Clustering and Labeling of a V2V Communication Dataset

based on CAM and DENM Messages with Malicious Data Injection

Analysis and Implications

ALESSANDRA BLASIOLI - PROJECT WORK

INTRO

VEHICULAR COMMUNICATION SYSTEMS

Computer networks
where vehicles and roadside units (RSU) serve as communicating nodes

VEHICLE-TO-VEHICLE COMMUNICATION (V2V)

Wireless information exchange about the speed and position of nearby vehicles.
Offering great promise in accident prevention.

INTRO

DATASETS

Two different types of messages: Cooperative Awareness Messages (CAM) and Decentralized Environmental Notification Messages (DENM).

OBJECTIVE OF THE WORK

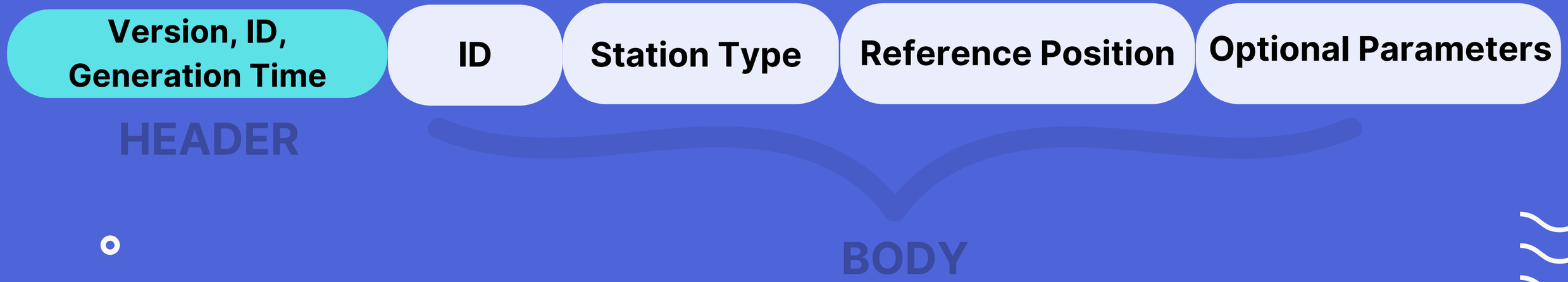
Data preprocessing, clustering, introduce noise into the provided dataset to simulate potential malicious reports, labeling the data based on clustering and identifying potential outliers.

CAM MESSAGES

Cooperative Awareness Message

defined by the European Telecommunications Standards Institute (ETSI) in 2011

- Basic awareness service by sending status data to **nearby nodes**
- Distributing messages about **presence**, **location**, and **fundamental status**



DENM MESSAGES

Decentralized Environmental Notification Message
defined by the European Telecommunications Standards Institute (ETSI) in 2011

- Notification service regarding **road status**
- Support active road **safety** applications

Version, ID, Generation Time

Management

Situation

Location

HEADER

BODY

DATA ANALYSIS

X-Means Algorithm

A **variant of K-Means algorithm**, determines **automatically** the optimal number of clusters in the data without requiring a predefinited specification, based on **recursion**. Part of pyclustering open-source library.

- Initially applies **standard K-Means** with an initial cluster count
- Assesses **clustering quality using measures**
- Checks if **splitting clusters** improves overall quality
- **Divides clusters with K-Means** if advantageous

DATA ANALYSIS

X-Means Algorithm

A **variant of K-Means algorithm**, determines **automatically** the optimal number of clusters in the data without requiring a predefinited specification, based on **recursion**. Part of pyclustering open-source library.

