Clusterização de Dados de Fibra Ótica em Séries Temporais

1st Caio Bonani Carvalho
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
caio.bonani@unifesp.br

2nd Luiz Fernando de Cristo Moloni
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
luiz.moloni@unifesp.br

***Resumo*—O projeto visa aplicar técnicas de clusterização a dados de alarmes provenientes de CTOs (Caixas de Terminação Óptica) instaladas em postes de telecomunicações. O objetivo é identificar padrões e otimizar a gestão e manutenção dessas infraestruturas, que são críticas para a distribuição de sinais e serviços de telecomunicações. Para isso, foram utilizadas diversas técnicas de análise de dados e algoritmos de clusterização, como K-means, DBSCAN, e Clusterização Hierárquica com Dynamic Time Warping (DTW). A metodologia incluiu a limpeza e preparação dos dados, seguidas pela aplicação das técnicas de clusterização e análise temporal. Os resultados indicaram que a clusterização baseada em coordenadas geográficas foi a mais eficaz, destacando-se o K-means e a Clusterização Hierárquica, enquanto o tempo não teve impacto significativo na clusterização dos alarmes. O estudo conclui que os clusters estão relacionados principalmente à localização geográfica dos alarmes, sugerindo que a otimização da manutenção pode ser alcançada através da análise espacial dos dados.**

I. Introdução e Motivação

Com o avanço da tecnologia e a crescente complexidade das redes de telecomunicações, o monitoramento e a manutenção eficiente das infraestruturas de postes de telecomunicações tornaram-se essenciais. Neste contexto, as caixas de telecomunicações, conhecidas como CTOs (Caixas de Terminação Óptica), desempenham um papel crucial na gestão e distribuição de sinais e serviços. No entanto, a operação contínua dessas caixas está sujeita a uma variedade de problemas que podem afetar negativamente a qualidade dos serviços e a integridade da infraestrutura.

Para gerenciar e mitigar esses problemas, é fundamental entender a frequência e a natureza dos alarmes gerados pelos CTOs. Alarmes são sinais críticos que indicam falhas ou anomalias nas caixas, e a capacidade de interpretar esses sinais pode ajudar na rápida identificação e resolução de problemas. No entanto, a análise de grandes volumes de dados gerados por esses alarmes pode ser desafiadora, exigindo métodos sofisticados para extrair insights valiosos.

A clusterização de dados é uma técnica poderosa que pode ajudar a identificar padrões ocultos e agrupar alarmes semelhantes, permitindo uma análise mais eficiente e direcionada. Através da clusterização, é possível descobrir agrupamentos naturais de problemas, identificar áreas de alta incidência e entender melhor as causas subjacentes dos alarmes. Este conhecimento pode então ser utilizado para otimizar a manutenção, prever falhas e melhorar a resposta a incidentes.

O objetivo deste trabalho é aplicar técnicas de clusterização aos dados de alarmes provenientes de CTOs em postes, visando descobrir padrões e agrupar os alarmes com base em suas características. Utilizaremos métodos avançados de análise de dados e algoritmos de clusterização para processar e interpretar os dados, com a esperança de revelar insights que possam aprimorar a gestão de manutenção e a operação das caixas de telecomunicações.

A motivação para este estudo reside na necessidade crescente de melhorar a eficiência operacional e a capacidade de resposta das equipes de manutenção. A análise detalhada dos dados de alarmes permitirá não apenas uma compreensão mais profunda dos problemas recorrentes, mas também a implementação de estratégias mais eficazes para prevenir e resolver falhas. Com isso, buscamos contribuir para a melhoria da qualidade dos serviços de telecomunicações e a redução dos custos operacionais associados à manutenção de infraestruturas críticas.

II. Trabalhos Relacionados

O uso de técnicas de clusterização em redes de telecomunicações tem sido amplamente estudado para melhorar a eficiência da gestão dessas infraestruturas. Kanagala e Krishnaiah (2016) realizaram um estudo comparativo das abordagens de clusterização K-means, DBSCAN e OPTICS, destacando a eficácia do DBSCAN em detectar clusters em dados espaciais e temporais, especialmente em redes complexas.

Schubert et al. (2017) revisitaram o algoritmo DBSCAN, evidenciando sua robustez em aplicações de telecomunicações, enquanto Boonchoo et al. (2019) propuseram otimizações no DBSCAN para melhorar a indexação e inferência em grandes volumes de dados de redes.

Um estudo relevante é o de Kalisch e Bühlmann (2007), que propôs um algoritmo para estimar grafos acíclicos direcionados em redes de alta dimensionalidade. Esta pesquisa é importante para a detecção de relações causais entre eventos em redes complexas, como as de telecomunicações. Além disso, Lozonavu et al. (2017) exploraram a descoberta de

relações entre alarmes em redes móveis utilizando mineração de padrões sequenciais. Esta técnica ajuda a identificar a sequência de eventos que levam a falhas, facilitando o diagnóstico e a prevenção de problemas futuros.

Outra contribuição significativa é o trabalho de Zhang et al. (2018), que propuseram um algoritmo de mineração de padrões de inundação de alarmes em redes, baseando-se em associações multidimensionais. Esta abordagem permite uma análise mais profunda e detalhada de grandes volumes de dados de alarmes, melhorando a precisão na identificação de falhas.

Esses estudos, junto com outros na área de descoberta de regras de correlação de alarmes, fornecem uma base sólida para a melhoria contínua das técnicas de gestão de falhas em redes de telecomunicações, mostrando a importância de métodos avançados de análise de dados para a manutenção da integridade e eficiência dessas redes.

