FINAL PROJECT

Geothermal Anomaly Detection Using Satellite-Derived NDVI, LST, and InSAR: A Machine Learning Approach in Solok, Indonesia

Elistia Liza Namigo

REA6GDWBY

Background (motive)

Indonesia holds 40% of global geothermal reserves — yet exploration is costly, labor-intensive, and invasive.

Remote Sensing (RS) offers a scalable and eco-friendly alternative using satellite-based thermal, vegetation, and deformation data.

But: <u>Ground-truth data is scarce in remote geothermal regions.</u>

👉 Unsupervised Machine Learning (e.g., KMeans, Isolation Forest) detects hidden patterns and anomalies without labeled data.

🎯 This project integrates RS + Unsupervised ML to detect geothermal anomalies in Solok, West Sumatra — a model for low-cost exploration across Indonesia.

Objective of the Model

To develop a **remote sensing-based geothermal anomaly detection pipeline** using unsupervised machine learning.

This model identifies and prioritizes locations in the Solok region with surface characteristics indicative of geothermal activity—specifically focusing on **thermal, vegetative, and surface deformation signals** derived from satellite data.

Comprehensive Model Description

- This project integrates **two unsupervised learning models**:

1. **KMeans Clustering**
    1. Groups data into 3 clusters in NDVI–LST–InSAR feature space.
    2. Anomaly score is calculated based on **Euclidean distance to cluster centroid**.
    3. Points far from centroids are flagged as **statistical anomalies**.
2. **Isolation Forest**
    1. Detects anomalies by evaluating how easily a point can be isolated.
    2. Effective for capturing **rare combinations** in multi-dimensional data.
    3. Output is binary: normal (1) or anomaly (-1).

- These models are applied independently and then **combined to isolate high-priority geothermal zones**, i.e., points classified as anomalies in both models.

Effectiveness Metrics

Since unsupervised models lack labels, indirect validation methods are used:

- **Anomaly Score**
  Distance from cluster centroid (KMeans) as a proxy for rarity.
- **Model Agreement**
  Points flagged by **both** KMeans and Isolation Forest are considered more reliable.
- **Geographic Coherence**
  Visual overlays of anomalies against known **geothermal sites**, **fumaroles**, and **hot springs** assess spatial correlation.
- **Cluster Summary Statistics**
  Mean NDVI, LST, and VV per cluster to interpret geothermal-likeness.

Dataset Overview(1)

- Source: Google Earth Engine (GEE)
- Study Area: Solok & Mt. Talang geothermal zone (West Sumatra)
- Resolution: ~500 m grid over 10 km × 10 km area
- Time Range: 2–3 years composite (monthly averaged)

Variables Used:

Variable	Source	Purpose
NDVI	Landsat 8	Proxy for vegetation health (0–1)
LST (°C)	Landsat 8 TIRS	Surface thermal emission
InSAR VV (mm)	Sentinel-1	Vertical ground deformation

Dataset Overview (2)

Preprocessing Steps:

- Dropped missing or invalid values
- Extracted coordinates from .geo if needed
- Filtered physically implausible values:
  - NDVI: [0, 1]
  - LST: [–20°C, 60°C]
  - VV: [–20 dB, +5 dB]
- Scaled features using StandardScaler

Methodology(1)

1. **Feature Selection**: NDVI, LST, and VV as key geothermal indicators.
2. **Feature Scaling**: Standardized to zero mean and unit variance.
3. **Unsupervised Learning**:

   - KMeans (k=3) clustering : to explore groupings.

   - Isolation Forest : for probabilistic anomaly detection.

4. **Anomaly Scoring**:

   - KMeans: distance from centroid

   - Joint Filtering: intersected both models to reduce false positives

Methodology(2)

5. Geospatial Mapping:

- Visualized anomalies in 2D (Folium) and 3D (Matplotlib)
- Overlaid with:
  - Known geothermal infrastructure
  - Volcanoes and hot springs
  - Anomaly heatmaps (folium plugins)

6. Ranking & Field Prioritization:

- Top-10 anomaly locations scored using composite index:
  Anomaly Score = LST - NDVI - VV
- Exported as black-dot markers for potential field surveys

🔺 LST ↑ 🔻 NDVI ↓ ↕ InSAR shift = Likely geothermal anomaly 🔥

- When **LST**, **NDVI**, and **InSAR** anomalies are considered **jointly**, they form a **multi-dimensional signature** of surface behavior that can strongly indicate geothermal activity.