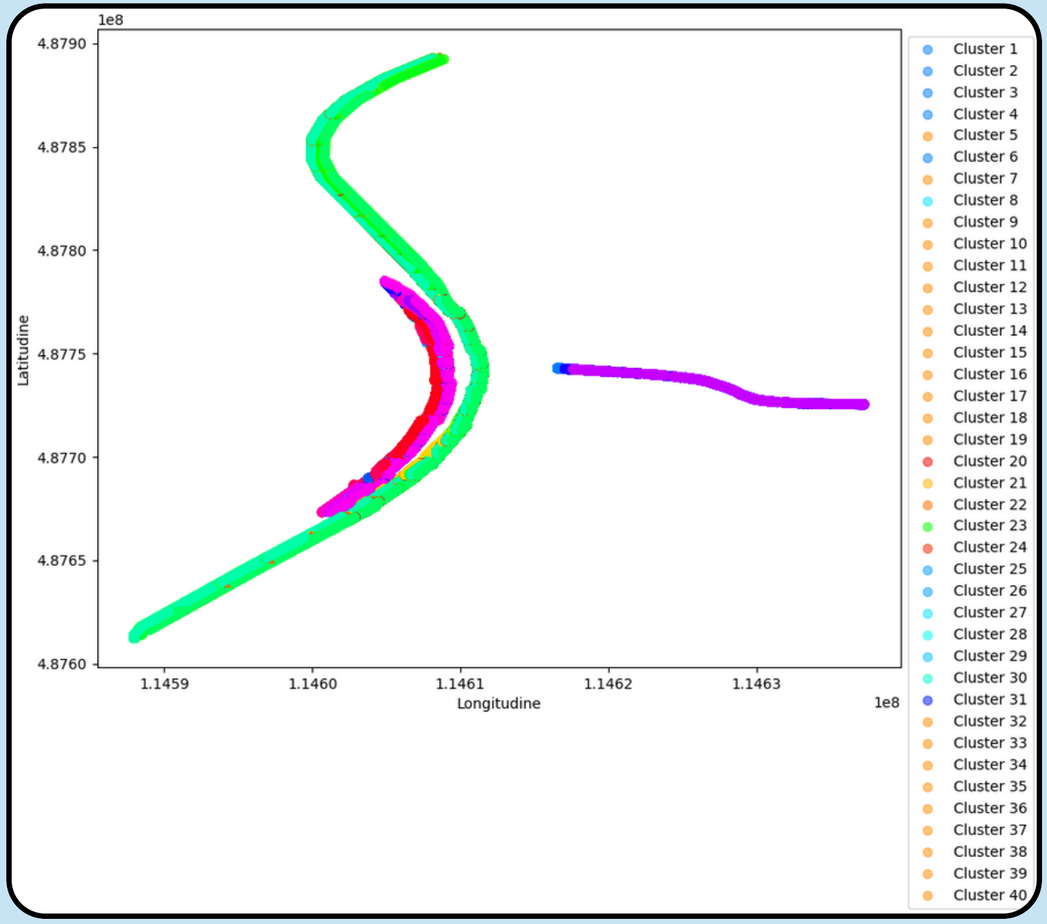
- Initially applies **standard K-Means** with an initial cluster count
- Assesses **clustering quality using measures**
- Checks if **splitting clusters** improves overall quality
- **Divides clusters with K-Means** if advantageous

Repeats division and evaluation for existing clusters and **allows potential creation of new sub-clusters for improved clustering**

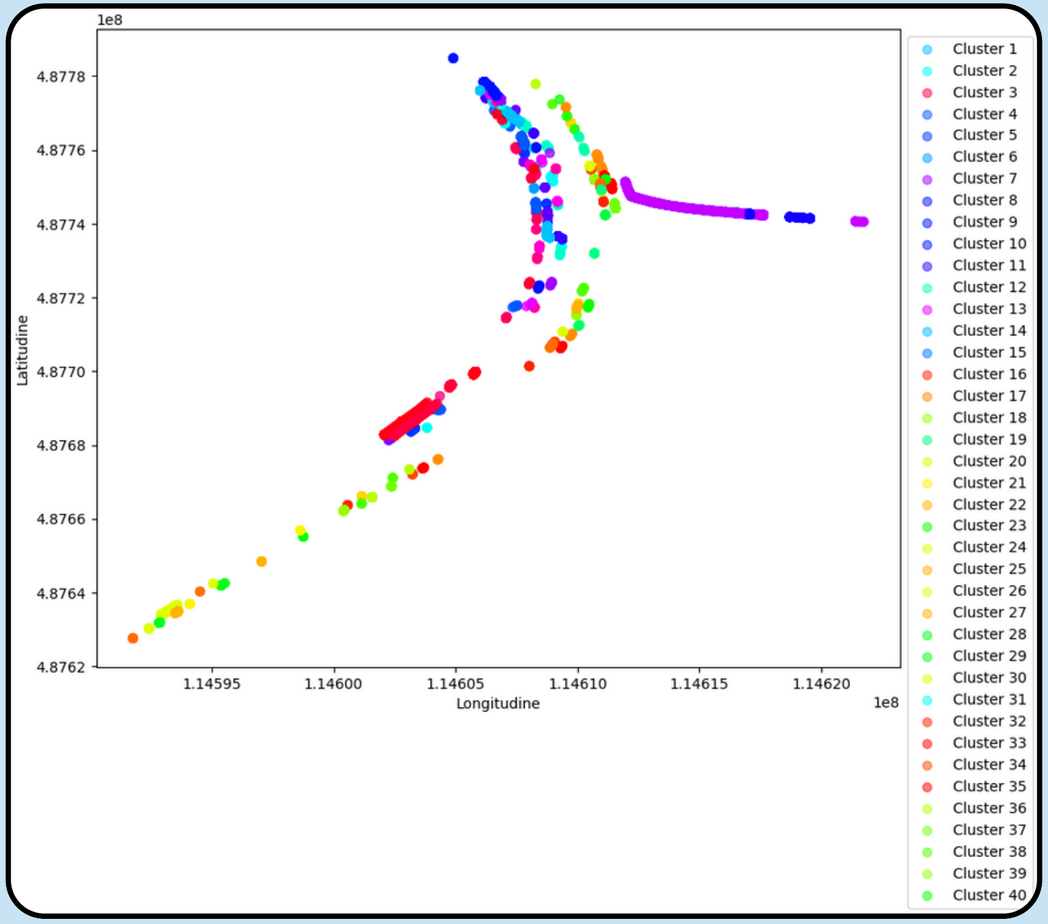
- Stops when **no further cluster division is possible**
- Halts if **the division doesn't significantly enhance results** compared to complexity