III. Conceitos Fundamentais

Para compreender a análise e a clusterização dos dados de alarmes de CTOs em postes, é essencial familiarizar-se com alguns conceitos fundamentais que sustentam a abordagem e a interpretação dos resultados. Esta seção aborda os principais conceitos relacionados à análise de dados, técnicas de clusterização e à natureza dos alarmes em CTOs.

1. Dados de Alarmes e CTOs

Os alarmes provenientes de CTOs são sinais críticos que indicam problemas ou falhas nas caixas de terminação óptica instaladas em postes de telecomunicações. Estes alarmes podem variar em tipo e gravidade, refletindo diferentes tipos de problemas, como falhas de conexão, interferências ou falhas de equipamento. A coleta e o monitoramento contínuos desses alarmes são essenciais para garantir a operação estável e eficiente das redes de telecomunicações.

2. Análise de Dados

A análise de dados refere-se ao processo de inspeção, limpeza e modelagem de dados com o objetivo de descobrir informações úteis e tomar decisões informadas. Este processo pode envolver várias etapas, incluindo a preparação dos dados, a aplicação de técnicas analíticas e a interpretação dos resultados. No contexto dos alarmes de CTOs, a análise de dados permite identificar padrões, tendências e anomalias, fornecendo insights valiosos sobre a frequência e a natureza dos problemas.

3. Clusterização

A clusterização é uma técnica de aprendizado de máquina não supervisionado que visa agrupar dados semelhantes em clusters ou grupos. O objetivo é encontrar estruturas naturais dentro dos dados, onde objetos no mesmo cluster são mais semelhantes entre si do que com os objetos em outros clusters. A clusterização é particularmente útil para explorar dados sem rótulos e identificar padrões ocultos. Em nosso estudo, aplicaremos a clusterização para agrupar alarmes com características semelhantes, facilitando a análise e a identificação de padrões de problemas.

4. Algoritmos de Clusterização

Existem vários algoritmos de clusterização, cada um com suas próprias características e aplicações. Alguns dos algoritmos mais comuns incluem:

- **K-means:** Um dos algoritmos de clusterização mais populares, que divide os dados em um número predefinido de clusters (K) com base na minimização da soma das distâncias dentro dos clusters.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Um algoritmo que identifica clusters de alta densidade e lida bem com dados ruidosos e de forma arbitrária.
- **Hierarchical Clustering:** Um método que constrói uma hierarquia de clusters, permitindo visualizar a estrutura dos dados em diferentes níveis de granularidade.

5. Análise de Séries Temporais

Como os alarmes de CTOs são registrados ao longo do tempo, a análise de séries temporais é crucial para compreender como a frequência e a gravidade dos alarmes variam ao longo do tempo. Séries temporais são conjuntos de dados coletados em intervalos regulares ao longo do tempo e podem revelar padrões sazonais, tendências e ciclos. Técnicas de análise de séries temporais podem ser aplicadas para entender melhor a dinâmica dos alarmes e identificar padrões temporais significativos.

6. Normalização dos Dados

A normalização dos dados é o processo de ajustar a escala dos dados para que diferentes variáveis possam ser comparadas de forma justa. Em nosso estudo, normalizaremos as coordenadas dos alarmes para garantir que as variáveis estejam na mesma escala, facilitando a aplicação de algoritmos de clusterização e a interpretação dos resultados.

7. Dynamic Time Warping

O Dynamic Time Warping (DTW) é uma técnica de análise que mede a similaridade entre duas séries temporais que podem variar em velocidade ou tempo. Ao contrário de métodos

tradicionais que exigem que as séries temporais tenham o mesmo comprimento ou estejam alinhadas temporalmente, o DTW permite que as séries sejam ”deformadas”para encontrar o melhor alinhamento entre elas. Essa característica torna o DTW particularmente útil para comparar padrões de alarmes de CTOs que podem ocorrer em diferentes intervalos de tempo ou em momentos de operação diferentes.

No contexto deste estudo, o DTW pode ser aplicado para identificar semelhanças em padrões temporais de alarmes em diferentes CTOs, mesmo quando esses padrões não estão perfeitamente alinhados no tempo. Isso pode ser crucial para detectar problemas recorrentes ou anomalias em momentos distintos de operação das redes de telecomunicações.

8. Visualização de Dados

A visualização de dados é uma parte crucial da análise de dados, permitindo que os padrões, tendências e anomalias identificados nos dados sejam apresentados de maneira clara e compreensível. No contexto dos alarmes de CTOs, a visualização de dados pode ajudar a identificar rapidamente áreas problemáticas e entender a distribuição espacial e temporal dos alarmes. Gráficos como heatmaps, gráficos de dispersão e diagramas de cluster podem ser particularmente úteis para ilustrar os resultados da clusterização e da análise de séries temporais.

IV. Objetivos

O principal objetivo deste trabalho é realizar a clusterização de dados sobre alarmes provenientes de CTOs em postes, com o intuito de identificar padrões e otimizar a gestão e manutenção das infraestruturas de telecomunicações. Para atingir esse objetivo, o estudo visa, primeiramente, limpar e preparar os dados de alarmes. Isso inclui a remoção de valores ausentes, a normalização das coordenadas e a transformação dos dados em um formato adequado para a clusterização. A qualidade dos dados é crucial para garantir que os algoritmos de clusterização funcionem corretamente e produzam resultados precisos.

Uma vez que os dados estejam preparados, o estudo aplicará diferentes técnicas de clusterização, como K-means e DBSCAN, para identificar agrupamentos naturais nos dados de alarmes. Esta abordagem permitirá descobrir padrões e agrupar alarmes com características semelhantes, facilitando a análise e interpretação dos dados.

