scenarios and variations of sampling techniques for the training were used. The best results were obtained from the RF algorithm using the *oversampling*, *undersampling* and ROSE re-sampling techniques.

Keywords: machine learning, *random forest*, *knn* and wastewater overflow.

¹ Universidade Federal da Bahia (UFBA), brennerbisiss@gmail.com

² Universidade Federal da Bahia (UFBA), karlaesquerre@ufba.br

³ Universidade Federal da Bahia (UFBA), robsonpessoa2007@gmail.com

⁴ Universidade Federal da Bahia (UFBA), tarssioesa@gmail.com



Introdução

Em sistemas de drenagem pluvial urbana combinados com o sistema de esgotamento sanitário problemas ambientais podem ocorrer em função do lançamento inadequado de efluentes. Este evento é conhecido como transbordamento de esgoto combinado (*Combined sewer overflow* - CSO), e não é um fenômeno restrito a zonas urbanas, também podendo ocorrer em indústrias que contenham um sistema de drenagem (YU et al., 2013).

O volume excedente de efluentes à capacidade útil dos sistemas está relacionado com o fenômeno da transformação da precipitação pluviométrica para vazão, que é conhecido como processo chuva-vazão (PAZ, 2004).

Estudos recentes de análise estatística de ocorrência de CSO foram realizados em Tóquio no Japão (YU et al., 2013, 2018), em Quebec no Canadá (MAILHOT; TALBOT; LAVALLÉE, 2015) e em uma área piloto no Reino Unido (ZHAO; BEACH; REZGUI, 2017).

Utilizando dados de monitoramento de campo e regressão LASSO, Zhao, Beach e Rezgui (2017) elaboraram modelos preditivos de CSO de forma automatizada, proporcionando uma nova opção metodológica para estimativas de modelos chuva-vazão sem emprego direto de modelagem hidrológico-hidráulica.

Estes modelos se configuram como ferramentas bastante pertinentes, principalmente, pela dificuldade que há em se obter dados locais de chuvas com frequência inferior a diária e a não disponibilidade de longas séries de dados para locais isolados.

Este tipo de desenvolvimento de conhecimento baseado em modelagem empírica utilizando algoritmos computacionais e metodologia como mineração de dados e aprendizado de máquina são importantes na descoberta de conhecimento e na tomada de decisões nas diversas possíveis áreas de aplicação, particularmente, quando se há dificuldade em obter parte dos dados, como a pluviometria (GE et al., 2017).

Aprendizado de Máquina Supervisionado

Os métodos de aprendizado de máquina supervisionado tem sido amplamente utilizados para várias aplicações nas indústrias (GE et al., 2017). Um dos métodos supervisionados é o *KNN* (*K-nearest neighbors*), que é um método não-paramétrico baseado na votação dos vizinhos mais próximos de acordo com métrica de distância para similaridade. (SILVA; PERES; BOSCARIOLI, 2016).

Outro algoritmo não-paramétrico é o de Árvore de Decisão (AD), que são simples e úteis para interpretação para modelos de regressão ou classificação, e levam grande vantagem



frente a métodos paramétricos, principalmente pela flexibilização dos pressupostos estatísticos.

O algoritmo AD tende a obter melhores resultados que métodos elementares como o *KNN* e resultados inferiores frente a regressão RIDGE-LASSO, situação que é contornada quando utilizado um modelo misto (*ensemble*), como a *Random Forest* (JAMES et al., 2013).

Modelos de aprendizado de máquina *ensemble* nada mais são que modelos conjugados. *Random Forest* é o método de árvore de decisão conjugado com aplicação de *bagging* para geração de árvores de maneira aleatória, sendo uma ferramenta eficaz para predição e robusta para *outliers* e ruído, bem como implementação em fenômenos com comportamento não-linear (BREIMAN, 2001).

Para a construção de um modelo preditivo de classificação em aprendizado de máquina o ideal é que os dados estejam balanceados, ou seja, a probabilidade de ocorrência de um evento deve ser igual ou próxima da probabilidade da não ocorrência do evento em interesse.

A construção de modelos preditivos que são treinados com conjuntos de dados desbalanceados de classe são altamente suscetíveis a produzir modelos de previsão imprecisos, entretanto é possível utilizar técnicas de reamostragem (*resampling*) para contornar esta questão a partir dos dados observados, como *under-sampling*, *over-sampling* e ROSE (*Random Over Sampling Examples*), que é baseada em *bootstrap* suavizado (TANTITHAMTHAVORN; HASSAN; MATSUMOTO, 2018).

Objetivo

O objetivo deste trabalho foi construir e analisar modelos preditivos para o fenômeno de extravasamento em bacia de contenção de efluentes industriais utilizando as técnicas *KNN* e *Random Forest*.

Material e Método

Estudo de Caso

O presente trabalho é um estudo de caso em um sistema de drenagem de efluentes industriais em uma refinaria de petróleo. O sistema de drenagem tem macro componentes como sistema de transporte, poços de sucção e bacias de contenção. Este pode ser particionado sob a perspectiva de sistema do diferentes tipos de efluentes, um sistema de drenagem para efluentes oleosos e um sistema de drenagem para efluentes contaminados.



