Análise de Dados Aplicada às Chamadas de Emergência em um Sistema de Distribuição Real

Arthur H. M. de Oliveira [*] **Karcius Dantas** [*] **Matheus Maia** [*]
Kézia Dantas [**] **Bruna Granadeiro** [***] **Raquel Zacarias** [***]

[*] *Universidade Federal de Campina Grande, PB (e-mails: arthur.mendes@ee.ufcg.edu.br, karcius@dee.ufcg.edu.br, matheus.maia@ccc.ufcg.edu.br)*
[**] *Universidade Estadual da Paraíba, PB (e-mail: kezia.vasconcelos@gmail.com)*
[***] *Light SA, RJ (e-mails: bruna.granadeiro@light.com.br, raquel.souza@light.com.br)*

Abstract: This work uses data analysis related to emergency calls from customers of a power distribution utility in Brazil, in order to investigate how temporal aspects and climatic characteristics influence the demands of the electricity grid and how these aspects are correlated with the number of calls received in a power distribution system over time. Therefore, a statistical analysis was performed based on the calculation of Pearson's coefficient correlation for data with numerical characteristics and on the chi-square test that verifies the association of categorical characteristics with the number of calls. Numerical and day-of-the-week classifications of climatic data were proposed. As a result, it was found that the impact of these manipulations has a greater influence on the number of calls received by the concessionaire, allowing a better understanding of the most significant root causes for the increase in the number of calls in energy distribution systems in the country.

Resumo: Este trabalho utiliza análise de dados relacionados às chamadas de emergência de clientes de uma concessionária de distribuição de energia no Brasil, a fim de investigar como os aspectos temporais e as características climáticas influenciam nas demandas da rede elétrica e como estes aspectos estão correlacionados com a quantidade de chamadas recebidas em um sistema de distribuição de energia ao longo do tempo. Para tanto, foi realizada uma análise estatística baseada no cálculo de correlação do coeficiente de Pearson para os dados com características numéricas e no teste qui-quadrado que verifica a associação das características categóricas com a quantidade de chamadas. Foram propostas categorizações dos dados climáticos de natureza numérica e dos dias da semana. Como resultado, foi constatado que o impacto dessas manipulações tem maior influência na quantidade de chamadas recebida pela concessionária, permitindo uma compreensão melhor sobre as causas raízes mais significativas para o aumento do número de chamadas em sistemas de distribuição de energia no país.

Keywords: Data analysis; Big Data; Electrical systems; Power distribution; Emergency calls.
Palavras-chaves: Análise de dados; Big Data; Sistemas elétricos; Distribuição de energia; Chamadas de emergência.

1. INTRODUÇÃO

Nas últimas décadas, o rápido desenvolvimento econômico e social fizeram aumentar os requisitos de segurança da rede elétrica e da qualidade no fornecimento de energia (Cai, 2016). Vários fatores tornam o ambiente operacional complexo, acarretando desafios para o setor. Por exemplo, as interrupções de energia causadas pela exposição do sistema ao ambiente hostil, são responsáveis por até 90% de todos os problemas de confiabilidade do cliente (Brown, 2002). A fim de minimizar as perdas induzidas por falhas e melhorar a credibilidade da distribuição é necessário identificar as principais causas das ocorrências para restaurar o serviço em tempo hábil. Por isso, a integração de novos dispositivos que permitem o monitoramento e controle da rede vêm sendo utilizados em sistemas de gerenciamento de interrupção (OMS - *Operation Management System*) de concessionárias por anos para facilitar as reações e restabelecer o serviço pós-falha. Além disso, alarmes de relés, sistemas de medição (SCADA - *Supervisory Controland Data Acquisition*), mensagens de leitura automatizada de medidores (AMR - *Automated Meter Reading*) e chamadas de clientes do sistema de distribuição também são usadas para inferir seções de falha (Sridharan and Schulz, 2001).

Uma vez que o fator de identificação da causa raiz varia de forma complexa de acordo com o comportamento estocástico da natureza nas áreas de serviço, torna-se difícil identifica-lo apenas por medições de grandezas elétricas.

[⋆] O presente estudo foi realizado no âmbito do Programa de P&D da ANEEL, PD-00382-0132/2020, Light SA/UEPB/UFCG/PaqTc-PB.

Muitas vezes para levantar hipóteses sobre a ocorrência do evento é necessário se deslocar até o local da falha e coletar evidências ao longo de quilômetros da linha de energia sobre quais aspectos possam justificar o acontecimento, baseados na observação e experiência humana. Nesse contexto, os dados armazenados nos sensores instalados em estações meteorológicas, que informam características sobre o clima de determinada região, podem descrever comportamentos que auxiliem na compreensão do problema. Essa situação oferece ao setor elétrico a oportunidade de utilizar análise de grandes dados (BDA - *Big Data Analytics*) para auxiliar os operadores do sistema na tomada de decisão em tempo real e na previsão de novas ocorrências nas redes de eletricidade (Tu et al., 2017).