Além disso, o trabalho se concentrará na análise e visualização dos resultados da clusterização. Serão criadas visualizações, como mapas e gráficos, para representar a distribuição dos clusters e os tipos de alarmes associados a cada grupo. Essas visualizações ajudarão a interpretar os clusters e a entender melhor a distribuição espacial e temporal dos alarmes, proporcionando insights valiosos para a gestão e manutenção das CTOs.

Outro aspecto importante do estudo é a análise da dinâmica temporal dos alarmes. Será realizada uma análise de séries temporais para investigar como a frequência e a gravidade dos alarmes variam ao longo do tempo, identificando padrões sazonais, tendências e ciclos. Compreender a dinâmica temporal pode ajudar a prever falhas e otimizar a programação de manutenção, melhorando a eficiência operacional.

A validação e comparação dos resultados também serão realizadas para garantir a robustez e a confiabilidade das análises. Serão utilizadas métricas de avaliação para medir a qualidade dos clusters e a precisão das análises, assegurando que as conclusões do estudo sejam baseadas em dados precisos e confiáveis.

Finalmente, com base nos resultados obtidos, serão propostas recomendações para melhorar a gestão e a manutenção das CTOs. As recomendações visam aplicar os insights obtidos para aprimorar a gestão prática das infraestruturas, reduzindo custos e melhorando a qualidade dos serviços de telecomunicações.

V. Metodologia

A. Base de Dados

Neste projeto foi utilizada a base de alarmes da empresa FiBrasil, dados estes que são gerados pelo *software* Alarm Manager. Os dados incluem informações diversas, entre elas:

- problema: descrição do problema que gerou o alarme;
- ts_registro: data, no formato UTC, de quando o alarme foi gerado;
- netwin_x: longitude de onde a CTO está localizada, informação que é retirada do inventário – Netwin;
- netwin_y: latitude da CTO;
- NETWIN_LOCALITY: cidade onde o alarme foi gerado.

Estes dados estão compilados em uma tabela SQL, que se encontra hospedada no serviço BigQuery do Google Cloud Plataform.

Para o acessá-los, além da forma tradicional via interface do BigQuery, foi criado um *script* python para realizar uma *query* utilizando o pacote ***google-cloud-bigquery***, do PiP.

IA.py

```
client=bigquery.Client.from_service_account
    ↪_json('fibrasil.json')

query =
"""
    SELECT
      subsystem,
      node_alias,
      local_code,
      problema,
      ts_registro,
      netwin_x,
      netwin_y,
      NETWIN_ADDRESS,
      NETWIN_LOCALITY
    FROM
       `fibrasil.vw_zone.vw_alarmes`
    WHERE
```

```
        DATE(DATETIME(ts_registro))<='2024-08-16'
"""

df = client.query(query).to_dataframe()
df.to_csv('alarmes.csv')
```

Nomes do arquivo .json e do campo FROM foram fantasiados, por questões de privacidade e segurança.

B. Análise de Dados

A primeira busca realizada no DataFrame foi quais as cidades com maior geração de alarmes.

IA.py

```
top_cities=df1['NETWIN_LOCALITY']
                .value_counts()
                .head(10)

print(top_cities)
```

Com isso foi descoberto que as 10 maiores cidades geradoras de alarmes são:

1) MANAUS
2) BELÉM
3) SAO LUÍS
4) SANTA LUZIA
5) PARAGOMINAS
6) VOLTA REDONDA
7) UBERLÂNDIA
8) SETE LAGOAS
9) GUARAPARI
10) CACHOEIRO DE ITAPEMIRIM

Também foi plotado um gráfico para identificar quais os maiores problemas que ocorrem na base.

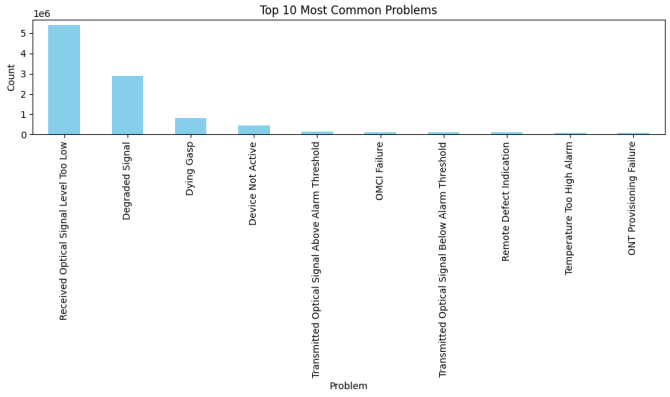

Figura 1. 10 maiores problemas

Após esta etapa foi decidido analisar a quantidade e os problemas dos alarmes em cada uma das dez cidades.