Returns **optimal clusters determined automatically**

DATA ANALYSIS



Clustering of the provided CAM dataset



Clustering of the provided DENM dataset

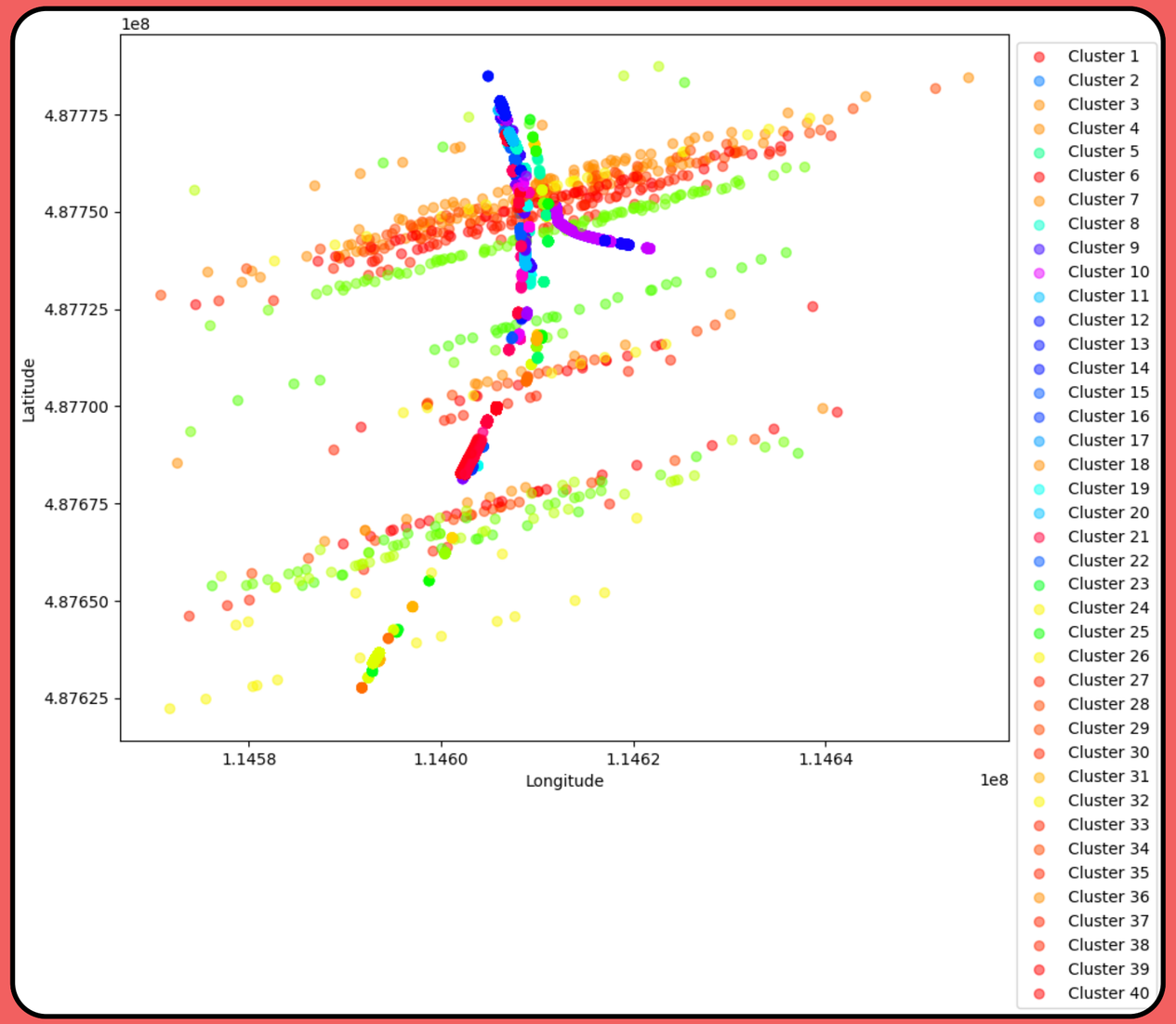
INTRODUCTION OF MALICIOUS DATA

Variation of coordinates using a **Gaussian distribution** (mean: 0, standard deviation: 1) multiplied by a factor of 100

Variations within a range of approximately 100 to 900 meters on the map

- Contamination of data based on the number of sources to simulate malicious vehicles
- Focus contamination on **sources with eventType value 97**, representing the most significant cluster
- Contamination of **20% (8 sources) of this specific eventType**

