IBM Report for Clustering Project

Code on GitHub

https://github.com/luisresende13/ibm-machine-learning-specialization/blob/main/Clustering/Otimiza%C3%A7%C3%A3o%20DBSCAN%20por%20Densidade.ipynb

1. **Main Objective of the Analysis**

The main goal of this analysis is to apply unsupervised learning, specifically clustering, to identify patterns within the dataset's flood related events. By focusing on event coordinates, the objective is to group similar events together based on their spatial distribution. This clustering approach aims to find spatial clusters of flood events and expand our understanding of the different types of flood events, including Water Blade, Water Bag, and Flood.

Understanding the spatial distribution of various flood event types can contribute to more effective resource allocation, risk mitigation strategies, and urban planning decisions to minimize the impact of such events on communities.

2. **Brief Description of the Data Set and Summary of Its Attributes**

The dataset originates from the Operations and Control Center in Rio de Janeiro, and contains information about city events. Each row in the dataset represents a specific city event with attributes such as event's type, location, timestamp, description, severity. The columns in the dataset include:

tipo (Type): Denotes the type or category of the event.
pop_id: Event type identification code.
latitude: The geographical latitude coordinates of the event location.
inicio (Start): The timestamp indicating when the event started.
titulo (Title): Descriptive title of the event.
fim (End): The timestamp indicating when the event ended.
aviso_id: Identifier associated with any warnings related to the event.
descricao (Description): Detailed description of the event.
informe_id: Identifier associated with any informational aspect of the event.
gravidade (Gravity): Indicates the severity or gravity level of the event.
id: Another identifier associated with the event.
longitude: The geographical longitude coordinates of the event location.
status: The status of the event.
bairro (Neighborhood): The neighborhood in which the event occurred.
prazo (Deadline): Deadline or time frame associated with the event.
pop_titulo: Title associated with the event type identifier.

The geographical coordinates (latitude and longitude) serve as the focal point for this analysis, aiming to discern patterns in the spatial distribution of flood events across the dataset.

3. **Brief Summary of Data Exploration and Actions Taken for Data Cleaning and Feature Engineering**

The focus of exploratory data analysis (EDA) was primarily to understand the structure and characteristics of the dataset. The EDA involved examining the distribution of flood event types, checking for missing values, and gaining insights into the temporal and spatial aspects of the events.

Visualization techniques: Scatter plots and histograms, were produced in order to understand the distribution of event occurrences for different geographical coordinates and time intervals.

Handling Missing Values: The cleaning process involved identifying and addressing missing or inconsistent values in the dataset.

Handling Missing Coordinates: Events with missing coordinates were excluded from the dataset before fitting the unsupervised model.

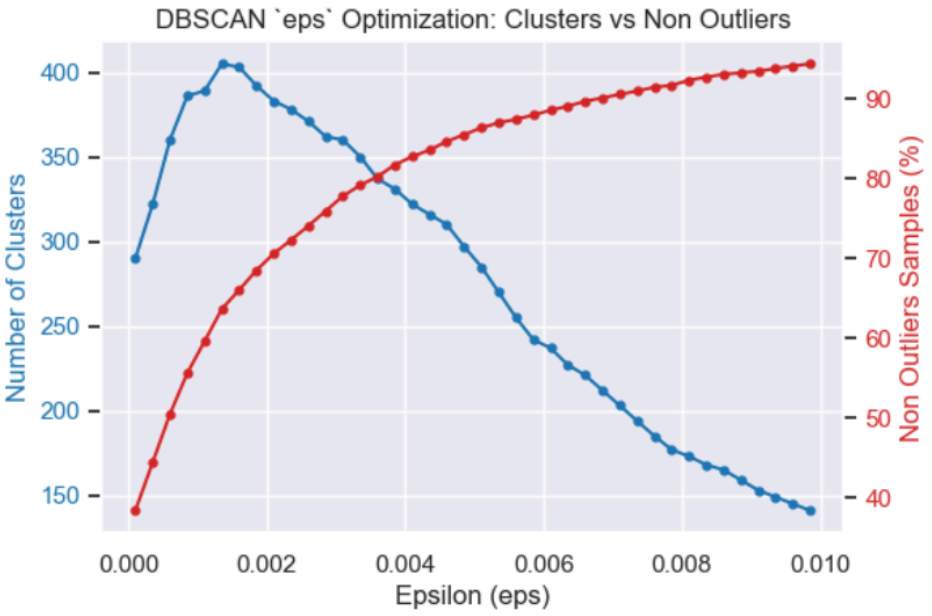
Outliers: Geographical outliers (with coordinates not belonging to the city area) were identified and removed.

In summary, the data cleaning stage was relatively straightforward given the high quality of the dataset and the simplicity of the analysis, which relied solely on the availability and correctness of the event's coordinates.