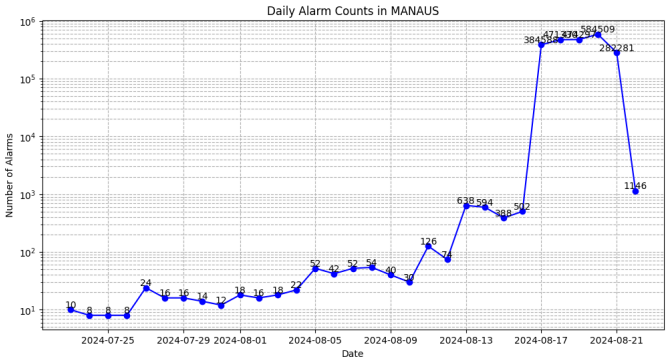

Figura 2. contagem diária de alarmes em Manaus

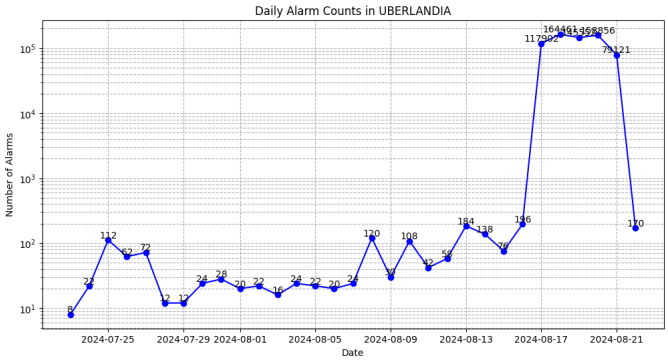

Figura 3. contagem diária de alarmes em Uberlândia

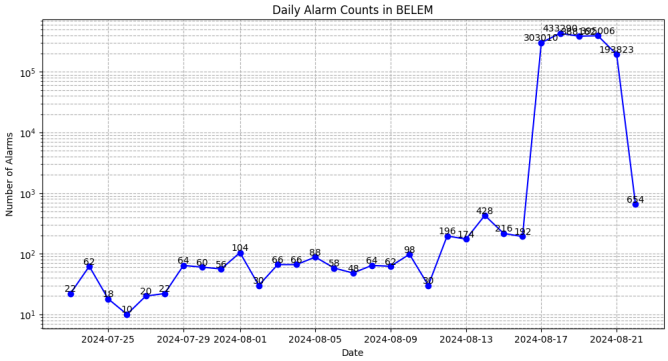

Figura 4. contagem diária de alarmes em Belém

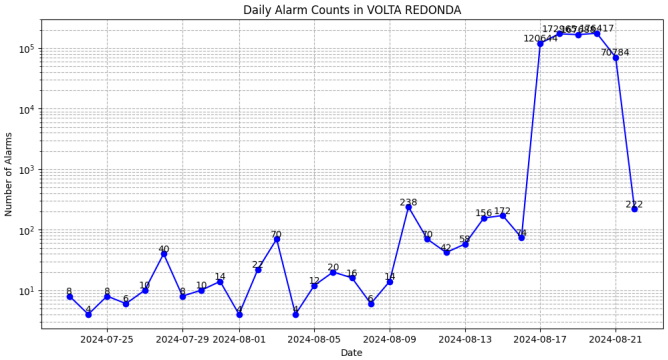

Figura 5. contagem diária de alarmes em São Luís

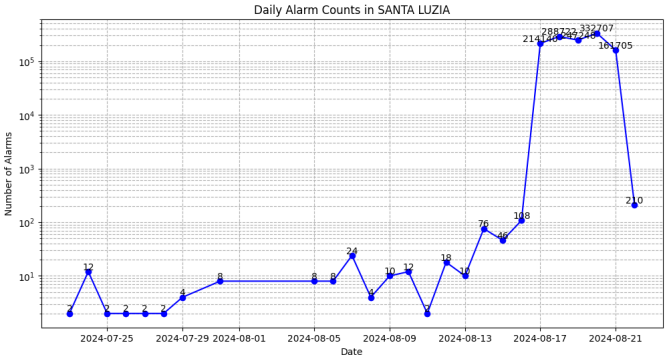

Figura 6. contagem diária de alarmes em Santa Luzia

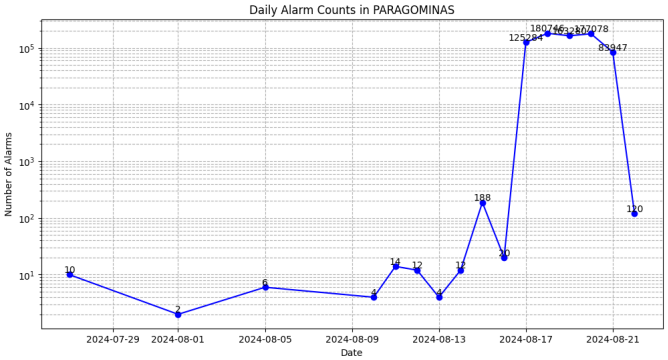

Figura 7. contagem diária de alarmes em Paragominas

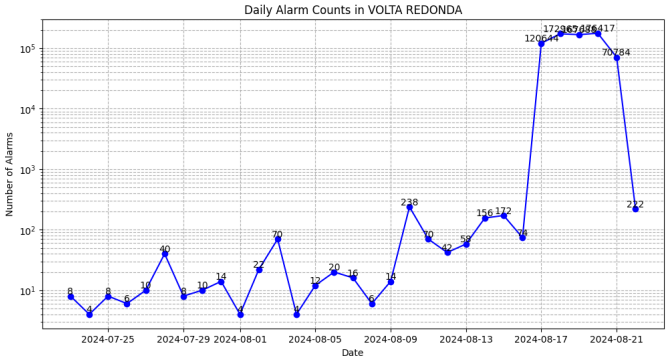

Figura 8. contagem diária de alarmes em Volta Redonda

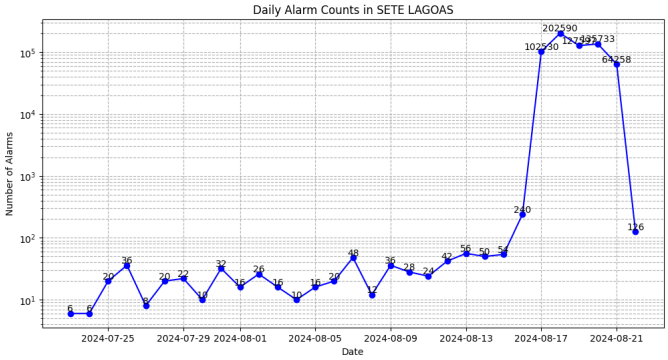

Figura 9. contagem diária de alarmes em Sete Lagoas

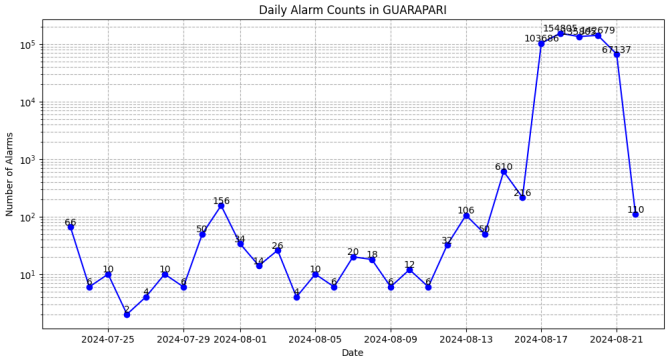

Figura 10. contagem diária de alarmes em Guarapari

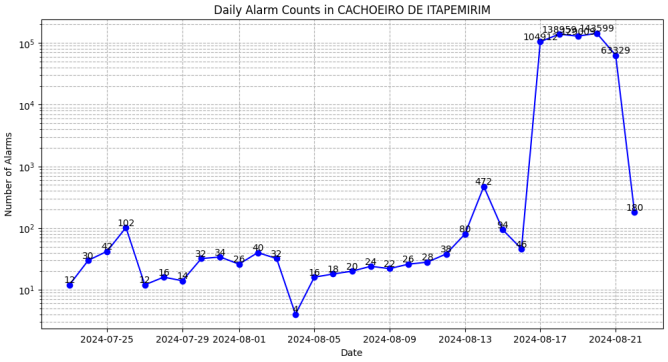

Figura 11. contagem diária de alarmes em Itapemirim