Oito bacias de contenção compõem a totalidade do sistema de drenagem, sendo quatro para o sistema de efluente contaminado (CET) e outras quatro para o sistema de efluente oleoso (OET). O efluente classificado como contaminado é aquele que pode ter eventual presença de hidrocarbonetos e contaminantes, e o efluente oleoso tem presença constante destes compostos (HODGSON; BENDIAK, 1987).

Os efluentes contaminados são gerados em toda a área industrial, sendo assim completamente vulneráveis a ação externa, como ocorrência de chuvas. Os efluentes oleosos são gerados a partir de descargas ou falhas de algum processo produtivo dentro da refinaria, podendo estar eventualmente exposto a interferência externa ao processo, como a ocorrência de precipitação pluviométrica. E, todo o efluente coletado deve ser aduzido para Estação de Tratamento de Efluentes Industriais (ETEI).

Neste artigo foi realizada a modelagem preditiva de uma bacia de contenção de efluente contaminado, CET 09. A CET 09 recebe em média uma vazão de entrada (Q_{in}) de 142 m³/h, tem capacidade útil de armazenamento (Vol_{Max}) de 6.800 m³ de efluente, contendo também dois conjuntos motor-bomba (CMB) que proporcionam uma vazão de recalque (Q_{dis}) de 285 m³/h.cmb, e esta bacia é de maior importância para o pleno funcionamento do sistema.

O período dos dados em estudo foi de 01/01/2011 à 14/09/2011. Os dados pluviométricos tem frequência diária, enquanto que os dados de nível percentual da bacia CET 09 e acionamentos (ON/OFF) de CMB foram coletados com frequência 1min⁻¹. A modelagem aqui relatada foi desenvolvida utilizando o *software* R versão 3.5.1 (R CORE TEAM, 2018).

Tratamento de Dados e Análise Exploratória

O tratamento dos dados brutos, preenchimento de dados faltantes da variável de nível e de acionamento de CMB foi realizado utilizando a última informação disponível. Para o preenchimento de seis informações faltantes foi realizado a interpolação pelo inverso do quadrado da distância utilizando dados diários obtidos das estações pluviométricas de Salvador, Alagoinhas, Feira de Santana e Cruz das Almas/BA, a partir de dados obtidos no Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP) do Instituto Nacional de Meteorologia (INMET).

A análise descritiva e exploratória dos dados foi realizada utilizando o conjunto de dados total para cada variável ou agrupando em menor escala de tempo, como horas ou dias, para análises específicas.

Modelagem Estatística de Ocorrência de Extravasamento



A modelagem preditiva aqui realizada é a de classificação para o dia posterior, pautada na predição da ocorrência ou não de extravasamento de efluente ou não, com abordagem considerando cada dia observado como um objeto.

Foram construídos vinte e quatro modelos de aprendizado de máquina, sendo doze modelos utilizando *Random Forest* e outros doze com *KNN*. Os doze modelos para cada algoritmo, foram divididos em três cenários, onde cada cenário possui um específico conjunto de variáveis independentes, e também foram aplicadas técnicas de reamostragem para compensação de desbalanceamento de informação de classes.

Para minimizar possíveis *bias*, para cada algoritmo e cenário foram construídos modelos preditivos considerando os dados reais e os dados com reamostragem, sendo empregado as técnicas de reamostragem de *under-sampling*, *over-sampling* e ROSE.

O conjunto de variáveis do Cenário 01 é baseado no fenômeno chuva-vazão, e para construção desse modelo foi utilizado apenas a informação de precipitação pluviométrica diária ($V_{c,i-1}$) acumulada [mm/dia] no dia anterior e variáveis relacionadas ao nível da CET 09 (mediana diária do nível, máximo nível diário, desvio-padrão do nível diário e a diferença do *range* do nível diário).

Para o Cenário 2, foi utilizado as mesmas variáveis do Cenário 1 com o incremento de variáveis relacionadas a ocorrência de chuvas. As variáveis inseridas são: volume de precipitação pluviométrica com defasagem de dois dias ($V_{c,i-2}$) [mm], e variáveis condicionais (se, então), como volume acumulado se ocorrido dois dias consecutivos de chuva ($V_{c,i-1} + V_{c,i-2}$), volume acumulado se ocorrido três dias consecutivos de chuva ($V_{c,i-1} + V_{c,i-2} + V_{c,i-3}$) e informação do processo operacional. Para as variáveis condicionais, em caso de ausência de dias consecutivos de chuva, a informação é zero.

A informação do processo operacional é a variável tempo diário de descarga, ou seja, o tempo diário que o total de bombas ficou ligado durante o dia. Esta informação é necessária devido a existência de dois CMB na CET 09, pois a operação ótima e a prática eficaz podem reduzir os riscos de inundação e transbordamentos (YAZDI; CHOI; KIM, 2016). Logo, o tempo diário é o somatório de tempo de trabalho de cada bomba.