Vários estudos têm sido realizados na área de sistemas elétricos de potência utilizando *Big Data* em diversos contextos diferentes. Por exemplo, os trabalhos de Xu and Chow (2006) e de Xu et al. (2006) têm por objetivo classificar a causa raiz de falhas em um sistema de distribuição devido à causas ambientais; Cai and Chow (2009) apresenta um diagnóstico de ocorrências da rede fazendo uso de relações espaciais e temporais a partir de uma estrutura de georreferenciamento (GIS); Doostan and Chowdhury (2017) abordam um método para identificar falhas em equipamentos no sistema de distribuição por meio de um classificador binário, onde as interrupções são categorizadas em com ou sem falha de equipamento; Sahai and Pahwa (2006) realizam um estudo sobre a predição de faltas com causa de origem animal em uma concessionária.

Diante do exposto, este trabalho tem o objetivo de utilizar técnicas de BDA para análise de dados reais relacionados às chamadas de emergência de clientes de uma concessionária de distribuição de energia no Brasil, a fim de investigar como os aspectos temporais e as características climáticas influenciam nas demandas da rede elétrica e como estes aspectos estão correlacionados com a quantidade de chamadas recebidas ao longo do tempo.

2. FUNDAMENTAÇÃO TEÓRICA

A necessidade de manusear grandes quantidades de dados, mesmo que de forma rudimentar, encontra-se presente desde muitos anos atrás. É possível encontrar registros no decorrer das décadas de atividades que trazem consigo o conceito do que seria posteriormente abordado como *Big Data*. Por exemplo, o mecanismo criado em 1943 para interceptar, processar e decifrar códigos nazistas durante a segunda guerra mundial ou o *Data Center* desenvolvido pelos Estados Unidos na década de 60, com o objetivo de armazenar declarações fiscais e impressões digitais (Kraus, 2013). Nos últimos anos, devido ao crescimento acelerado da quantidade de dados produzidos, bem como da capacidade computacional disponível, a discussão acerca da utilização de métodos para processamento de quantidades massivas de informações começou a ter maior importância, visto que soluções tradicionais de análise e tratamento de dados passaram a ser insuficientes.

Ao se avaliar os desafios e oportunidades relacionados à aplicação dessa tecnologia, pode-se destacar a definição proposta por Douglas (2001) dos 3Vs (Volume, Velocidade e Variedade). Essa nomenclatura foi amplamente utilizada e adaptada por pesquisadores e grandes empresas, como a IBM e Microsoft. Apesar desta definição inicial, inúmeros outros estudos abordam o assunto e defendem a expansão para novas classificações, incluindo, por exemplo, os conceitos de Valor e Veracidade, expandindo o modelo para 4Vs ou 5Vs (Zhou et al., 2016; Wen et al., 2018; Codal and İlter, 2020). A seguir, são apresentadas definições para cada um dos pilares que compõem os 5Vs:

- **Volume**: geração de grandes quantidades de dados, provenientes da introdução de dispositivos e sensores na rede elétrica, com outras fontes de informação.
- **Velocidade**: refere-se à rapidez do processamento de dados para garantir o equilíbrio entre oferta e demanda do serviço em tempo real.
- **Variedade**: refere-se à diversidade das fontes de dados, podendo ser incluídas mídias sociais, dados meteorológicos, além das grandezas comuns ao setor elétrico, como tensões, correntes, frequência e potência.
- **Veracidade**: refere-se à quão bem os dados representam as grandezas reais.
- **Valor**: refere-se a busca de informações valiosas em um conjunto de dados massivos.

No setor elétrico, a integração de diversos e modernos equipamentos de campo, com taxas de aquisição de dados cada vez mais altas e provenientes de diferentes fontes, acarretou no aumento da complexidade das redes de distribuição e transmissão. Isso fez o emprego de técnicas de BDA tornar-se uma necessidade. Hoje em dia, as redes de energia elétrica incluem inovações nas áreas de medição, controle, comunicação e ciência da informação para buscar fornecimento confiável e de qualidade ao consumidor final (Bhattarai et al., 2019). Dentre as principais aplicações no contexto de sistemas elétricos de potência, percebem-se grandes contribuições acadêmicas voltadas para as áreas de detecção de faltas (Huang et al., 2019), atípicos e padrões(Alves et al., 2017), estimação de estado (Peppanen et al., 2015), gerenciamento e planejamento de operação (Zhou et al., 2016), detecção de fraudes (Rocha et al., 2016), previsão de carga e padrões de consumo (Li et al., 2016), resposta a demanda (Munshi and Yasser, 2017), entre diversos outras.

