Instituto Nacional de Pesquisas Espaciais – INPE
Programa de Pós-Graduação em Computação Aplicada – PGCAP

Machine Learning-Based Flood Risk Assessment at Street Level: A Graph Embedding Approach in the Tamanduateí River Basin

Andre Nogueira Sousa

Professors:
Prof. Dr. Leonardo B. L. Santos Profa. Dra. Juliana Aparecida Anochi
Course:
Advanced Topics in Environmental Modeling

04/12/2025

Outline

Outline

Introduction: The Urban Forecasting Challenge

- **Main Problem (Thesis):** Use of Machine Learning models to create flood alerts at the urban *street* level.
- **Final Goal (Thesis):** Develop a model to predict the occurrence (binary) and/or intensity (regression) of floods on streets with a predictive *lead time*.
- **Main Methodology (Thesis):** Spatio-Temporal Graph Neural Networks (**ST-GNNs**).

- Study Area: Tamanduateí River Basin (SP)
  - High population and infrastructure density.
  - History of recurrent and severe flood events, causing significant losses.
  - Availability of detailed geographic data and an inventory of occurrences (2019).

Outline

Project Timeline: Accomplished in the Previous Classes

- Initial research on **Flood Prediction in Roads (Tamanduateí Basin)**.
- Exploratory Data Analysis (EDA) of geographical data and 2019 flood reports.
- Identified features correlated with the binary target (flood occurrence).
- Used geoprocessing tools and simple **binary classification models** for baseline static correlations.

Project Timeline: Current Project Increment (This Semester)

- **Objective:** Establish a static baseline model and create the core dataset structure.
  - **1. Feature Engineering:** Explore and extract novel features from the geographic model.
  - **2. Road Network Graph Creation:** Develop the **Graph representation of the road network** in the basin.
  - **3. Data Integration:** Combine tabular features with the graph structure.
  - **4. Initial Prediction Model:** Train a simple **binary classification model** (Flood/No Flood) for each road segment in 2019.

Project Timeline: Scope & Next Steps

- **Initial Model:** No temporal dependency (no rainfall data / time series data) is included yet.
- **Current Goal:** Baseline model for **understanding static correlations** and preparing the database structure.
- **Future Work:** Incorporate **temporal data** (e.g., accumulated rainfall) and apply **Graph Neural Networks (GNNs)** for real-time prediction.

Outline

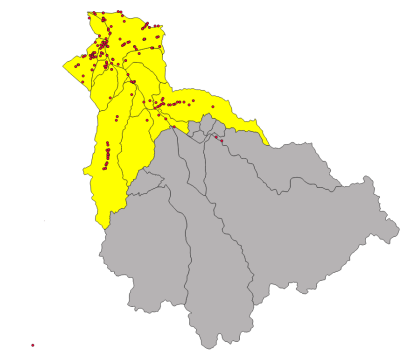
Data Sources: Defining Area of Interest

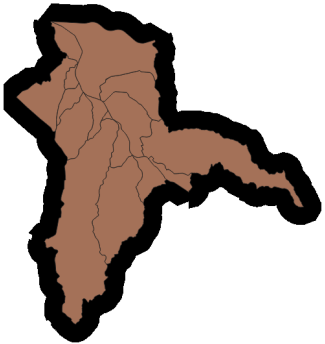
1. **Digital Terrain Model:** Elevation of each point of the interest area
2. **River and Tributaries:** Location of the river and its tributaries
3. **Microbasins:** Divisions in the basin used only to split our data
4. **Flood Label:** Point location of floods occurred in 2019
5. **Street Network:** Graph with the road network of the region

Outline

Data Preparation: Area of interest

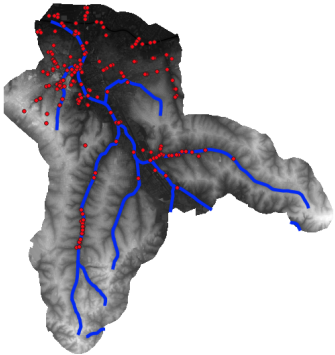
- Region where we have flood occurrence data
- Using Microbasins areas to limit the area
- Applying 1 km buffer around it to avoid edge effect on the graph