Como pode ser observado, todas cidades tiveram o mesmo comportamento – explosão nos problemas do dia 17 de Agosto de 2024 em diante. Não somente isso, mas o problema que as CTOs alertavam também foram iguais, como visto nos gráficos:

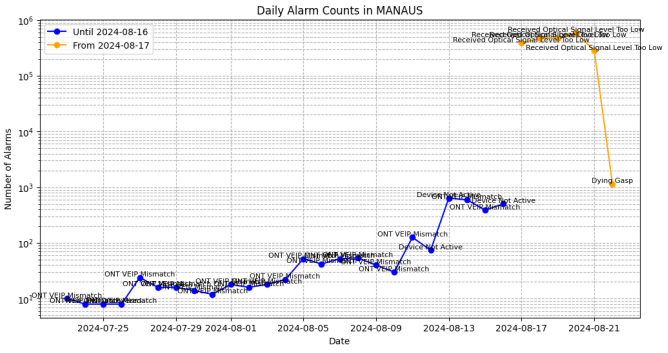

Figura 12. problemas - Manaus

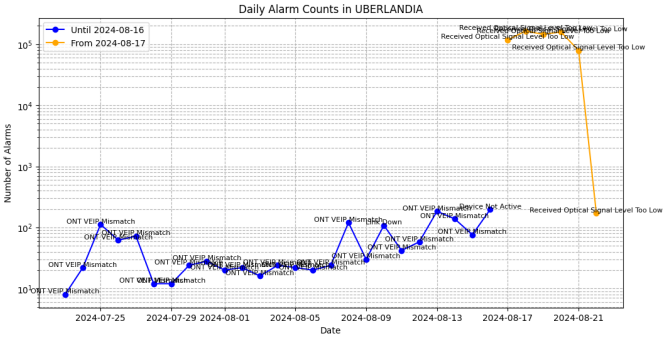

Figura 13. Problemas - Uberlândia

Os gráficos para as outras cidades seguem o mesmo padrão.

Não foi possível identificar a causa para este salto em alarmes, em todas cidades e exatamente no mesmo dia. Algumas das suspeitas incluem a inclusão de um novo *software* que também está incluindo dados de alarmes extras na base e problemas relacionados ao Alarm Manager, onde este indica alarmes errados.

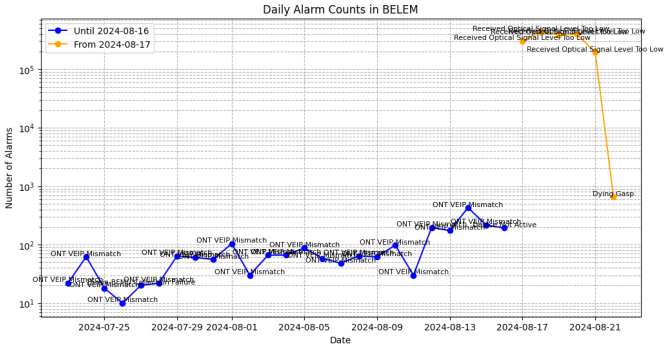

Figura 14. problemas - Belém

Independentemente, a organização dos dados atuais são prejudiciais à análise, pois o aumento está na casa de duas ordens de grandeza(10^3 para 10^5). Dessa forma, foi decidido descartar os dados do dia 17/08/2024 em diante.