INTRODUCTION OF M



Clustering Following
Data Contamination

EVENT TYPE ANALYSIS

3 Specific Event Types

CauseCodeType_dangerousEndOfQueue = 27
CauseCodeType_collisionRisk = 97
CauseCodeType_trafficCondition = 1

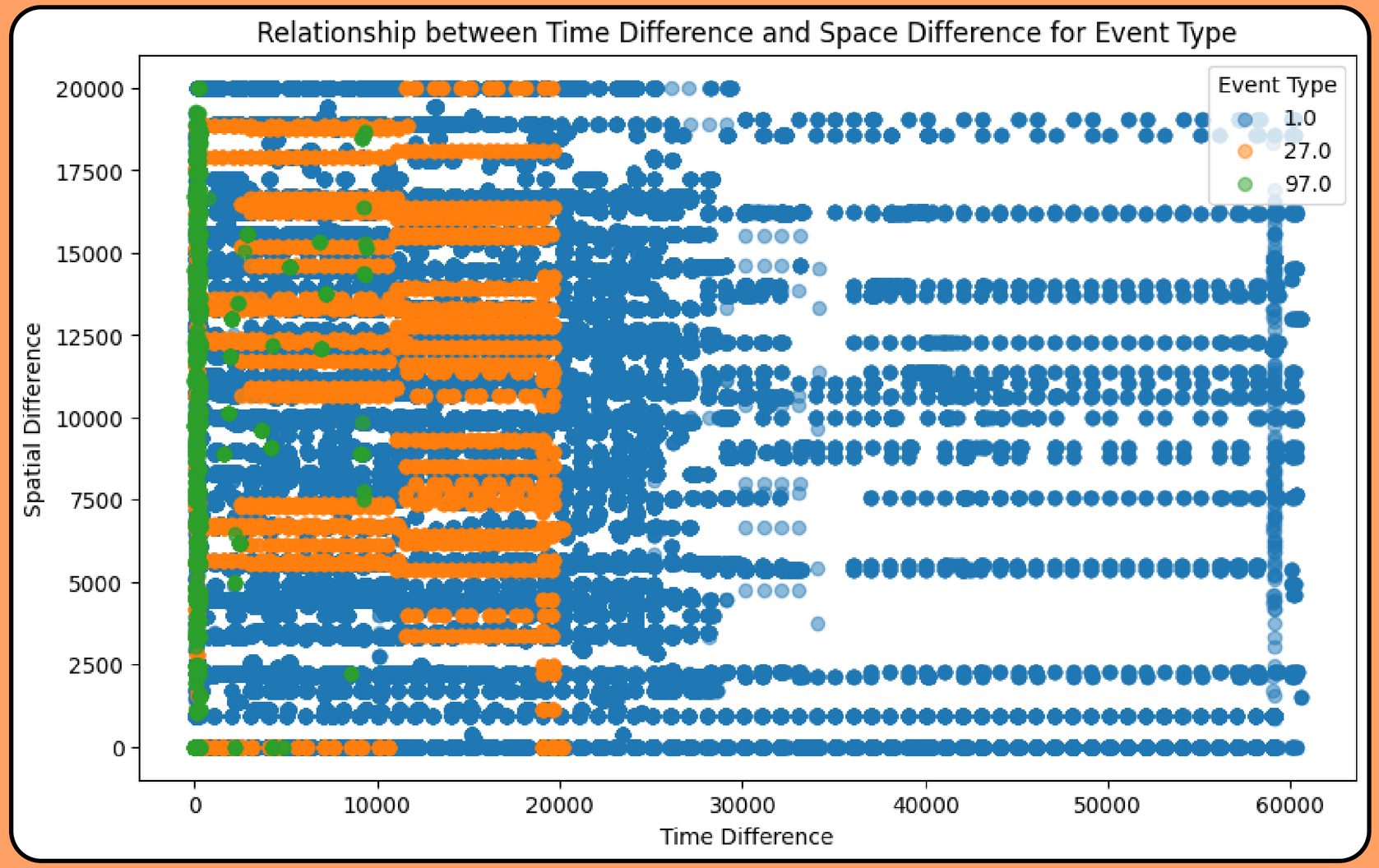
Time difference between simulation time and detection time

The time differences between **when a vehicle perceives an event and when it reports it.**