Data Preparation: Slicing Properties to the Area of interest

- Given that we don't want the whole basin we need to slice the properties on the area + buffer defined
- Load the graph for the area + buffer defined



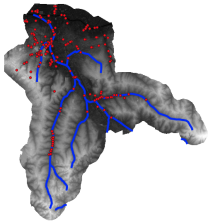
Outline

Feature Engineering:

- **Distance to River**: for each pixel on the area of interest the euclidian distance to the closest river is calculated
- **Slope**: for each pixel on the area of interest the slope of the region is calculated from the DTM
- **Topographic Wetness Index (TWI):** Calculated as a function of the **specific catchment area** (SCA) and the **local slope** (β), it estimates the tendency of a terrain to accumulate water (i.e., soil moisture distribution).

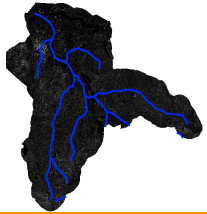
$$\text{TWI} = \ln\left(\frac{\text{SCA}}{\tan\beta}\right)$$

Feature Engineering: Feature Extraction from DTM

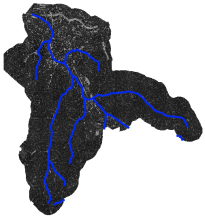


Distance

Slope



TWI

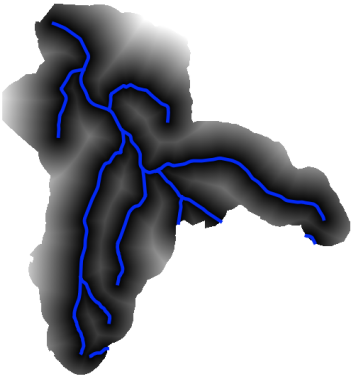


Feature Engineering: Mapping Features to Graph

- **Objective:** Transfer the continuous information from geospatial raster layers to the discrete road network graph.
- **Methodology:**
  - **Overlay:** The graph edges are superimposed on the feature raster layers.
  - **Aggregation:** For each edge (road segment), the **mean value** of the underlying pixels is calculated.
  - **Assignment:** This mean value is stored as a static **property** (attribute) of the edge in the graph.

Feature Mapping: Distance to River

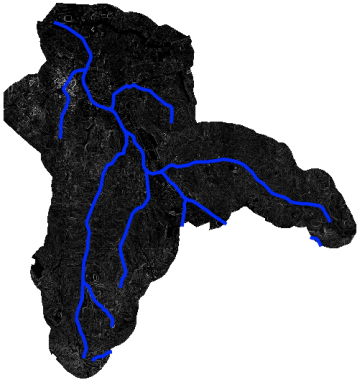
Feature Raster



Graph Property

Feature Mapping: Slope

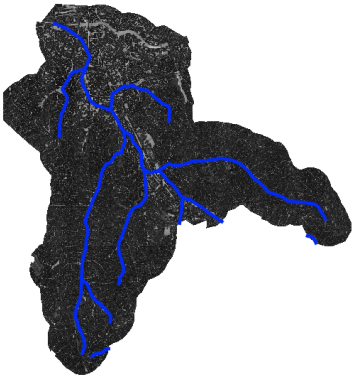
Feature Raster



Graph Property

Feature Mapping: Topographic Wetness Index (TWI)

Feature Raster



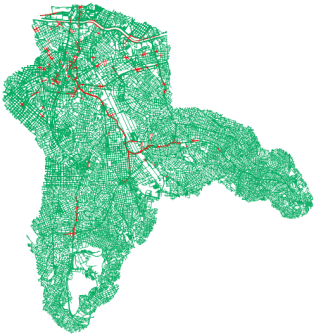
Graph Property

Target Variable Generation: Flood Labels

- **Labeling Process:** Mapping historical flood points to the graph.
- **Criterion:** If a flood point is located within **30 meters** of a road segment, that edge is labeled as a positive class (Flood Event).