A informação da chuva com defasagem de dois dias é importante devido a baixa resolução dos dados de chuva. A chuva acumulada a nível diário dificulta a compreensão quanto a ocorrência e implicação no fenômeno chuva-vazão, uma vez que não se conhece o tempo de concentração da bacia hidráulica. Em anexo é possível visualizar um recorte da estruturação destes dados em uma tabela dinâmica.

O Cenário 3 são as variáveis do Cenário 2 com a adição de variáveis relacionadas ao fenômeno de transbordamento. As variáveis adicionais ao Cenário 3 são: Tempo diário da



CET 09 em transbordamento [min.], volume diário de extravasamento [m^3] e máximo nível [%] diário ($i-1$) no intervalo de 23:00hr à 23:59hr, dia anterior. Esta última, tem como objetivo de estruturar informação para o momento de transição entre os dias.

O tempo e o volume de extravasamento por estruturas unitárias são variáveis que foram amplamente pesquisadas por (YU et al., 2013, 2018), estas variáveis tendem a proporcionar a informação de resposta para variáveis como intensidade e volume de chuvas.

A variável do máximo nível no último intervalo horário do dia foi escolhida em função da observação da frequência de extravasamento em intervalos de hora e da característica de predição diária de extravasamento.

O volume diário de extravasamento foi mensurado realizando balanço de massa considerando que em situações onde o nível da bacia é maior que 95% e o CMB está desligado, ocorre o extravasamento.

Para a construção dos modelos os dados foram pré-processados utilizando a padronização *z-score* com centralização, onde cada modelo foi dividido em 75% dos dados para treinamento e 25% para teste. Para validação estatística foi realizado a validação cruzada *k-fold*, sendo $k = 10$.

Avaliação dos Modelos de Predição

A predição para classificação utilizando *KNN* e *Random Forest* fornece resultados probabilísticos (votos) para as classes. E, no presente trabalho convencionou-se utilizar o *threshold* (limiar) padrão, onde para discriminação de dois grupos a regra de classificação é caso a probabilidade de ocorrência de extravasamento fosse maior que 50%, então classifica-se a predição efetivamente como extravasamento, em caso contrário classifica-se como operação normal.

Para avaliar a performance dos modelos preditivos construídos foram utilizados os scores da matriz de confusão de cada modelo, conforme exemplificado na Tabela 1.

Tabela 1 – Matriz de confusão obtida após teste do modelo preditivo.

Evento predito	Evento observado	
	Extravasamento	Operação Normal
Extravasamento	Verdadeiro Positivo (VP)	Falso Positivo (FP)
Operação	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Fonte: Os Autores, 2019

Delineado o evento de interesse, que é o fenômeno de extravasamento da CET 09, a predição correta deste é um verdadeiro positivo (VP), FP é um falso positivo (quando o modelo incorretamente prevê extravasamento e o real é operação normal), VN é um verdadeiro



negativo (corretamente predito a operação normal operação), e FN é um falso negativo (quando uma predição de operação normal de bem é realizada quando tem ocorrência de extravasamento).

Organizando as informações de categorias preditas e observadas na matriz de confusão, foi empregado as métricas de Acurácia, Sensibilidade, Especificidade, *Kappa* (COHEN, 1960), *F1 Score* e ORSS (*Odds Ratio Skill Score* – índice de taxa de probabilidade) foram empregados para avaliação da escolha dos modelos.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} = \frac{VP + VN}{n} \quad (1)$$

$$Sensibilidade = \frac{VP}{VP + FN} \quad (2)$$

$$Especificidade = \frac{VN}{VN + FP} \quad (3)$$

$$Kappa = \frac{p_o - p_e}{1 - p_e} \quad (4)$$

$$\therefore p_o = Acurácia \wedge p_e = \left(\frac{VP + FP}{n} \times \frac{VP + FN}{n} \right) + \left(\frac{FN + VN}{n} \times \frac{FP + VN}{n} \right)$$

$$F1 = 2 \times \frac{\frac{VP}{VP + FP} \times Sensibilidade}{\frac{VP}{VP + FP} + Sensibilidade} \quad (5)$$

$$ORSS = \frac{(VP \times VN - FP \times FN)}{(VP \times VN + FP \times FN)} \quad (6)$$

O ORSS é um dos índices a ser aplicado para avaliar modelagem que envolve dados predições relacionada a climatologia (STEPHENSON; THORNES, 2001), sendo este índice também implementado no trabalho de (MAILHOT; TALBOT; LAVALLÉE, 2015) para predição de CSO.

O algoritmo *KNN* foi implementado utilizando Distância Euclideana para o cálculo de similaridade entre objetos, e o número de vizinhos mais próximos (*K*) compreende o intervalo de $K = \{1, 2, \dots, 24, 25\} \in \mathbb{N}$, que é um intervalo que engloba o possível máximo número de vizinhos ótimos conhecidos empiricamente e citado por (HASSANAT; ABBADI; ALTARAWNEH, 2014). Para o algoritmo *Random Forest* foi considerado o número de 1500 árvores aleatórias, e o número de preditores (*mtry*), disponível para divisão em cada nó das árvore, foi concebido pelo pacote *caret* (KUHN, 2018).