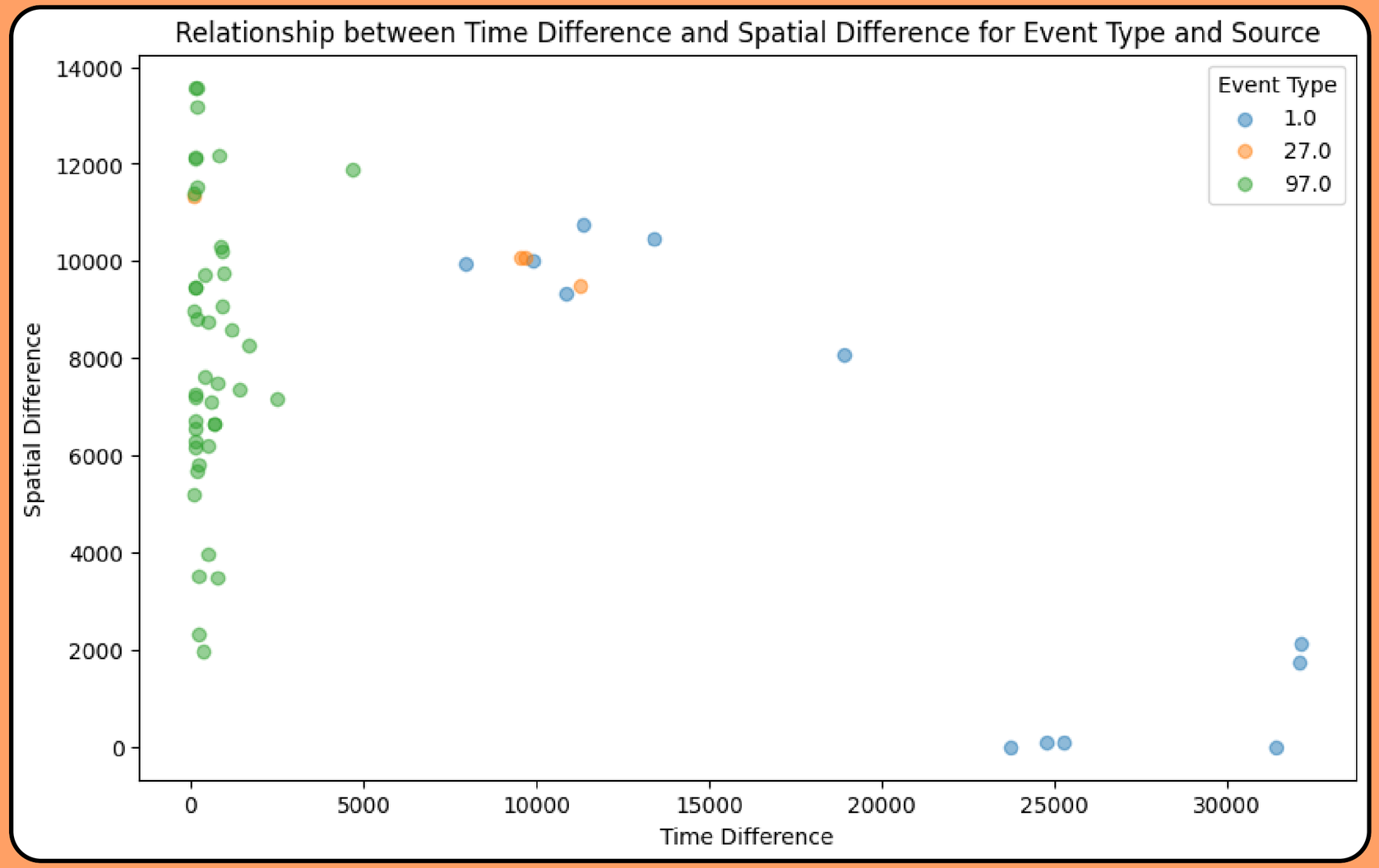
Spatial variation between CAM and DENM messages from the same source within a minimal time gap

Graphical representation

EVENT TYPE
A


Relationship between Time Difference and Space Difference for Event Type
Event Type
1.0
27.0
97.0
Spatial Difference
20000
17500
15000
12500
10000
7500
5000
2500
0
0
10000
20000
30000
40000
50000
60000
Time Difference

EVENT TYPE A


Relationship between Time Difference and Spatial Difference for Event Type and Source
Spatial Difference
Time Difference
Event Type
1.0
27.0
97.0
14000
12000
10000
8000
6000
4000
2000
0
0
5000
10000
15000
20000
25000
30000