Raw Flood Points

Labeled Graph Edges



Outline

Database Generation: Graph Embeddings (Node2Vec)

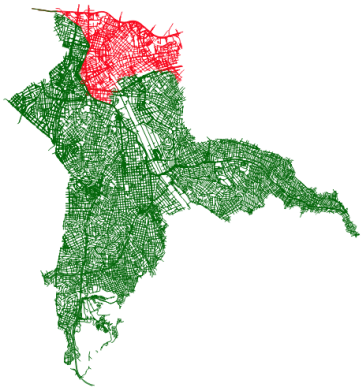
- Graph Characteristics:
  - Nodes: 48739, Edges: 73076
  - Number of Connected Components: 1
  - Density: 0.000062
  - Mean Clustering Coefficient: 0.0354
- **Objective:** Capture the structural and topological context of the street network, generating feature vectors for the machine learning model.
- **Algorithm: Node2Vec**
  - **Random Walks:** Uses random walks across the graph to generate node sequences.
  - **Skip-Gram:** Applies Skip-Gram to the sequences to learn a vector representation (dim = 64) for each node.
- **Edge Feature Construction:**
  - **Embedding Edge:** The embedding vector for each edge (street segment) is calculated as the **mean** of the vectors of its two connecting nodes.
  - **Final Dataset:** The 64-dimensional embedding vector is concatenated with the static geospatial properties (TWI, Slope, Distance, Label) to form the final training dataset.

Outline

Training: Geographic Train-Test Split Strategy

- **Spatial Leakage Prevention:** Standard random splitting leads to data leakage due to spatial autocorrelation. A **Geographic Split** based on Microbasin IDs was adopted to ensure independence.

- **Split Configuration:**
  - **Training Set:** 36,426 streets (83.55%).
  - **Test Set (Hold-out):** 7,172 streets (16.45%).
  - Selected Microbasin for Test: **'AC 1.1.12'** (High flood rate).

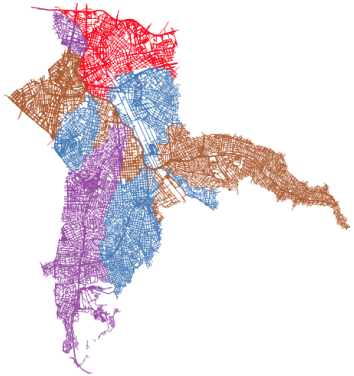


Training: Geographic Cross-Validation (3-Fold)

- **Methodology:** A 3-Fold GroupKFold was applied to the Training set.
- **Balance:** Folds were strategically grouped to maintain representative flood rates and street volumes.

Distribution Statistics:

Group	Streets	Flood Rate	% Data	% Flood
Test (Hold-out)	**7,172**	**6.09%**	**16.45%**	**27.45%**
Fold 1	11,802	3.76%	27.07%	27.89%
Fold 2	13,385	4.51%	30.70%	37.94%
Fold 3	11,239	0.95%	25.78%	6.72%
Total	**43,598**	**3.65%**	**100%**	**100%**



Model Training & Optimization

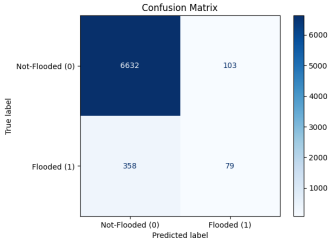
- **Algorithm: CatBoostClassifier** (Gradient Boosting on Decision Trees).
- **Validation Strategy:**
  - **Strategic Split:** Train (83.55%) / Test (16.45%) based on microbasins to test generalization on unseen regions.
  - **Cross-Validation: GroupKFold** ($k = 3$) ensuring streets from the same microbasin stay in the same fold to prevent data leakage.
- **Hyperparameter Tuning (Optuna):**
  - **Objective:** Maximize **ROC AUC**.
  - **Trials:** 50 iterations.
  - **Best Parameters Found:**
    - `iterations`: 238
    - `learning_rate`: 0.0147
    - `depth`: 10
    - `l2_leaf_reg`: 4.81

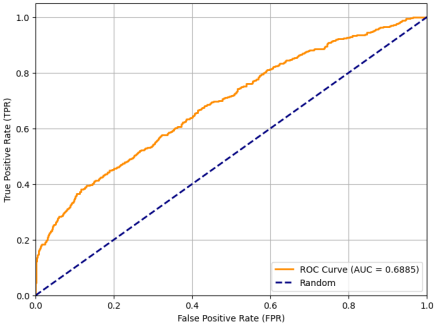
Outline

Model Evaluation: Test Set Metrics

- **Test Set:** Unseen microbasin (*AC 1.1.12*).
- **Class Imbalance:** Addressed using `scale_pos_weight`.