IA.py

```
df['ts_registro']=pd.to_datetime(
                          df['ts_registro'])

start_date='2024-07-16'
end_date='2024-08-16'

df=df[(df['ts_registro'] >= start_date) &
    ↪ (df['ts_registro'] <= end_date)]

df=df.dropna()

df.set_index('ts_registro', inplace=True)
```

Um detalhe observado na base foi a presença de dados anteriores à 01/07/2024, inclusive alguns dados de 2022 e 2023, algo não esperado, pois a base deveria compreender apenas o período do mês anterior. Por isso, além de uma variável do final do período, foi adicionada uma variável de início. E, finalmente, foi feita a limpeza de células vazias (*Null* ou NaN – Not a Number) e *"setado"* a coluna de data – ts_registro – como o *index*.

Foram feitas análises desse novo DataFrame transformado:

- gráfico de dispersão;
- análise temporal;
- contagem dos tipos de problemas.

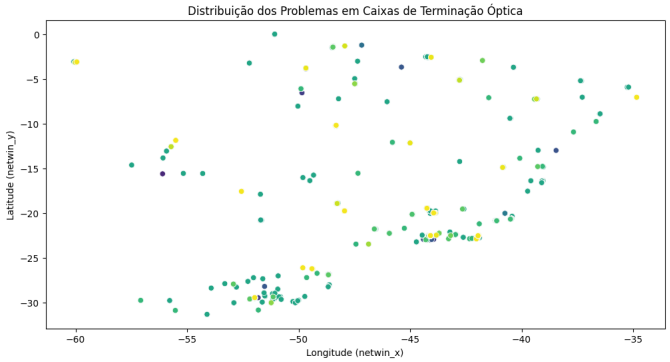

Figura 15. Gráfico de Dispersão

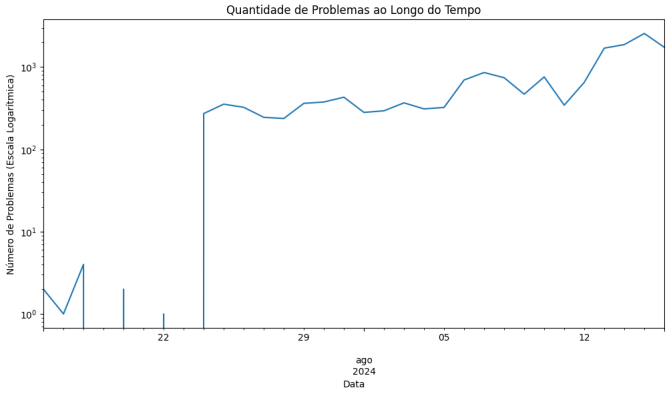

Figura 16. Análise Temporal - em escala Logarítmica

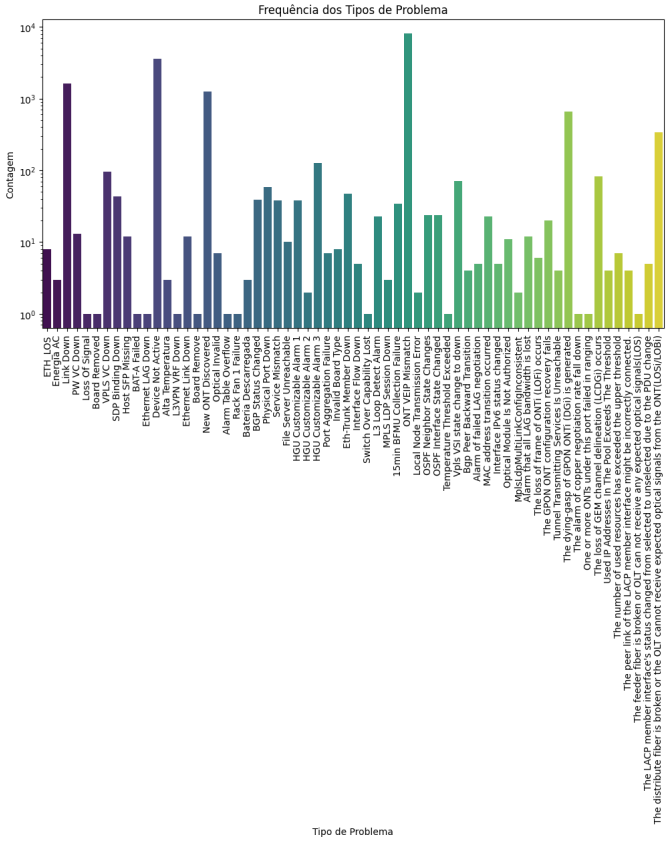

Figura 17. Frequência dos Problemas - em escala Logarítmica

C. Clusterização

Para realizar a Clusterização dos Dados foram tentadas algumas Técnicas diferentes, como:

- K-Means baseado somente nas coordenadas;
- DBSCAN baseado somente nas coordenadas;
- Hierarchical Clustering com Dynamic Time Warping (DTW);
- DBSCAN com DTW;
- K-Means com DTW.

Nos casos de clusterização sem o DTW, foram passados uma tupla como o X recebido pelo modelo:

IA.py

```
coords=df[['netwin_x', 'netwin_y']]

kmeans=KMeans(n_clusters=3,random_state=42)

df['k_cluster']=kmeans.fit_predict(coords)
```

Já nos casos de clusterização de Séries Temporais, utilizando o DTW, a abordagem foi diferente. Foi necessário calcular a matriz de distância, utilizando a ***dtaidistance***.