4. **DBSCAN Algorithm Hyperparameter Tuning**

The unsupervised learning algorithm DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was used. DBSCAN is well-suited for our spatial clustering task, particularly in cases where clusters show varying shapes and densities. An optimization procedure was done to determine the optimal value for the epsilon parameter (eps), a critical parameter for the clustering outcome.

The optimization used an exhaustive exploration of epsilon values ranging from 0.0001 to 0.01 with an increment of 0.00025. For each epsilon value, DBSCAN was applied, and the resulting clusters were evaluated based on various metrics. The goal was to find a balance between minimizing the number of identified clusters and ensuring a high percentage of included samples in those clusters. The epsilon value with the best clustering outcome was then selected for further analysis.



5. **Summary of Training Three Different Cluster Models**

Three distinct cluster models were trained using epsilon values within the optimal range (0.0015 ~ 0.0040) determined using optimization. The training was done using the DBSCAN algorithm. These models were created with epsilon values of 0.0015, 0.0020, and 0.0025, respectively.

Number of events of each type used in training:

- Bolsão d'água em via (Water bag): 4397
- Alagamento (Flood): 314
- Lâmina d'água (Water blade): 211
- Alagamentos e enchentes (Flood and overflow): 151
- Enchente (Overflow): 5
- Total: 5078

Model with Eps=0.0015:

- Total Clusters: 404
- Included Samples: 3300
- Included Samples Percentage: 64.99%
- Average Samples per Cluster: 8.17
- Average Area per Cluster: 0.016686 km²
- Average Density per Cluster: 1679.993809 events/km²
- Median Samples per Cluster: 5.0
- Median Area per Cluster: 0.006880 km²
- Median Density per Cluster: 768.444438 events/km²

Model with Eps=0.0020:

- Total Clusters: 388
- Included Samples: 3541
- Included Samples Percentage: 69.73%
- Average Samples per Cluster: 9.13
- Average Area per Cluster: 0.034760 km²
- Average Density per Cluster: 1295.934439 events/km²
- Median Samples per Cluster: 5.0
- Median Area per Cluster: 0.011690 km²
- Median Density per Cluster: 464.270838 events/km²

Model with Eps=0.0025:

- Total Clusters: 376
- Included Samples: 3733
- Included Samples Percentage: 73.51%
- Average Samples per Cluster: 9.93
- Average Area per Cluster: 0.057057 km²
- Average Density per Cluster: 912.864584 events/km²
- Median Samples per Cluster: 5.0
- Median Area per Cluster: 0.019856 km²
- Median Density per Cluster: 282.429572 events/km²

These metrics show the trade-off between the granularity of clusters and the percentage of included samples. The higher epsilon values result in fewer clusters with a higher percentage of included samples.

6. Spatial Visualization of Flood Clusters

6.1 Specific Locations at Different Epsilon Values

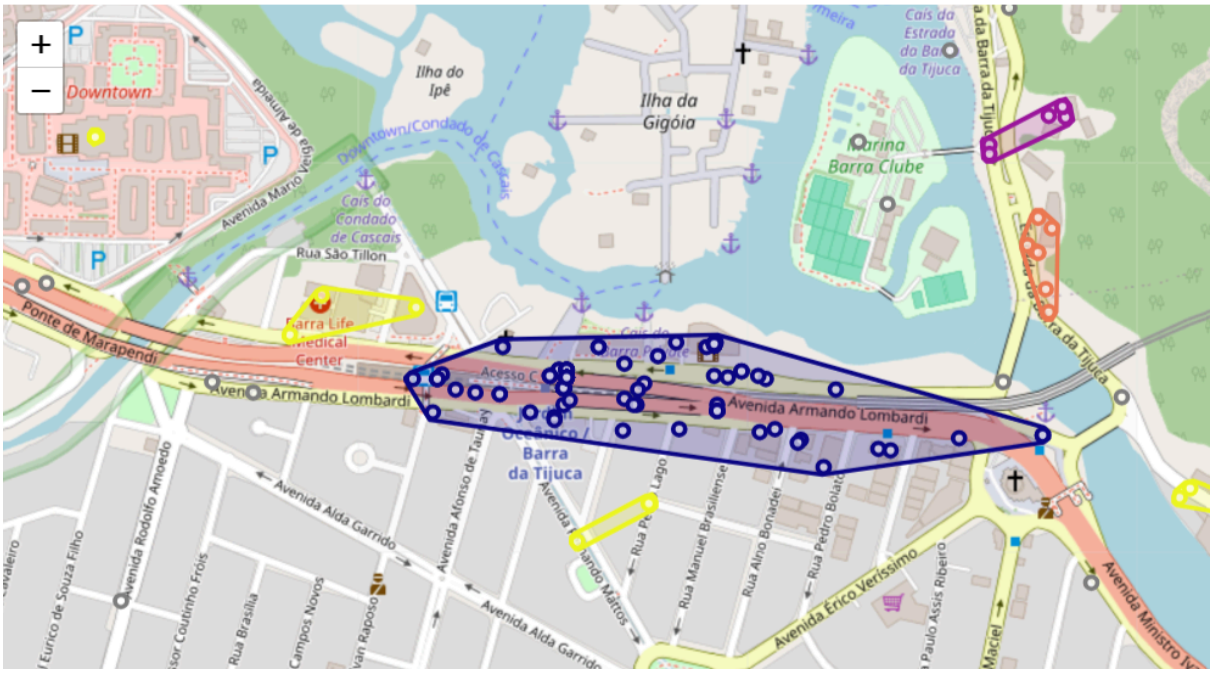
Interactive maps were created on three specific locations, “Avenida Armando Lombardi”, “Rua do Catete”, and “Lagoa”, to illustrate the optimization process. For each location, maps were created for the three epsilon values: 0.0015, 0.0020, and 0.0025.