R Packages

O principal pacote utilizado para este estudo foi o *caret* (KUHN, 2018), pois a partir dele foi realizado a construção e avaliação dos modelos desenvolvidos. Pacotes como *dplyr*



(WICKHAM et al., 2019), *tidyr* (WICKHAM; HENRY, 2019) e *data.table* (DOWLE; SRINIVASAN, 2019) foram utilizados na manipulação e estruturação de dados. O pacote *psych* (REVELLE, 2018) foi utilizado para obter informações de estatística descritiva básica, bem como o as funções bases (*base*) do R e do pacote *stats* (R CORE TEAM, 2018).

Resultados e Discussão

Dos 257 dias em análise, em 154 dias houve ocorrência de qualquer precipitação pluviométrica (> 0 mm), entretanto as chuvas ocorridas na região da refinaria foram em sua maioria de baixo volume, configurando uma maior densidade a esquerda da distribuição, leptocúrtica com assimetria positiva, como pode ser visualizado na Figura 1.

Verificou-se também que nos meses centrais do ano há uma maior distribuição de chuvas com tendência a altura diária de 25 mm. No período que compreende os meses de Abril à Junho são os meses onde ocorrem as maiores alturas de chuva, como também pode ser interpretado na Figura 2.

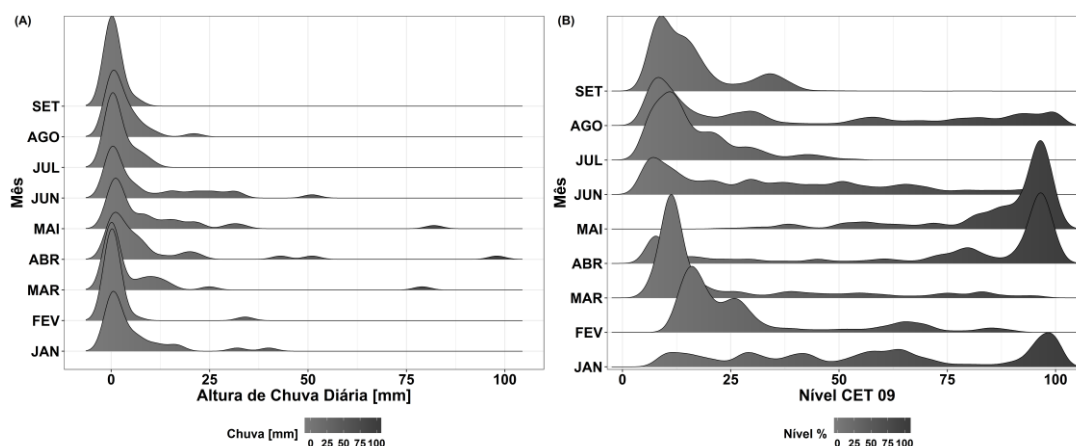


Figura 1 – Distribuições de densidade. (A) Chuva mensal [mm]. (B) % Nível CET 09 mensal

Fonte: Os Autores, 2019

O nível da CET 09 tem um outro perfil de distribuição, o nível da bacia pode ser estratificada em níveis baixos e níveis altos, indicando que níveis centrais são apenas transitórios, ou seja, não tendo um período de estacionariedade para estes.

Na Figura 1 percebe-se também que a mudança da assimetria da distribuição do nível da CET 09 ocorre de maneira mais acentuada nos períodos com maiores índices pluviométricos.



Quanto a operação do sistema de drenagem, apenas os meses de Julho e Setembro os CMB tem um tempo maior de *stand-by* ao que o tempo de trabalho, Figura 2. Tal fato que simboliza a redução dos dias de chuvas nesses meses durante esse período central do ano.

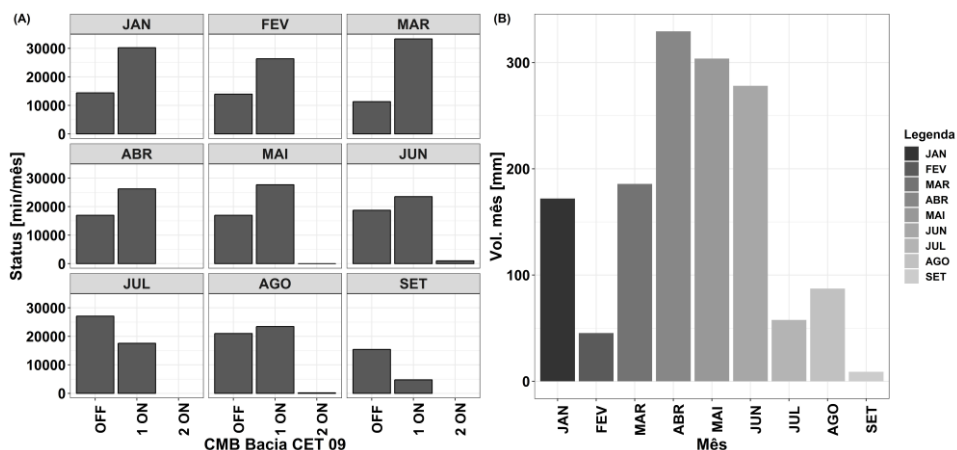


Figura 2 – Frequências acumuladas observadas (A) Tempo de trabalho (ONN) e tempo em *stand-by* (OFF) do CMB na CET 09. (B) Volume acumulado de chuva [mm]