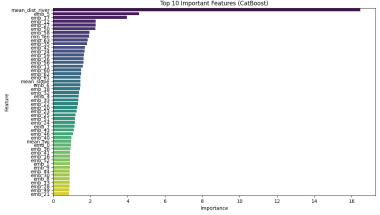
Metric	Value
Accuracy	93.57%
ROC AUC	**0.6885**
F1-Score	0.2553





Feature Importance Analysis

- **Top Predictor: Distance to River** is the most dominant feature ($\approx 16.5\%$), confirming hydrological intuition.
- **Graph Embeddings:** Multiple embedding dimensions (*emb_5*, *emb_17*, *emb_12*) rank highly, proving the network structure captures relevant spatial patterns.
- **Terrain Features: TWI** and **Slope** appear with moderate importance (ranks $\approx 20 - 37$), contributing to local flood susceptibility.



Outline

Conclusion: Static Baseline Results

- **Successful Data Integration:** The methodology for mapping continuous geospatial features (TWI, Slope, Distance) and network structure (Node2Vec embeddings) to the discrete road graph was implemented, creating the static dataset.
- **Model Performance:** The optimized CatBoost model established a viable baseline for binary flood classification, demonstrating an ROC AUC ≈ 0.69 on an unseen microbasin.
- The use of **Geographic Split by Microbasin** and **GroupKFold** prevented data leakage (spatial leakage), ensuring that the model was tested in an unseen region.
- **Feature Insights:**
  - **Hydrological Dominance: Distance to River** emerged as the single most critical predictor ($\approx 16.5\%$ importance).
  - **Structural Relevance:** Graph embeddings captured significant topological information, ranking among the top features and validating the use of the graph structure.

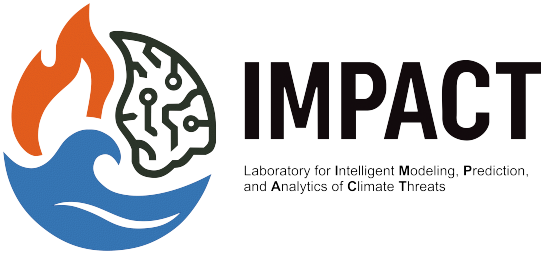
Outline

Future Work: Moving to Spatio-Temporal Prediction

- **Phase I: Incorporating Temporal Data and improving mapping**
  - **Data Collection:** Acquire high-resolution, time-series rainfall data (e.g., accumulated rain over 1h, 3h, 6h) for the Tamanduateí basin.
  - **Edge Feature Refinement:** The current aggregation method (**average**) for **a single value per edge** is a simplification for long streets. Evaluate the use of **multiple sampling points** per edge or more sophisticated spatial aggregation methods.
  - **Feature Creation:** Integrate temporal data as dynamic node/edge features, transforming the dataset into a sequence structure.
- **Phase II: Spatio-Temporal Graph Neural Networks (ST-GNN)**
  - **Model Transition:** Move from the static CatBoost model to a dynamic ST-GNN architecture.
  - **Prediction Goal:** Achieve real-time flood prediction (binary classification) for road segments with a specific predictive lead time (e.g., 30-60 minutes).
  - **Refinement:** Explore advanced loss functions or sampling techniques to further address the severe class imbalance issue.

IMPACT Lab

- Machine Learning-Based Flood Risk Assessment at Street Level: Source Code and Data. GitHub repository, 2025. Available at: https://github.com/andrenogueira88/inpe-tamanduatei-graph-embeddings-floods



Thank you!

André Nogueira Sousa

andrenogsousa@gmail.com