3. METODOLOGIA

Neste capítulo serão introduzidos os processos necessários para a realização da análise dos dados. Detalhes das etapas são apresentados a seguir no fluxograma da Figura 1.

Figura 1. Fluxograma das principais etapas realizadas.

3.1 Coleta de dados

Inicialmente foram coletadas amostras ao longo do ano de 2020 dos registros das chamadas dos clientes em um

banco de dados de uma distribuidora de energia situada no Estado do Rio de Janeiro, Brasil. Os dados consistem em relatos de ocorrências com respeito às condições de operação do sistema, e são provenientes de atendimento virtual, por meio de chamadas via SMS, twitter, facebook, ou físico, pelas chamadas telefônicas atendidas pelo *call center*, que registram as solicitações na base de dados da empresa.

Posteriormente foi realizada uma busca à respeito de dados provenientes de estações meteorológicas. O site do Instituto Nacional de Meteorologia (INMET) [1] foi escolhido, pois registra informações que podem ser adquiridas via Interface de Programação de Aplicação (API - *Application Programming Interface*). Dessa forma, foram realizadas coletas ao longo do mesmo período de tempo das chamadas dos clientes da concessionária. Extraiu-se das bases disponíveis na pesquisa do site do INMET as características mostradas na Tabela 1.

Tabela 1. Principais *features* disponíveis no site do INMET.

Características	Unidade de medida
Precipitação por hora*	mm
Temperatura Máxima por hora*	°C
Velocidade Máxima do Vento por hora*	m/s

* Fuso horário UTC +0 (*Coordinated Universal Time*).

A Figura 2 mostra a representação das fontes dos dados coletados.

Figura 2. Bases coletadas.

3.2 Análise e tratamento de dados

Dentre diversas informações disponíveis na base de dados da empresa, a fim de melhor caracterizar os eventos, foram coletados três tipos de dados: hora e data da ocorrência, subestação a qual a ocorrência está vinculada e o tipo de ocorrência. A Tabela 2 mostra as características utilizadas.

Tabela 2. Colunas utilizadas da base de chamadas.

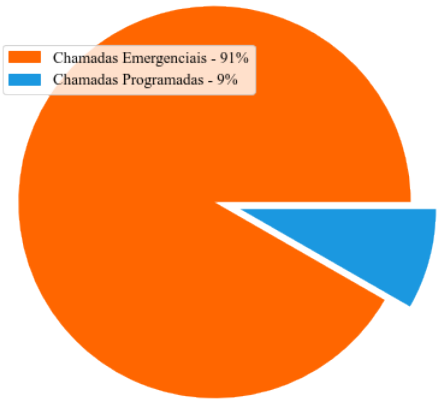
Data e hora da reclamação	Subestação	Tipo de ocorrência
dh_recla	*subesta*	*tipo_ocorrencia*

Dentre essas informações, verificou-se que 4,97% dos valores para subestações eram nulos. Também foi feito um levantamento sobre a quantidade de tipos das características e notou-se 110 subestações e 2 tipos de ocorrências, emergenciais e programadas, como mostra a Tabela 3. Além disso, observou-se a proporção dos dados que contêm natureza emergencial, totalizando 91%, conforme ilustrado na Figura 3.

Tabela 3. Análise de dados da base de chamadas.

	Valores nulos (NaN's)	Tipos
Data e Hora da Reclamação	-	
Subestação	4,97%	110
Tipo de ocorrência	-	2

Figura 3. Percentual de chamadas de emergência.



Atendendo às observações mencionadas antes, foram aplicados filtros no banco de registro das chamadas para remover os valores nulos da característica *"Subestação"* e para remover os dados que não apresentam natureza emergencial na característica *"Tipo de ocorrência"*.

Logo após à aplicação desses filtros, foi extraído da coluna data e hora, a informação específica sobre a hora que ocorreu o evento, posteriormente utilizada junto à análise dos dados meteorológicos para visualizar o comportamento dessas características durante a ocorrência. Além disso, foram adicionadas novas colunas com informações sazonais para investigar padrões de comportamento do consumidor e demanda do serviço ao longo do dia da semana, do mês específico e da estação do ano. Assim, obteve-se as seguintes novas colunas para a base de dados:

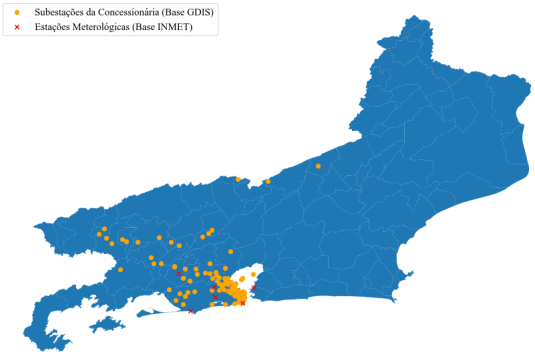
- **Hora do dia** (00:00, 01:00, 02:00, 03:00, etc.);
- **Dia da semana** (Segunda, Terça, Quarta, etc.);
- **Mês do ano** (Janeiro, Fevereiro, Março, etc.);
- **Estação do ano** (Verão, Outono, Inverno, Primavera).