Fonte: Os Autores, 2019

A partir do resultado do balanço de massa para verificação de ocorrência de extravasamentos, percebeu-se que a maior frequência de extravasamentos está predominantemente no período que compreende as maiores alturas de chuva [mm/dia]. Percebe-se também que há uma redução do tempo de trabalho dos CMB nos períodos em questão. Figura 3. Essa não conformidade operacional reflete possivelmente a importância da operação ótima do CMB, como sinalizado por (YAZDI; CHOI; KIM, 2016).

A representação da série histórica do nível transformado para mediana é uma boa representação do comportamento do fenômeno. Isto acontece devido a assimetria da distribuição de probabilidade.

Com a análise exploratória também foi possível avaliar da CET 09 conforme metodologia de Montserrat et al. (2015) construindo os gráficos do volume de chuva [mm] *versus* informação sobre extravasamento, Figura 4. Percebe-se que ocorreu extravasamento em dias com eventos de chuva menor que 2 mm, caracterizando que a CET 09 não tem capacidade de assimilar aumento de vazão em episódios de chuva.

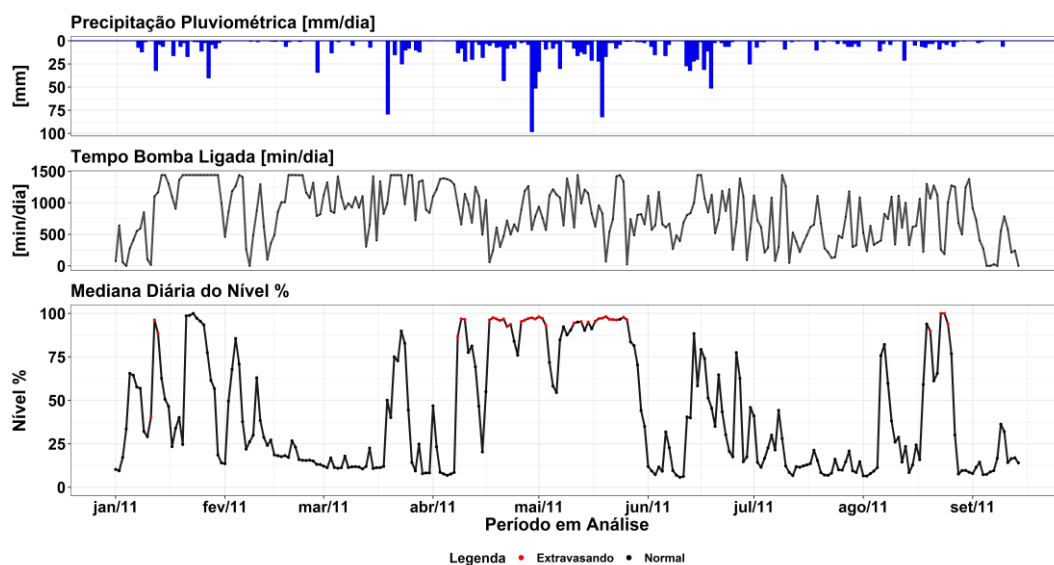


Figura 3 – Séries temporais de altura de chuva [mm/dia], mediana do % do nível, tempo de trabalho dos CMB [min/dia]

Fonte: Os Autores, 2019

Os resultados dessa análise descritiva auxiliaram na definição de variáveis da modelagem, previamente apresentadas na seção Modelagem Estatística de Ocorrência de Extravasamento. Na próxima seção apresentaremos os resultados da modelagem e previsão do comportamento dos eventos de extravasamento. As informações aqui apresentadas e os códigos desenvolvidos em R podem ser encontradas nos Anexos deste documento.