The maps are interactive, allowing users to zoom in and click on individual clusters to access their information. The popups for each cluster present the cluster label, rank, number of samples, area, and density.

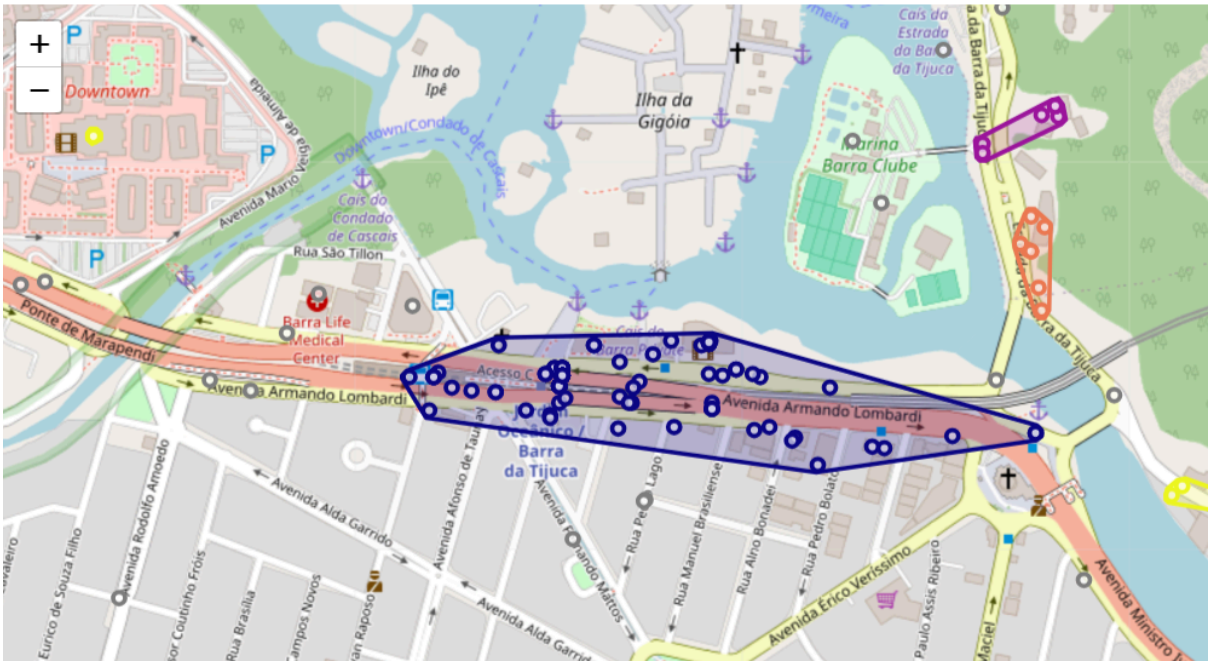
Each map is identified by a unique map title, containing the DBSCAN epsilon value, total clusters, percentage of included samples, and the location.

The colormap used in the maps helps to differentiate clusters based on the number of samples.