Em seguida foram adicionadas as localizações das subestações, em termos de latitude e longitude. A Figura 4 mostra essa disposição juntamente com as bases meteorológicas disponíveis no site do INMET ao longo do Estado do Rio de Janeiro.

Notou-se que a quantidade de chamadas na região metropolitana era notoriamente superior quando comparada à regiões mais afastadas ou interioranas. Por isso, ao observar a disposição das bases disponíveis do INMET, verificou-se que a maioria estava próxima da região metropolitana. Esse fator foi levado em consideração a fim de reduzir o risco da quantidade de dados da zona interiorana prejudicar a análise realizada, uma vez que suas subestações estão distantes das estações meteorológicas dispo-

[1] https://portal.inmet.gov.br/

Figura 4. Base Cartográfica.



níveis. Essa distância representa a veracidade dos dados referentes ao comportamento meteorológico da região.

Visto que as estações do banco INMET passam por períodos de manutenção ao longo do ano, e dessa forma os servidores são desligados, há necessidade de completar as lacunas referentes aos dados faltantes, que muitas vezes perduram por um período maior de tempo, como por exemplo, 15 dias. Uma opção para o preenchimento desses valores nulos, estratégia utilizada neste trabalho, é verificar as informações da estação meteorológica mais próxima disponível, dado que a retirada de um longo período de tempo de dados prejudicaria as análises. Essa manipulação faz com que a distância entre as estações meteorológicas também seja um fator de confiabilidade sobre o comportamento das características da região.

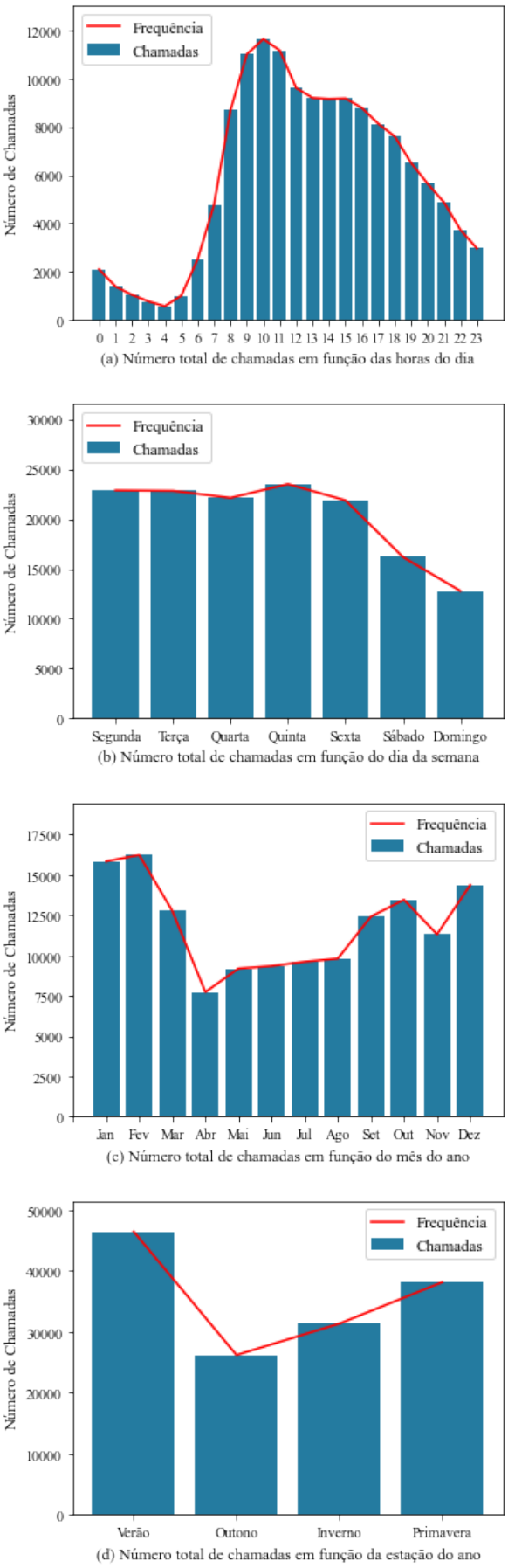
3.3 Junção das bases

Após realizar analises e as devidas manipulações nos dados das duas bases, o próximo passo foi unir essas características para extrair informações do grande volume de dados. Assim, foram utilizados os valores da latitude e longitude das subestações da concessionária a fim de buscar a estação meteorológica mais próxima da ocorrência, obtendo mais precisão sobre o comportamento climático de determinada região. Dessa forma, todo evento registrado no sistema terá as respectivas informações climáticas associadas pela distância mais próxima aquela chamada para a realização das análises necessárias.