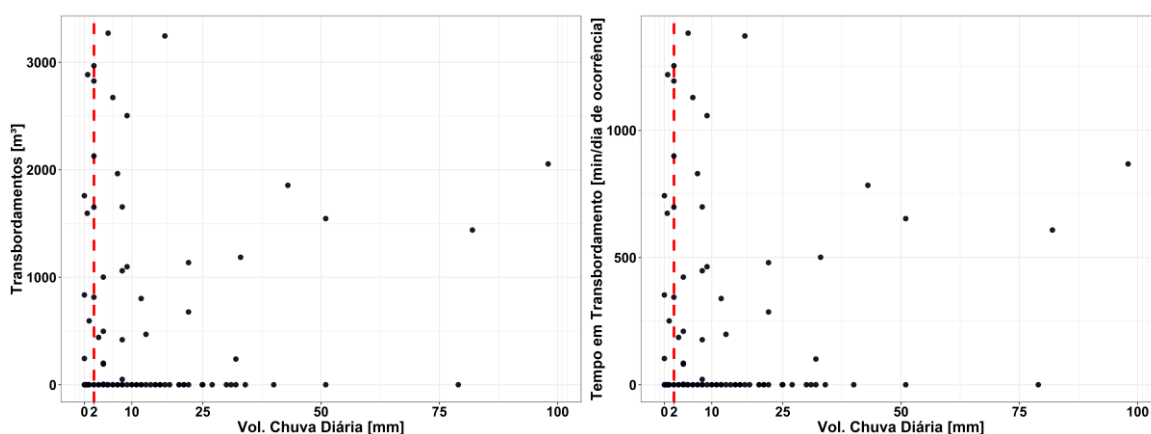


Figura 4 – Verificação do limiar de extravasamento

Fonte: Os Autores, 2019



Modelagem Preditiva

De posse dos atributos independentes, e do atributo dependente categórico de interesse, foram construídos 24 modelos preditivos utilizando *Random Forest* (RF) e *KNN*. Para cada modelo foi gerado uma matriz de confusão com *scores* de VP, VN, FP, que podem ser visualizadas no material em Anexo.

No primeiro momento, devido ao desbalanceamento de informação (14,4% Extravasando *versus* 85,6% Operação Normal), as principais métricas para avaliação inicial foram *Kappa*, sensibilidade e acurácia. A escolha da métrica *Kappa* é importante porque apresenta ponderação sobre todos os *scores* da matriz de confusão.

Como definido na metodologia, três cenários foram definidos de acordo com o conjunto de variáveis empregadas para a construção de cada modelo.

Para os modelos construídos no cenário 1 os melhores resultados foram para os modelos com o algoritmo *KNN* forma com as técnicas de reamostragem *undersampling* e ROSE. Ambos tiveram VP igual a 9, e ausência de *score* para FN, ou seja, todas as predições de extravasamento foram realizadas totalmente de acordo com o ocorrido histórico. Entretanto o modelo com ROSE teve também um menor *score* de FP, mas possui elevado número ótimo de vizinhos ($K=15$), refletindo a complexidade do modelo e um possível *overfitting*.

Dos modelos RF no cenário 1, destacam-se os modelos sem emprego de técnica de reamostragem e com *oversampling*. Ambos tem bom desempenho para o índice *Kappa* (aproximadamente igual a 0,75) e acurácia de 0,93, e possuem, respectivamente, o bom *score* com FN igual a 2 e 1. Os modelos com técnicas de *undersampling* e ROSE predizem de maneira assertiva todas as observações de extravasamento, mas contam com demasiado FN.

No cenário 2 com *KNN* os modelos apesar de apresentarem excelentes resultados de acurácia, aproximadamente 0,90, possuem baixos valores de sensibilidade. Esta situação reflete um alto *score* de FN, o que é extremamente indesejável no contexto de predição de extravasamento.

Como os modelos RF no cenário 2 ocorre algo similar para o modelo sem emprego de técnica de reamostragem e o modelo com *oversampling*, que é concretizado pelos módicos valores de *Kappa* (0,57 e 0,56). O modelo com *undersampling* e ROSE tem bom acurácia e sensibilidade, mas aponta um desequilíbrio com valores de *Kappa* menor que 0,60.

Com o conjunto de variáveis no cenário 3 utilizando *KNN* apenas o modelo com *undersampling* apresentou índices satisfatórios, entretanto o elevado número de vizinho, $K=24$, torna o modelo muito complexo. Diferentemente do que acontece com RF com *undersampling* e ROSE, que apresentaram resultados excelentes, bem equilibrados. O modelo RF com *undersampling* se destaca também por apresentar $mtry = 2$, ou seja, apenas dois nós são necessários em cada árvore de decisão para que o modelo possa prever o resultado.

Na Figura 5 são apresentados os resultados para os modelos construídos com resultados de maior destaque, destacando que todos os modelos com os melhores resultados são baseados no algoritmo *Random Forest* e com emprego de técnica de reamostragem, ratificando a informação de melhor desempenho da RF frente ao *KNN* (JAMES et al., 2013).