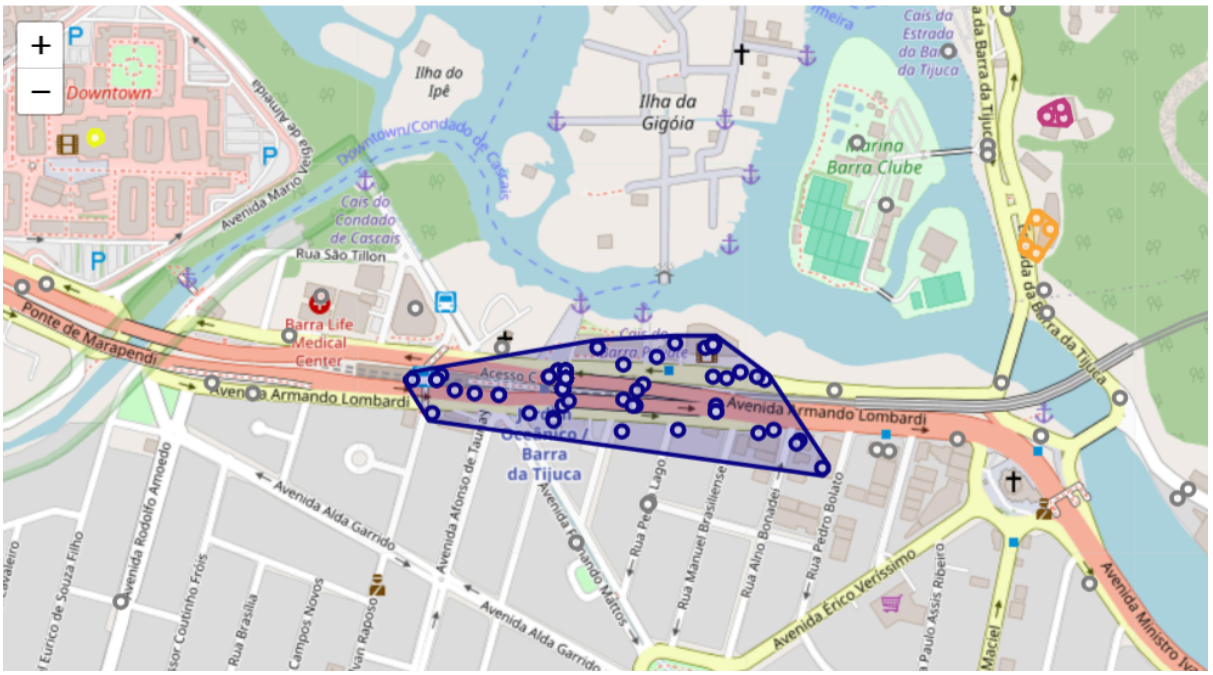
DBSCAN | EPS: 0.0025 | CLUSTERS: 376 | SAMPLES: 73.51 % | LOCATION: Avenida Armando Lombardi



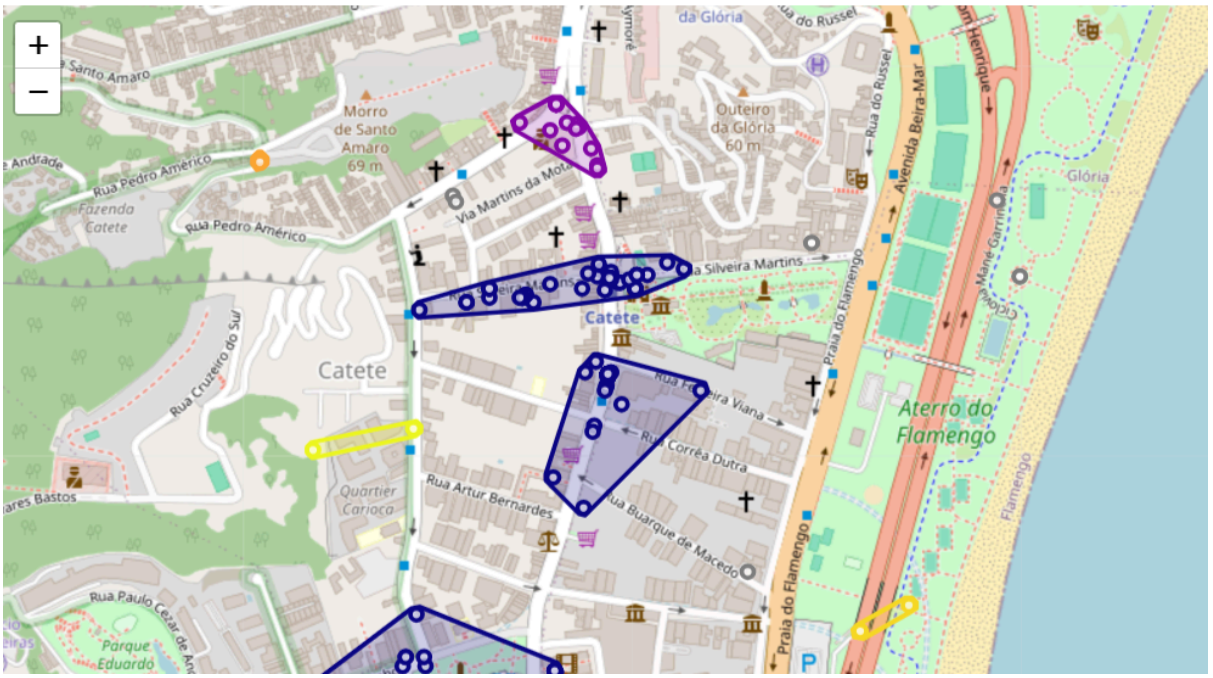
DBSCAN | EPS: 0.002 | CLUSTERS: 388 | SAMPLES: 69.73 % | LOCATION: Avenida Armando Lombardi