4. RESULTADOS

A fim de investigar o comportamento do número de chamadas, foram ilustradas na Figura 5 características sazonais como a hora do dia, o dia da semana, o mês e a estação do ano em relação a quantidade de chamadas. A Figura 5(a) mostra a disposição das chamadas ao decorrer das horas do dia. Nota-se que no período da madrugada o número de chamadas é menor que no período comercial (8 às 18 horas), com aumento ao longo da manhã e redução à noite. Já as Figuras 5(c) e 5(d) mostram um comportamento parecido entre si. Pode-se destacar que a quantidade de chamadas que ocorrem no Verão é maior que das demais estações do ano. Diversos fatores podem influenciar nessa sazonalidade, como período de férias, maior movimentação turística na região, eventos como Carnaval, características climáticas para cada mês e estação do ano, entre outros.

Figura 5. Análise de características sazonais.



(a) Número total de chamadas em função das horas do dia

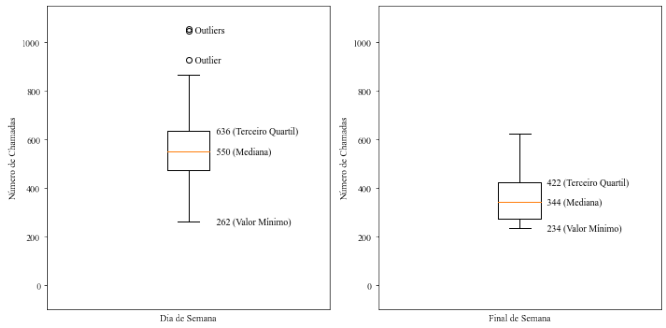
(b) Número total de chamadas em função do dia da semana

(c) Número total de chamadas em função do mês do ano

(d) Número total de chamadas em função da estação do ano

Analisando a curva de comportamento da Figura 5(b) referente aos dias da semana, percebe-se que existe uma queda do número de chamadas total dos dias de semana (ou dias comerciais) em comparação com os dias do final de semana (sábado e domingo). A fim de explorar essa característica, a Figura 6 apresenta um boxplot com a dispersão do número de chamadas dessa distinção.

Figura 6. Boxplot do número de chamadas diárias em função dos dias de semana ou final de semana.



Ao observar o comportamento da mediana das chamadas, nota-se que os dias de semana apresentam um valor superior ao final de semana, em torno de 100 chamadas diárias a mais. Percebe-se também que existe uma diferença nos valores de pico das chamadas e na presença de *outliers* durante a semana, com valores superiores até a mil chamadas em um único dia. Esse tipo de informação demonstra a necessidade de considerar o agrupamento da coluna *"dia da semana"* em uma nova coluna que discrimina de forma binária se é final de semana ou não, visando um impacto maior dessa característica em relação ao número de chamadas do que considerar a setorização entre os sete dias da semana como seria o usual. A Tabela 4 mostra o resultado dessa setorização em relação ao dia da semana.

Tabela 4. Categorização da precipitação diária.

Dia da Semana	Identificação de final de semana
Segunda	Não
Terça	Não
Quarta	Não
Quinta	Não
Sexta	Não
Sábado	**Sim**
Domingo	**Sim**

Com relação à análise meteorológica, a natureza estocástica dos dados climáticos expressos de forma numérica inviabiliza observar a relação direta dos mesmos com a quantidade total de chamadas. Dessa forma, foi necessário realizar uma categorização desses dados, o que foi feito baseado no site do Alerta Rio [2], a fim de setorizar os valores brutos em informações que podem ser relacionadas com a quantidade de ocorrências.

Inicialmente foi realizada uma setorização dos dados pluviométricos, considerando mm/h ao longo do dia, com relação ao tipo de chuva, conforme mostra a Tabela 5. Para isso, foi considerado o valor máximo de precipitação que ocorreu no dia ao longo de uma hora.

[2] http://alertario.rio.rj.gov.br/

Tabela 5. Categorização da precipitação diária.