IA.py

```
from dtaidistance import dtw

dist_matrix=dtw.distance_matrix_fast(coords.values)
```

VI. Resultados

Após rodado todos os diferentes algoritmos de clusterização citados foi obtido as seguintes visualizações:

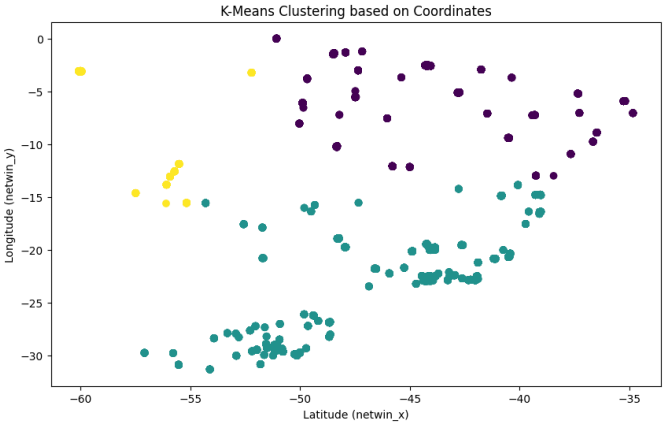

Figura 18. K-Means Espacial

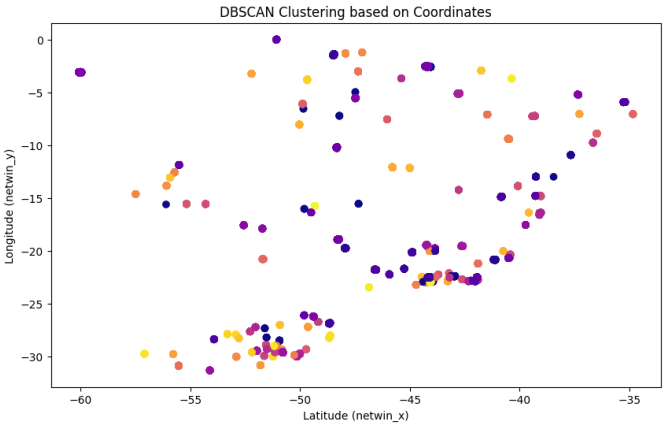

Figura 19. DBSCAN Espacial

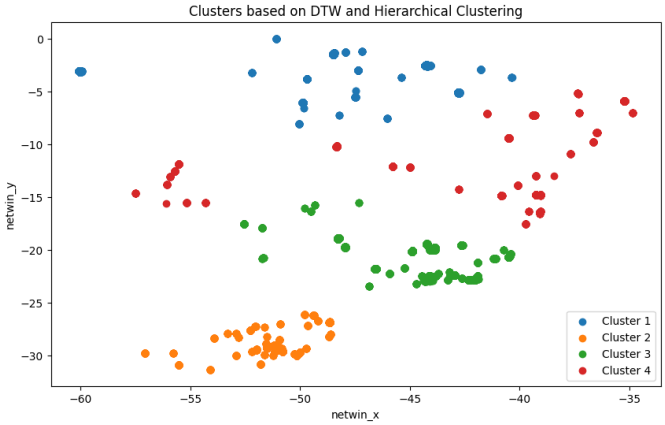

Figura 20. Hierarchical Clustering com DTW

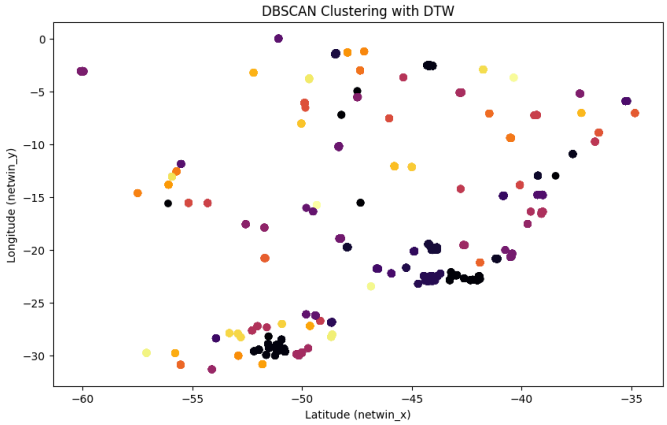

Figura 21. DBSCAN com DTW

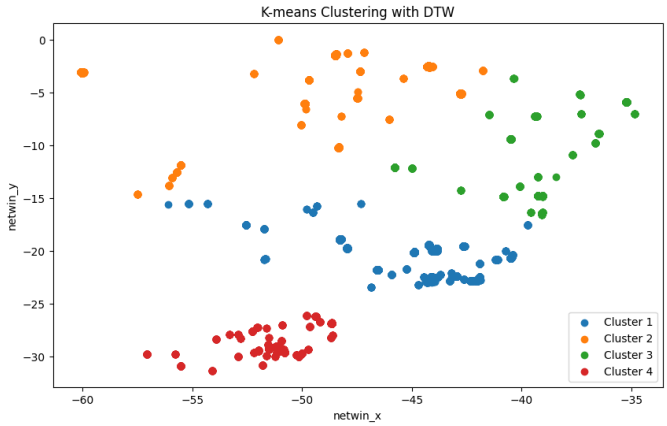

Figura 22. K-Means com DTW

É observado que o algoritmo Density-Based Spatial Clustering of Applications with Noise – DBSCAN – não se saiu muito bem na clusterização dos dados, tanto em sua versão padrão(espacial) quanto utilizando DTW. Enquanto, o K-Means (espacial e DTW) e o Hierarchical Clustering tiveram resultados satisfatórios e parecidos, onde conseguiram clusterizar todos os alarmes em regiões. Porém, é de se notar que o fator tempo não interferiu na clusterização. A versão do K-means com DTW produziu um cluster a mais, sendo alarmes provenientes da região Sul do Brasil, comportamento esperado.