OUTLIER DETECTION

- **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) **algorithm**

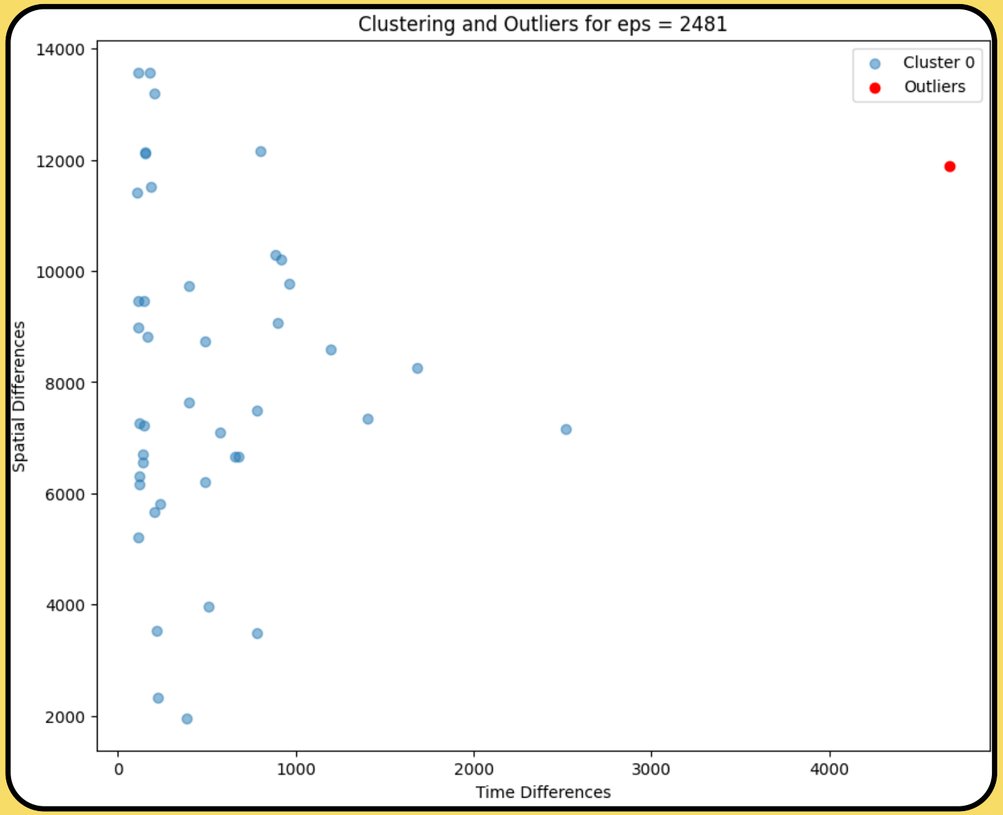
⟶ Groups points based on the data density in space

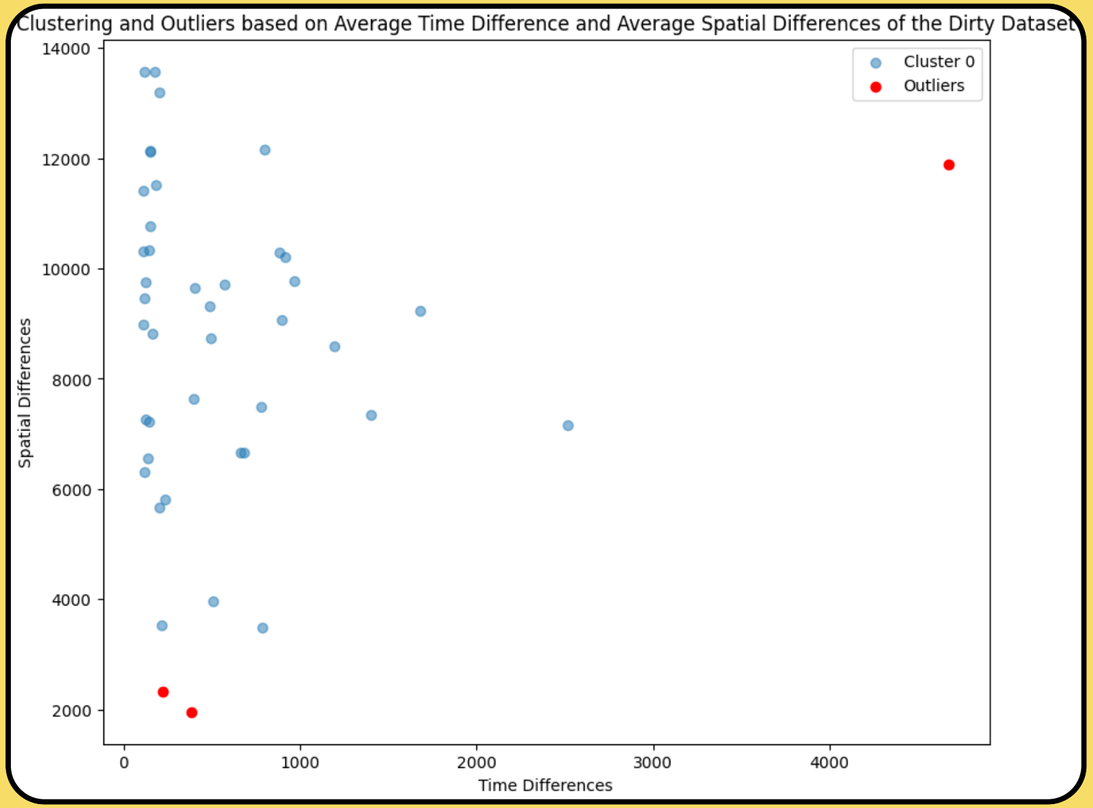
⟶ Two parameters:
  - **epsilon (ε)**, the maximum distance between two points to consider them part of the same cluster;
  - **minPoints**, the minimum number of points required to form a cluster.