Categorização do tipo de chuva	Precipitação máxima (mm/h)
Sem chuva	0,0
Chuva fraca	Maior que 0,0 até 5,0
Chuva moderada	Maior que 5,0 até 25,0
Chuva forte	Maior que 25,0

A precipitação é uma das características que apresenta comportamento bastante variável com valores nulos na maioria dos casos. Por isso, a necessidade de categorização em setores pode trazer uma informação mais entendível para relacionar com à quantidade de chamadas. De modo análogo, a categorização da variação da temperatura em relação ao dia anterior é mostrada na Tabela 6. Para tal, foi utilizado o maior valor da temperatura ocorrido no dia e comparado com o maior valor ocorrido no dia anterior.

Tabela 6. Categorização da temperatura.

Categorização da temperatura	Variação da temperatura (°C)
Estável	Até 2,9
Declínio	Diminuindo mais de 3,0
Elevação	Aumentando mais de 3,0

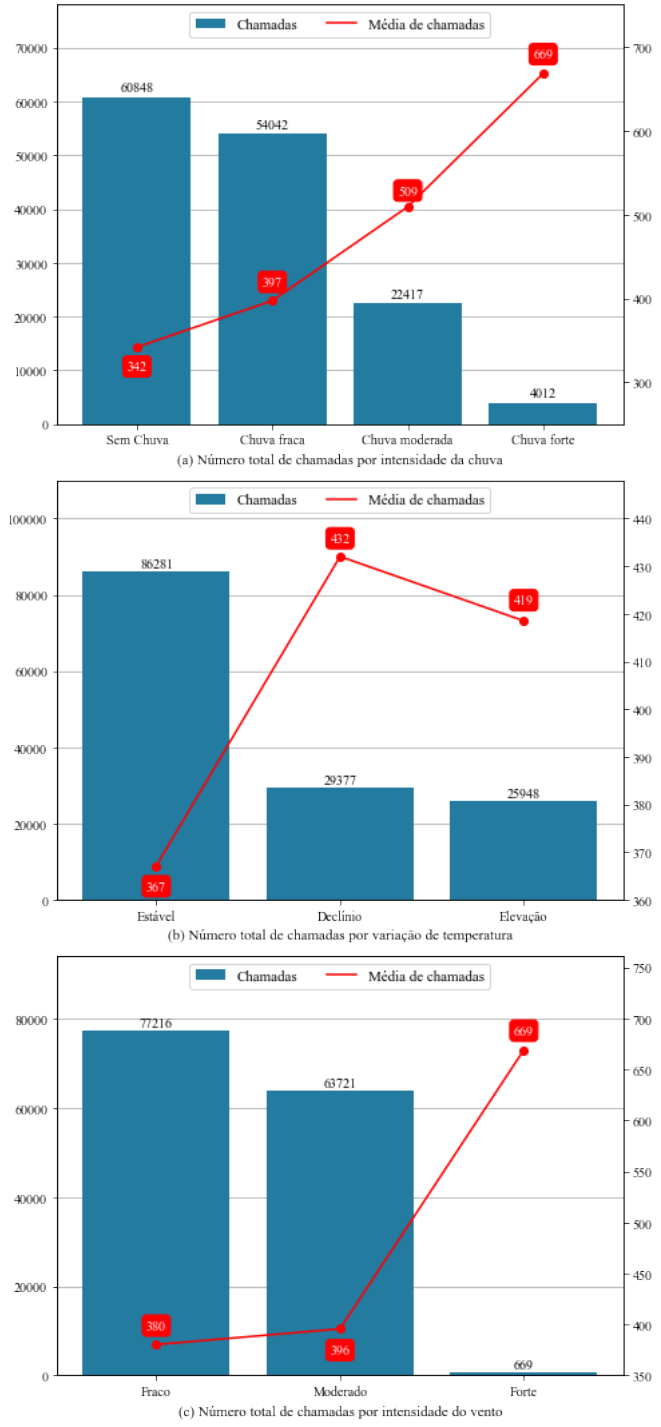
Para tratar a categorização dos dados relacionados ao vento, foi necessário realizar a manipulação da unidade de medida dessa grandeza. A princípio esses valores estavam em m/s e a categorização segundo o Alerta Rio é feita em km/h. Assim, foi criada uma coluna auxiliar com o valor da velocidade do vento em km/h. A setorização dessa característica é mostrada na Tabela 7.

Tabela 7. Categorização do tipo de vento.