Dessa forma, podemos dizer que o fator tempo não tem ligação alguma com os alarmes, muito menos a região junto do problema. Não foram identificados problemas mais presentes em regiões ou estados, assim a clusterização só foi possível se dar de forma espacial, ou seja, todos os clusters estão relacionados com as posições geográficas de seus pontos, agrupando-os por proximidade, como pode ser observado pelo mapa gerado utilizando a biblioteca ***folium***.

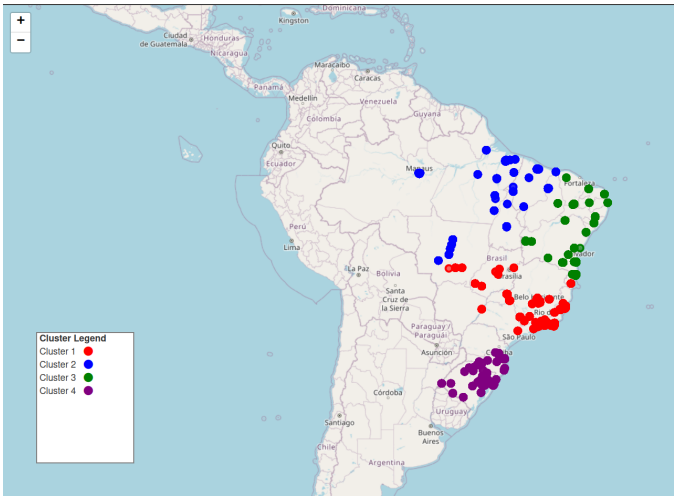

Figura 23. Mapa gerado pelo folium. Alarmes Clusterizados.

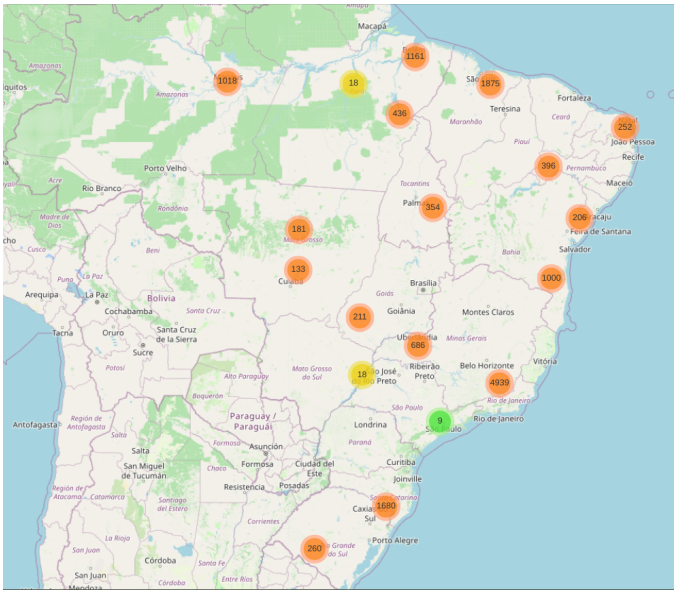

Figura 24. Mapa gerado pelo folium. Quantidade de Alarmes x Clusters - 1

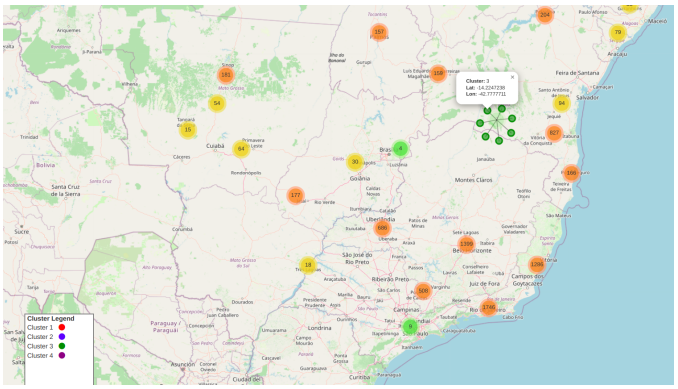

Figura 25. Mapa gerado pelo folium. Quantidade de Alarmes x Clusters - 1

REFERÊNCIAS

[1] H.K. Kanagala, and V.J.R. Krishnaiah, "A comparative study of K-means, DBSCAN and OPTICS,"*2016 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2016, pp. 1-6. http://dx.doi.org/10.1109/ICCCI.2016.7479923

[2] E. Schubert, J. Sander, M. Ester, H.P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN,"*ACM Transactions on Database Systems*, vol. 42, no. 3, pp. 1-21, 2017. http://dx.doi.org/10.1145/3068335

[3] T. Boonchoo, X. Ao, Y. Liu, W. Zhao, F. Zhuang, and Q. He, "Grid-based DBSCAN: Indexing and inference,"*Pattern Recognition*, vol. 90, pp. 271-284, 2019. http://dx.doi.org/10.1016/j.patcog.2019.01.034

[4] Kalisch, M., Bühlmann, P. (2007). "Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm."*Journal of Machine Learning Research*, 8, 613-636. http://www.jmlr.org/papers/volume8/kalisch07a/kalisch07a.pdf

[5] Lozonavu, P., Olteanu, M., Blaga, P. (2017). "Sequential Pattern Mining for Alarm Correlation in Mobile Networks."*2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pp. 233-238. http://dx.doi.org/10.1109/SAMI.2017.7880308

[6] Zhang, J., Wu, Y., Zhang, T. (2018). "Multidimensional Association Rule Mining for Alarm Floods in Telecommunication Networks."*Journal of Computer Networks and Communications*, 2018. http://dx.doi.org/10.1155/2018/7360248