- Experiment varying **epsilon values** to find the optimal configuration

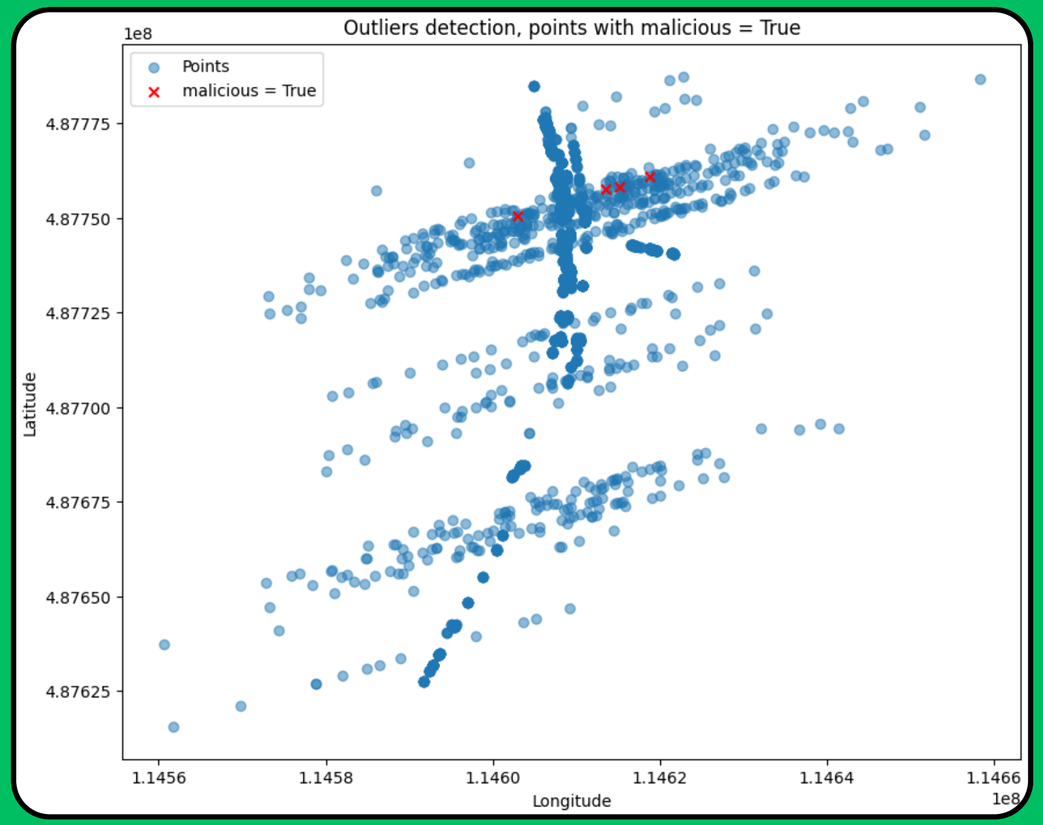
⟶ The best epsilon is the fewest possible outliers, the data we are working with is clean!