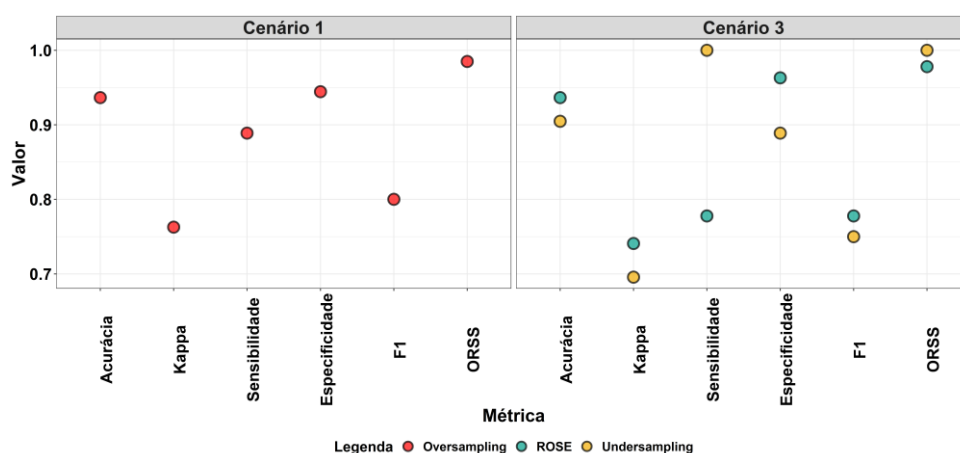


Figura 5 – Análise dos principais modelos preditores construídos

Fonte: Os Autores, 2019

Os modelos escolhidos apresentam excelentes resultados, todos com acurácia $> 0,90$ e bom equilíbrio entre as demais métricas. Entretanto o modelo RF com técnica de *undersampling* com as variáveis do cenário 3 apresenta sensibilidade igual a 1, ou seja, ele acerta todos os momentos que ocorre o fenômeno de extravasamento na CET 09, como pode ser visto na Tabela 2.

É de extrema importância a ocorrência de baixos valores ou a ausência de Falsos Negativos em modelagens desse tipo, pois o ideal é que caso ocorra falha de predição, a falha deve ser um FP. Esta ação visa proporcionar uma boa ferramenta preditiva que forneça suporte a tomada de decisão aos profissionais envolvidos na operação do sistema de drenagem industrial.



A interpretação de modelagem preditiva como suporte a tomada de decisão é importante também devido a característica estocástica da principal variável externa que promove perturbações na CET 09, a chuva, possíveis incertezas que rondam o modelo, e a possibilidade influenciar possíveis regras e atividades operacionais na indústria, pois diferentemente do que acontece no CSO, o sistema de drenagem pluvial em área industrial é monitorado e operado.

Conforme pode ser visualizado na matriz de confusão apresentada na Tabela 2, o modelo apresenta excelente sensibilidade (máximo VP), elucidando novamente a grande capacidade preditiva do modelo.

Tabela 2 – Matriz de confusão para RF *undersampling* no cenário 3.

Evento predito	Evento observado	
	Extravasamento	Operação Normal
Extravasamento	9	6
Operação Normal	0	48

Fonte: Os Autores, 2019

Observou-se que o modelo RF discutido na Tabela 2 classificou corretamente o fenômenos de extravasamento com boa distinção de classes, promovendo verdadeiros positivos com valores de probabilidade entre ~0,70 e ~1,0, revelando uma boa segmentação.

A variável mais importante para o modelo preditivo escolhido é o máximo nível percentual diário da CET 09 no intervalo de 23:00 às 23:59, como pode ser visualizado na Figura 6 baseado no Índice Gini.

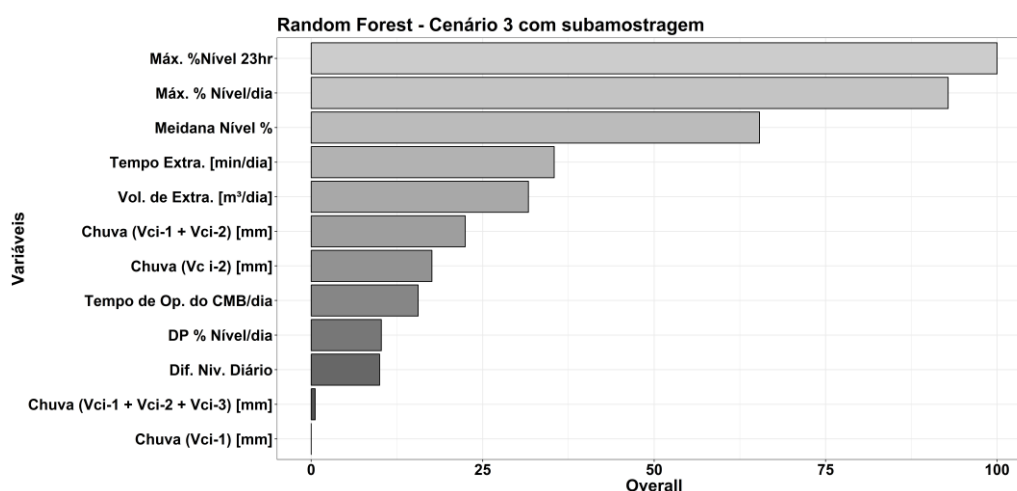


Figura 6 – Valores de importância por variável

Fonte: Os Autores, 2019



Os baixos valores de importância para chuva diária ($V_{C_{i-1}}$) acumulada [mm/dia] no dia anterior de refletem a necessidade de um monitoramento mais detalhado desse fenômeno. E as variáveis mais importantes ratificam que o sistema extravasa devido aos altos níveis de efluente acumulado.

Conclusão

A complexidade e a não linearidade que envolvem o fenômeno de extravasamento, e a ausência de dados relacionados a precipitação pluviométrica em menor escala de tempo tornam a modelagem empírica deste fenômeno um verdadeiro desafio. Não obstante, foi realizado a modelagem com horizonte preditivo para o dia posterior quanto a classificação se a bacia de contenção de efluentes em uma refinaria, CET 09, vai extravasar ou não.