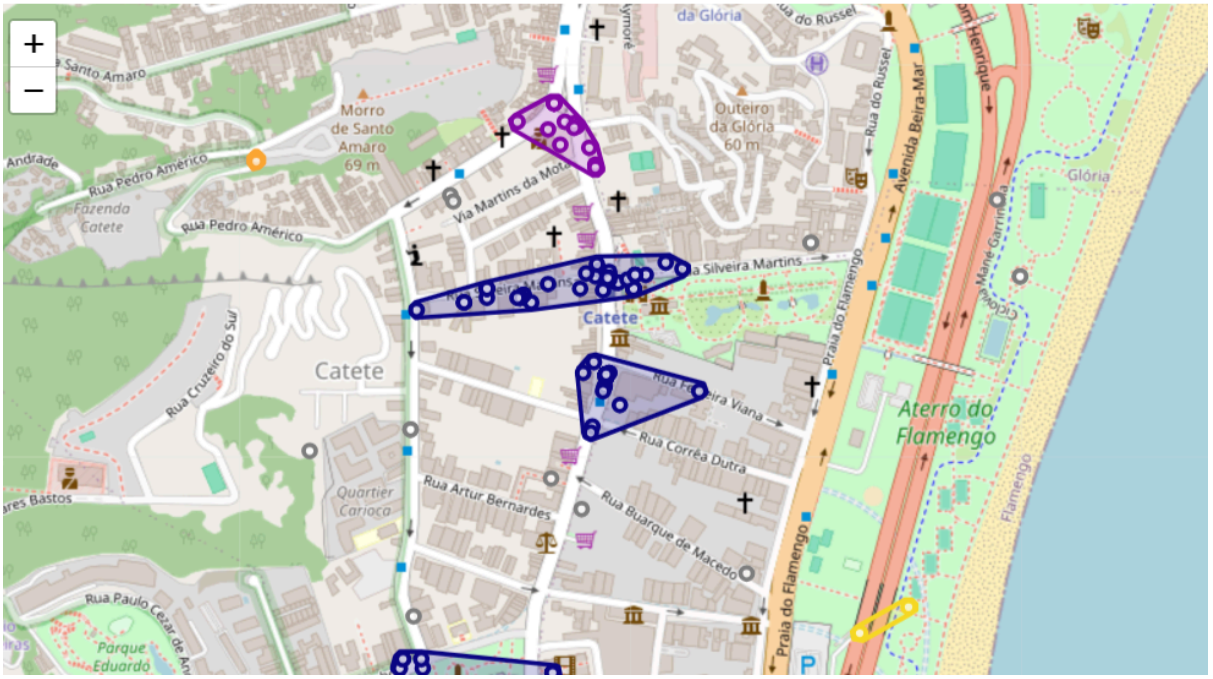
DBSCAN | EPS: 0.0015 | CLUSTERS: 404 | SAMPLES: 64.99 % | LOCATION: Avenida Armando Lombardi



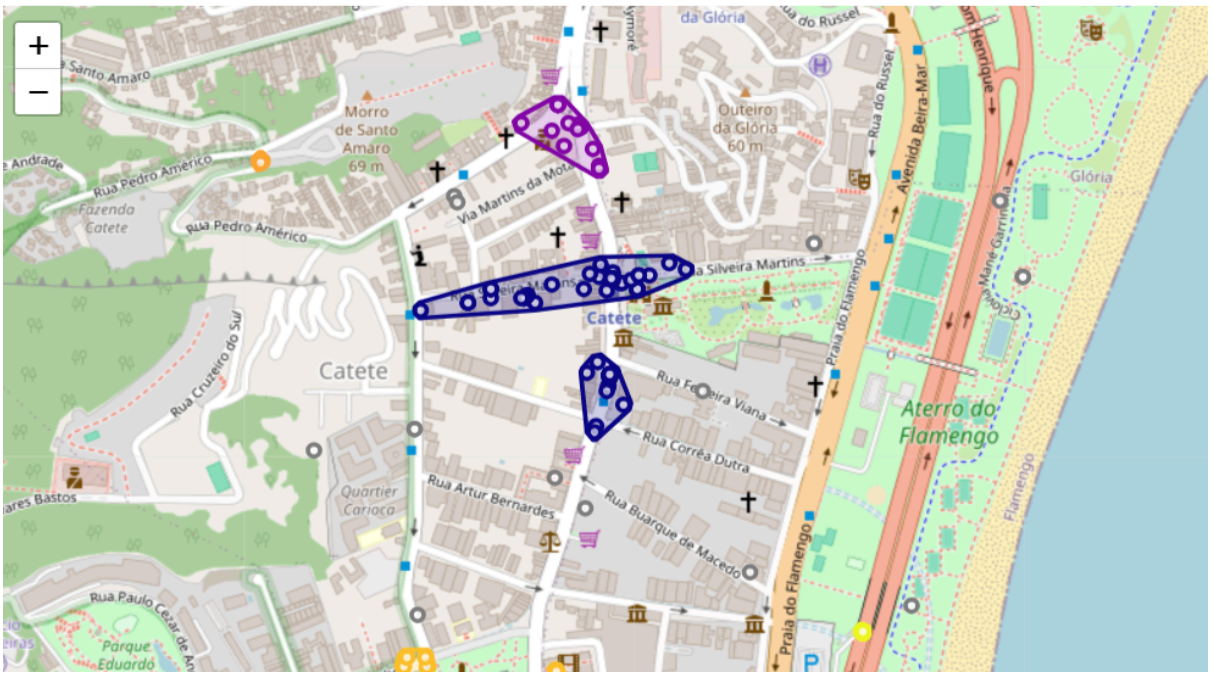
DBSCAN | EPS: 0.0025 | CLUSTERS: 376 | SAMPLES: 73.51 % | LOCATION: Rua do Catete, Catete