OUTLIER
D

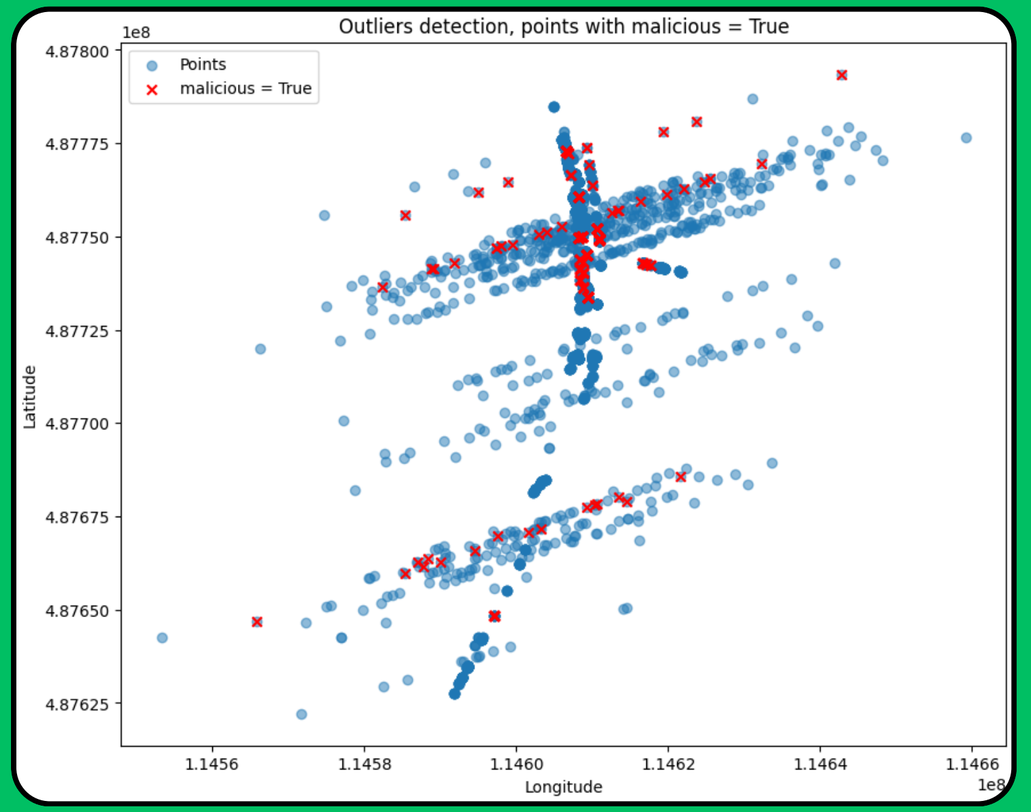

Clustering and Outliers for eps = 2481
Cluster 0
Outliers
Spatial Differences
Time Differences
14000
12000
10000
8000
6000
4000
2000
0
1000
2000
3000
4000


Clustering and Outliers based on Average Time Difference and Average Spatial Differences of the Dirty Dataset
Cluster 0
Outliers
Spatial Differences
Time Differences
14000
12000
10000
8000
6000
4000
2000
0
1000
2000
3000
4000

LABELING - EXPERIMENTS

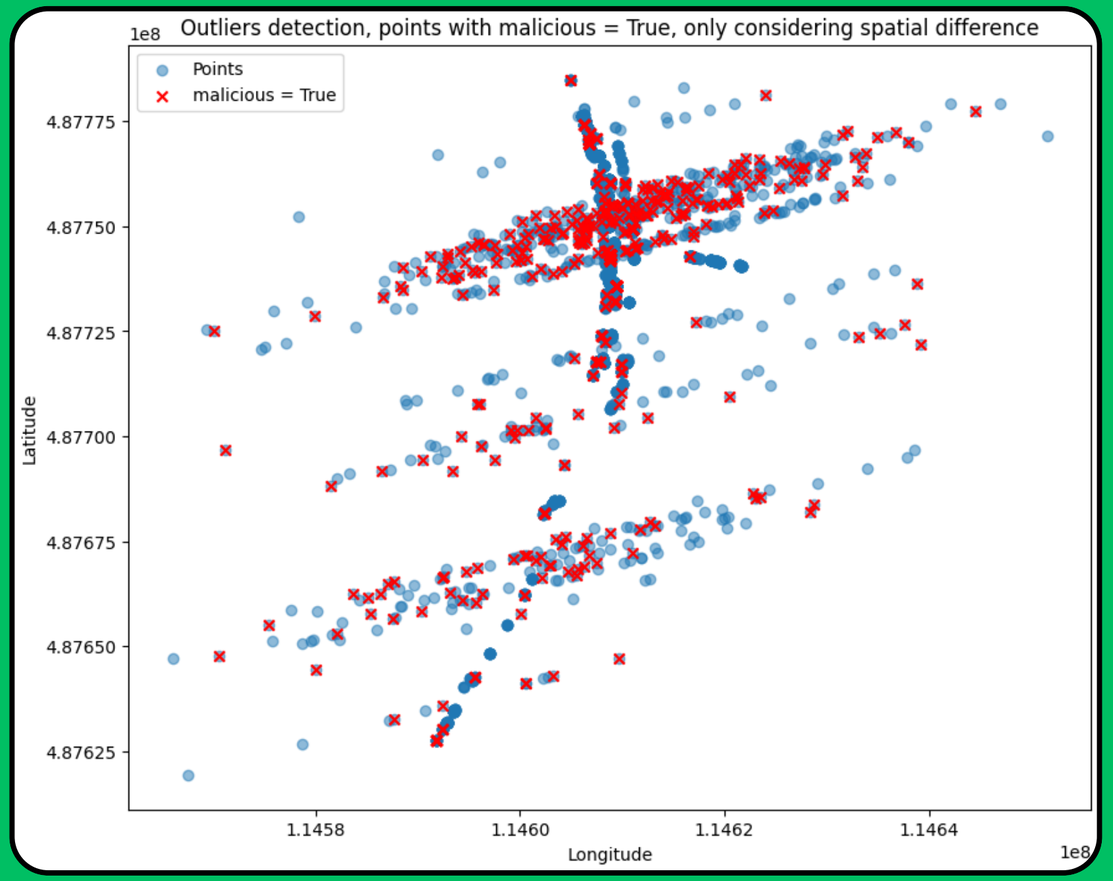


Outliers detection for epsilon = 2481 with the dirty dataset



Outliers detection for epsilon = 1281 with the dirty dataset

LABELING - EXPERIMENTS



TP	FN	FP	TN
38.66%	61.34%	14.44%	85.56%

Outliers detection, points with malicious = True, only considering spatial difference

CONCLUSIONS

Challenges in outlier identification, especially for **points near trajectories**.

Substantial improvement observed by removing the **temporal component**.

Best-performing experiment captured only 38.66% outliers.

Consider utilizing a **more extensive dataset** with diverse event types to enhance the robustness and generalizability of the model.

Explore more advanced **machine learning and deep learning** techniques.

Thanks for the attention!