Categorização do vento	Velocidade do vento (km/h)
Fraco	Até 18,4
Moderado	Maior que 18,4 até 51,9
Forte	Maior que 51,9

A Figura 7 mostra os resultados obtidos com as categorizações realizadas. A Figura 7(a) ilustra a relação entre a quantidade de chamadas de cada uma das setorizações relacionadas à chuva com as respectivas médias por setor. Percebe-se que há mais dias sem chuva ou com chuva fraca na maior parte da base de dados analisada. Porém, ao observar o aumento da média das chamadas ao longo das categorizações, nota-se que ao passo que há intensificação da chuva na região também há um aumento na quantidade média diária de chamadas. De modo análogo, ao avaliar o comportamento do número de chamadas por setor na Figura 7(b), percebe-se a grande parte dos dados com característica estável. Ocorre um aumento discreto na média das chamadas ao passo que a temperatura sofre uma alteração em comparação com o dia anterior. Por último, a Figura 7(c) mostra a relação da quantidade de chamadas em relação à intensidade do vento na região. Observa-se que os dias com vento fraco ou moderado representam a maioria das ocorrências no banco de dados. Por outro lado, nota-se um aumento na média das chamadas ao passo que essa intensificação da velocidade do vento aumenta. A característica com vento forte apareceu em apenas um dia durante o ano de 2020 e a quantidade de chamadas, que também é a média, foi de 669, maior que a média das outras características.

Figura 7. Quantidade de chamadas e média por categoria.



(a) Número total de chamadas por intensidade da chuva

(b) Número total de chamadas por variação de temperatura

(c) Número total de chamadas por intensidade do vento

4.1 Análise Estatística

A fim de mensurar a correlação entre os dados numéricos com a quantidade de ocorrências na base de chamadas, utilizou-se o Coeficiente de Pearson (ρ), segundo Filho et al. (2009), para realizar o cálculo da correlação linear. De modo geral, foi analisada a intensidade da correlação entre essas características. Assim, quando o valor do módulo do coeficiente ρ está entre o intervalo 0,1 e 0,35 é considerada a intensidade da correlação fraca. Por outro lado, se o valor do módulo de ρ estiver entre o intervalo 0,35 e 0,65 é considerada a intensidade da correlação forte. A Tabela 8 mostra os resultados das características numéricas dos dados climáticos analisados.

Tabela 8. Correlação linear das *features* numéricas com a quantidade de chamadas.

Características	Coeficiente de Pearson (ρ)	Correlação
Precipitação	0,34	Fraca
Velocidade do vento	0,18	Fraca
Temperatura	0,26	Fraca

Percebeu-se que as características numéricas da precipitação, em mm, da velocidade do vento, em m/s, e da temperatura, em graus Celcius, apresentaram correlação de intensidade fraca com a quantidade de chamadas de modo geral. Isso se deve principalmente à causa estocástica da natureza dos dados. Por isso, setorizar essas características auxilia na identificação de comportamentos que os dados expressos de forma bruta não demonstram facilmente.

Dessa forma, foi realizado o teste qui-quadrado, segundo Rana and Singhal (2015), para verificar a associação das *features* categóricas com o número de chamadas. Este trabalho considerou a hipótese de significância de 5%. Assim, o valor da probabilidade de significância (P), resultado do teste, apresenta um limiar em 0,05. Logo, interpreta-se os intervalos de P do seguinte modo:

- **Para $P > 0,05$**: não se rejeita a hipótese nula ou, em outras palavras, aceita-se a hipótese nula e se conclui que não há associação entre as variáveis.
- **Para $P < 0,05$**: há associação entre as variáveis.

A Tabela 9 mostra as respectivas probabilidades de significância encontradas para as características.

Tabela 9. Associação das *features* categóricas com a quantidade de chamadas.

Características	P	Associação
Estação do ano	0,15	Não
Final de semana	0,047	**Sim**
Mês do ano	0,45	Não
Dia da semana	0,083	Não
Intensidade da chuva	0,00033	**Sim**
Variação da temperatura	0,18	Não
Vento	0,0015	**Sim**

Notou-se que algumas características sazonais como a estação do ano, o mês e o dia da semana não apresentaram associação com a quantidade de chamadas. Porém, ao avaliar a coluna de descrição binária proposta com a informação se é final de semana ou não, percebeu-se que houve associação. Isso demonstra a importância dessa manipulação ao longo da análise das características. Além disso, ao verificar as setorizações dos dados meteorológicos, percebeu-se que a categorização da chuva e do vento apresentaram associação com o número de chamadas. Portanto, os resultados mostram que as manipulações decorrentes da análise dos dados auxiliaram na busca pela forma melhor associada à quantidade de chamadas.

5. CONCLUSÃO

Neste trabalho foram apresentadas as análises de dados reais relacionados às chamadas de emergência de clientes de uma concessionária de distribuição de energia elétrica, localizada no Estado do Rio de Janeiro. Foram investigados aspectos temporais e as características climáticas que influenciam nas demandas da rede elétrica e sua correlação com a quantidade de chamadas recebidas ao longo do tempo. Foi necessário levantar hipóteses e buscar tratar a informação de diferentes formas para obter êxito quanto ao foco da aplicação. Verificou-se que a setorização das grandezas climáticas como chuva e vento foram essenciais para obter uma associação com a quantidade de chamadas de emergência em um sistema de distribuição de energia. Do mesmo modo, a setorização binária referente ao comportamento dessas chamadas em relação ao final de semana também atingiu um resultado de associação a partir da aplicação de análise do *Big Data*. Os resultados apresentados contribuem para um melhor entendimento das ocorrências em um sistema de distribuição de energia elétrica e podem auxiliar as equipes de campo que atendem as chamadas de emergência a se programarem e otimizarem a prestação de serviço, a fim de recompor o fornecimento de energia em um menor espaço de tempo e melhorar os índices de qualidade da empresa.