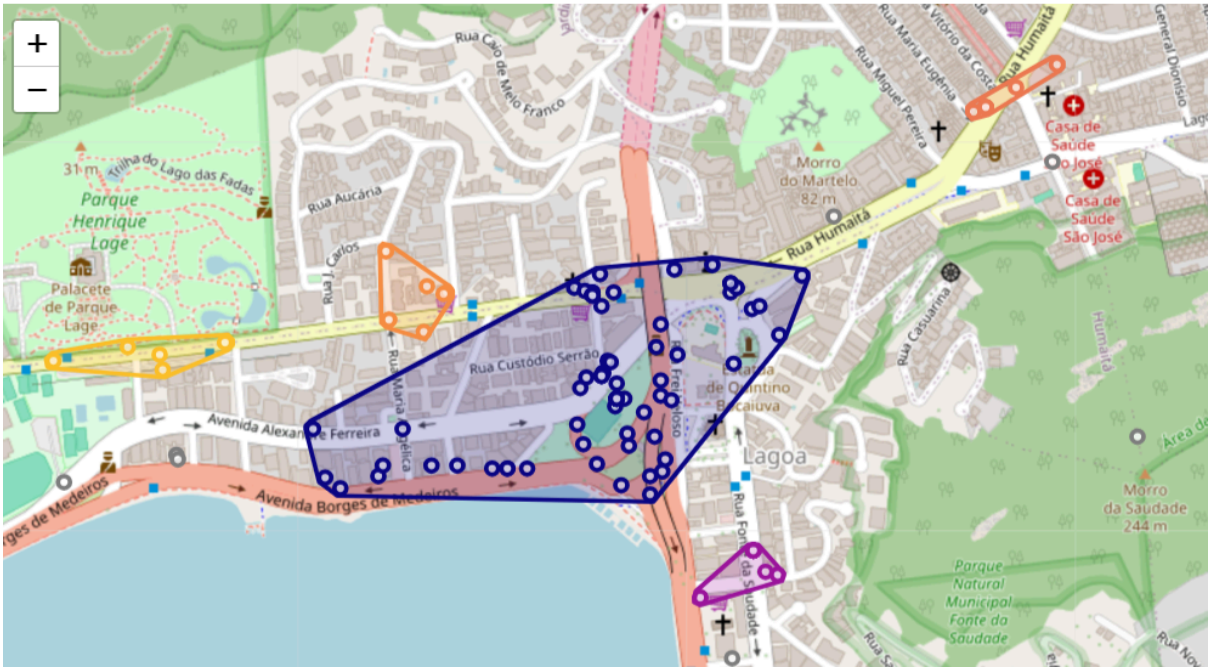
DBSCAN | EPS: 0.002 | CLUSTERS: 388 | SAMPLES: 69.73 % | LOCATION: Rua do Catete, Catete