Os melhores resultados de modelos construídos para a modelagem sobre a CET 09 foram obtidos a partir do algoritmo *Random Forest* com emprego de técnica de *oversampling*, *undersampling* e ROSE. O melhor resultado foi para o conjunto de variáveis do cenário 3 com o emprego de técnica de reamostragem do tipo *undersampling*.

Contudo, com o intuito de proporcionar melhor *benchmarking*, indica-se a necessidade de realizar novos estudos a partir de séries temporais mais longas, realizar o emprego da técnica de regressão logística de modo a proporcionar interpretação direta aos coeficientes de regressão.

Agradecimentos

O presente trabalho foi apoiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo Programa de Pós-Graduação em Engenharia Industrial (PEI) e pelo grupo GAMMA (*Growing with Applied Modeling and Multivariate Analysis*) da Universidade Federal da Bahia.

Referências

- BREIMAN, L. RANDOM FORESTS Leo. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.
- COHEN, J. **A Coefficient of Agreement for Nominal Scales**. **Educational and Psychological Measurement**, v. 20, n. 1, p. 37–46, 2 abr. 1960.
- DOWLE, M.; SRINIVASAN, A. **data.table: Extension of `data.frame`**, 2019. Disponível em: <<https://cran.r-project.org/package=data.table>>
- GE, Z. et al. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. **IEEE Access**, v. 5, p. 20590–20616, 2017.



- HASSANAT, A. B.; ABBADI, M. A.; ALTARAWNEH, G. A. Optimal K parameter for KNN Classifier with square root. **International Journal of Computer Science and Information Security**, v. 12, n. 8, p. 33–39, 2014.
- HODGSON, J. E.; BENDIAK, L. C. Stormwater Management for Petroleum Refineries. **Canadian Water Resources Journal**, v. 12, n. 3, p. 38–47, jan. 1987.
- JAMES, G. et al. **An Introduction to Statistical Learning**. New York, NY: Springer New York, 2013. v. 103
- KUHN, M. **caret: Classification and Regression Training**. R package version 6.0-81, 2018. Disponível em: <<https://cran.r-project.org/package=caret>>
- MAILHOT, A.; TALBOT, G.; LAVALLÉE, B. Relationships between rainfall and Combined Sewer Overflow (CSO) occurrences. **Journal of Hydrology**, v. 523, p. 602–609, abr. 2015.
- MONTSERRAT, A. et al. Using data from monitoring combined sewer overflows to assess, improve, and maintain combined sewer systems. **Science of the Total Environment**, v. 505, p. 1053–1061, 2015.
- PAZ, A. R. **Hidrologia Aplicada**. Caxias do Sul (RS): UERGS, 2004.
- R CORE TEAM. R: **A Language and Environment for Statistical Computing**. Vienna, Austria. R Foundation for Statistical Computing, 2018. Disponível em: <<https://www.r-project.org>>
- REVELLE, W. psych: **Procedures for Psychological, Psychometric, and Personality Research**. Evanston, Illinois Northwestern University, 2018. Disponível em: <<https://cran.r-project.org/package=psych>>
- SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados Com Aplicações em R**. 1. ed. ed. Rio de Janeiro: Elsevier, 2016.
- STEPHENSON, D. B.; THORNES, J. How to judge the quality and value of weather forecast products. **Meteorological Applications**, v. 8, p. 307–314, 2001.
- TANTITHAMTHAVORN, C.; HASSAN, A. E.; MATSUMOTO, K. The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models. **IEEE Transactions on Software Engineering**, p. 1–20, 2018.
- WICKHAM, H. et al. **dplyr: A Grammar of Data Manipulation**, 2019. Disponível em: <<https://cran.r-project.org/package=dplyr>>
- WICKHAM, H.; HENRY, L. **tidyr: Easily Tidy Data with “spread()” and “gather()” Functions**, 2019. Disponível em: <<https://cran.r-project.org/package=tidyr>>
- YAZDI, J.; CHOI, H. S.; KIM, J. H. A methodology for optimal operation of pumping stations in urban drainage systems. **Journal of Hydro-environment Research**, v. 11, p. 101–112, jun. 2016.
- YU, Y. et al. Cluster analysis for characterization of rainfalls and CSO behaviours in an urban drainage area of Tokyo. **Water Science and Technology**, v. 68, n. 3, p. 544–551, ago. 2013.
- YU, Y. et al. Simple Method for Calculating Hydraulic Behavior of Combined Sewer Overflow from Rainfall Event Data. **Journal of Water Resources Planning and Management**, v. 144, n. 10, p. 04018061, out. 2018.
- ZHAO, W.; BEACH, T. H.; REZGUI, Y. Automated Model Construction for Combined Sewer Overflow Prediction Based on Efficient LASSO Algorithm. **IEEE Transactions on Systems, Man, and Cybernetics: Systems**, p. 1–16, 2017.

Anexo

Todo o código desenvolvido para este artigo bem como outras figuras podem ser acessados a partir do *link* disponibilizado abaixo:

http://rpubs.com/brennerbiasi/Silva_etal_SER_IV_2019