6. AGRADECIMENTOS

Agradecemos a Light S.A. pela disponibilidade dos dados utilizados neste trabalho.

REFERÊNCIAS

Alves, W., Martins, D., Bezerra, U., and Klautau, A. (2017). A hybrid approach for big data outlier detection from electric power scada system. *IEEE Latin America Transactions*, 15, 57–64.

Bhattarai, B., Paudyal, S., Luo, Y., Mohanpurkar, M., Cheung, K., Tonkoski, R., and Zhang, X. (2019). Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions. *Energy informatics*, 2, 141–154.

Brown, R. (2002). Electric power distribution reliability, 2008. *North Carolina, USA*.

Cai, D. (2016). Electric power big data and its applications. *International Conference on Enery, Power and Electrical Engineering (EPEE)*, 181–184.

Cai, Y. and Chow, M. (2009). Exploratory analysis of massive data for distribution fault diagnosis in smart grids. *2009 IEEE Power Energy Society General Meeting*, 1–6.

Codal, K. and İlter, H. (2020). Compression of smart meter big data: A survey. *Renewable and Sustainable Energy Reviews*, 11, 11–26.

Doostan, M. and Chowdhury, B.H. (2017). Power distribution system equipment failure identification using machine learning algorithms. *2017 IEEE Power Energy Society General Meeting*, 1–7.

Douglas, L. (2001). 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6, 1.

Filho, F., Britto, D., and Júnior, J.A.S. (2009). Desvendando os mistérios do coeficiente de correlação de pearson (r). *Revista Política Hoje*, 18, 115–146.

Huang, M., Wei, Z., Sun, G., and Zang, H. (2019). Hybrid state estimation for distribution systems with ami and scada measurements. *IEEE Access*, 7, 120350–120359.

Kraus, R. (2013). Statistical déjà vu: The national data center proposal of 1965 and its descendants. *Journal of Privacy and Confidentiality*, 5.

Li, R., Li, F., and Smith, N.D. (2016). Multi-resolution load profile clustering for smart metering data. *IEEE Transactions on Power Systems*, 31, 4473–4482.

Munshi, A.A. and Yasser, A.R.M. (2017). Big data framework for analytics in smart grids. *Electric Power Systems Research*, 151, 369–380.

Peppanen, J., Reno, M.J., Thakkar, M., Grijalva, S., and Harley, R.G. (2015). Leveraging ami data for distribution system model calibration and situational awareness. *IEEE transactions on smart grid*, 6, 2050–2059.

Rana, R. and Singhal, R. (2015). Chi-square test and its application in hypothesis testing. *ournal of the Practice of Cardiovascular Sciences*, 1, 69.

Rocha, D.V., Sepúlveda, G.P.L., and Flores, H.A.R.. (2016). Mineração de dados aplicada à detecção de fraudes nas redes de distribuição de energia elétrica. *VI Congresso Brasileiro de Engenharia da Produção. Paraná-PR.*

Sahai, S. and Pahwa, A. (2006). A probabilistic approach for animal-caused outages in overhead distribution systems. *2006 International Conference on Probabilistic Methods Applied to Power Systems*, 1–7.

Sridharan, K. and Schulz, N. (2001). Outage management through amr systems using an intelligent data filter. *IEEE Transactions on Power Delivery*, 16, 669–675.

Tu, C., He, X., Shuai, Z., and Jiang, F. (2017). Big data issues in smart grid–a review. *Renewable and Sustainable Energy Reviews*, 79, 1099–1107.

Wen, L., Zhou, K., Yang, S., and Li, L. (2018). Compression of smart meter big data: A survey. *Renewable and Sustainable Energy Reviews*, 91, 59–69.

Xu, L. and Chow, M. (2006). A classification approach for power distribution systems fault cause identification. *IEEE Transactions on Power Systems*, 21, 53–60.

Xu, L., Chow, M., and Taylor, L.S. (2006). Data mining and analysis of tree-caused faults in power distribution systems. *2006 IEEE PES Power Systems Conference and Exposition*, 1221–1227.

Zhou, K., Fu, C., and Yang, S. (2016). Big data driven smart energy management: From big data to big insights. *Renewable and Sustainable Energy Reviews*, 56, 215–225.