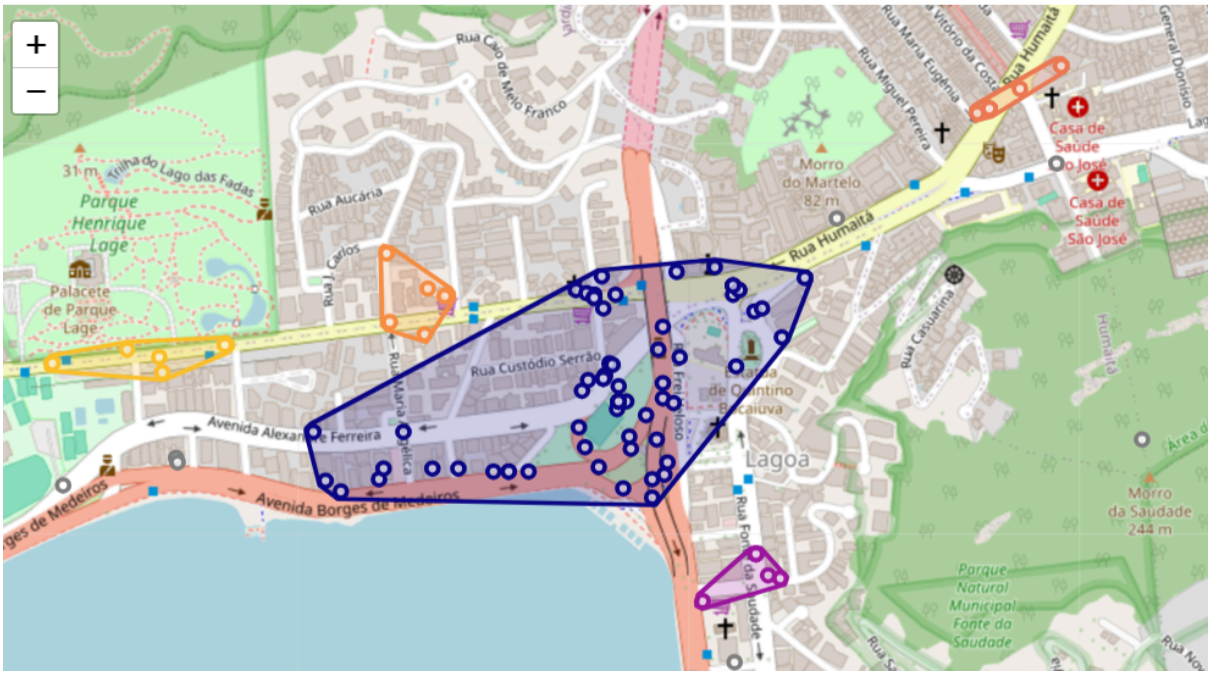
DBSCAN | EPS: 0.0015 | CLUSTERS: 404 | SAMPLES: 64.99 % | LOCATION: Rua do Catete, Catete

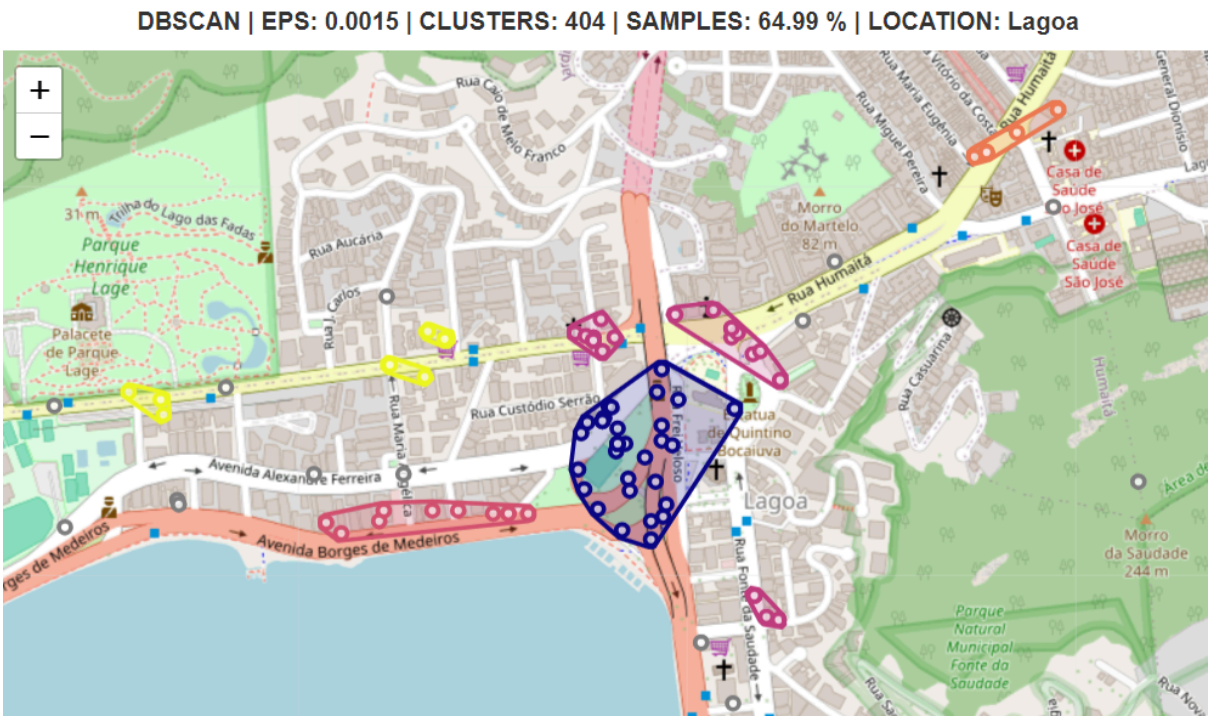


DBSCAN | EPS: 0.0025 | CLUSTERS: 376 | SAMPLES: 73.51 % | LOCATION: Lagoa



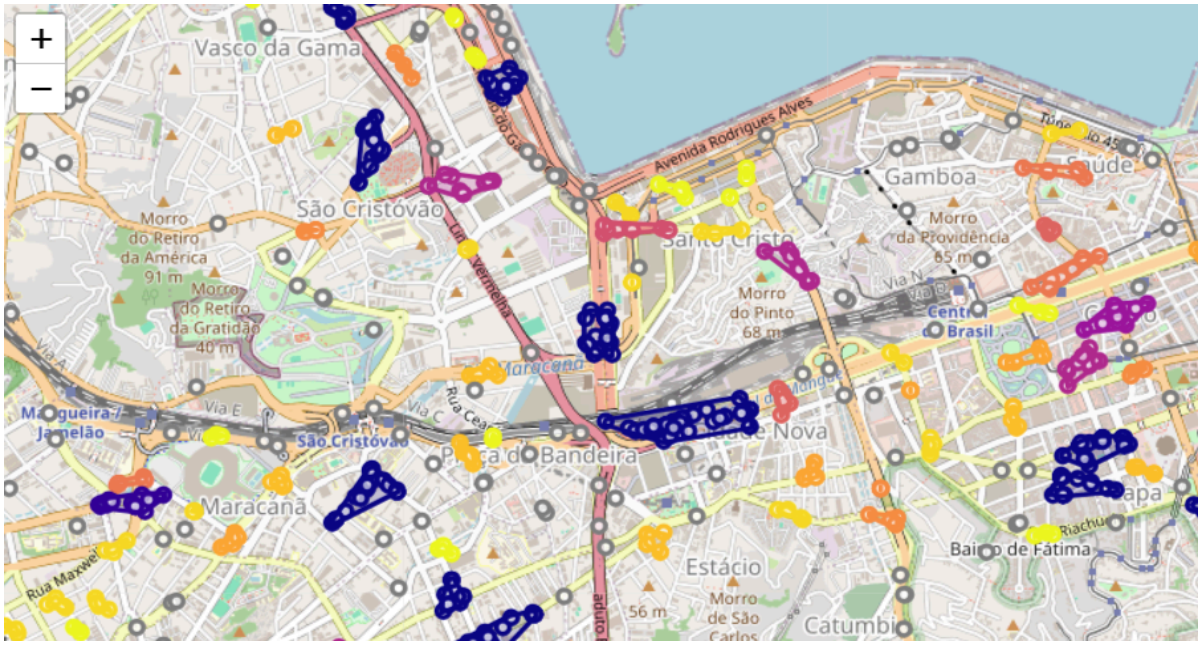
DBSCAN | EPS: 0.002 | CLUSTERS: 388 | SAMPLES: 69.73 % | LOCATION: Lagoa





7. Recommended Final Cluster Model

The map below shows the result of the final DBSCAN model with the selected epsilon value (EPS=0.0025). The value of epsilon for the final model was chosen based on visual evaluation and the balance it presents between the total number of clusters (376) and the percentage of included samples (73.51 %). The map shows the identified flood-related clusters in the center region of Rio de Janeiro. The colormap reflects the number of samples of each cluster, providing a visual hierarchy of the clusters' significance.



Final DBSCAN model with the selected epsilon value (EPS=0.0025).

8. Key Findings and Insights

Optimal Epsilon Selection: The optimization of the epsilon parameter (eps) was important for finding the right balance between cluster granularity and inclusiveness. The epsilon range used in optimization (0.0001 to 0.01) showed that as epsilon increased, the number of clusters decreased, but the percentage of included samples in those clusters increased.

Training Three Cluster Models: Each model (Eps=0.0015, 0.0020, 0.0025) presented a trade-off between cluster granularity and the percentage of included samples, generating insights into the sensitivity of the model to epsilon adjustments.

Cluster Model Evaluation Metrics: Metrics such as average samples per cluster, average area per cluster, and average density per cluster served as essential indicators of the quality and characteristics of the identified clusters.

Spatial Visualization of Flood Clusters: Interactive maps created for three specific locations ("Avenida Armando Lombardi," "Rua do Catete," and "Lagoa") illustrated the clustering process at different epsilon values.

Recommended Final Cluster Model: The final DBSCAN model, with the selected epsilon value (EPS=0.0025), represents a balance between the total number of clusters (376) and the percentage of included samples (73.51%).

9. Next Steps

There are several options for further exploration and improvement of the model's results:

Temporal Analysis: Use temporal features to analyze how flood events evolve over time, identifying patterns, trends, and potential seasonality.

Feature Engineering: Use additional features such as weather conditions, land use, or urban infrastructure, which may enhance the model's ability to discern subtle spatial patterns.
Dynamic Epsilon Selection:

Real-Time Monitoring: Develop a real-time monitoring system that integrates live data streams, enabling timely detection and response to emerging flood events.

Collaboration with Domain Experts: Collaborate with domain experts, local authorities, or environmental agencies to validate and refine the model.

Communication and Visualization: Communication and visualization tools to make the model's outputs accessible to a broader audience.

Continued exploration of these next steps will contribute to the improvement and practical use of the clustering model, ultimately supporting effective disaster management strategies.