Parameter	What It Signals	Geothermal Context
LST ↑	Elevated ground surface temperature	Suggests subsurface heat flow
NDVI ↓	Vegetation stress or absence	Indicates thermal damage or toxic gas emissions
InSAR shift (↑ or ↓)	Vertical ground deformation	Implies pressurization, inflation, or subsidence due to fluid movement

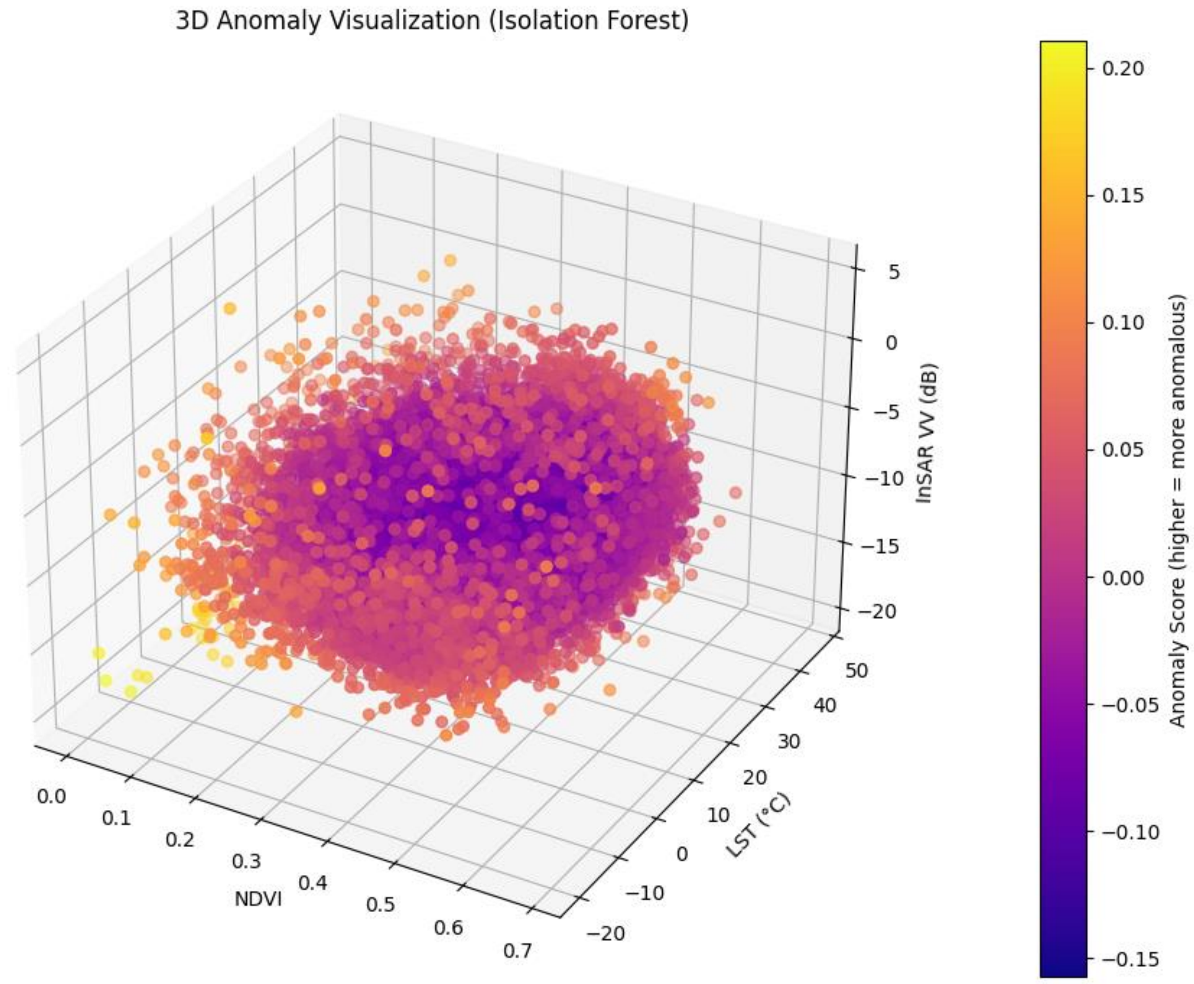
Individually, each of these parameters **may have false positives** (e.g., urban heat, deforestation, landslide deformation). But **together**, they form a **corroborative anomaly signature**.

Results

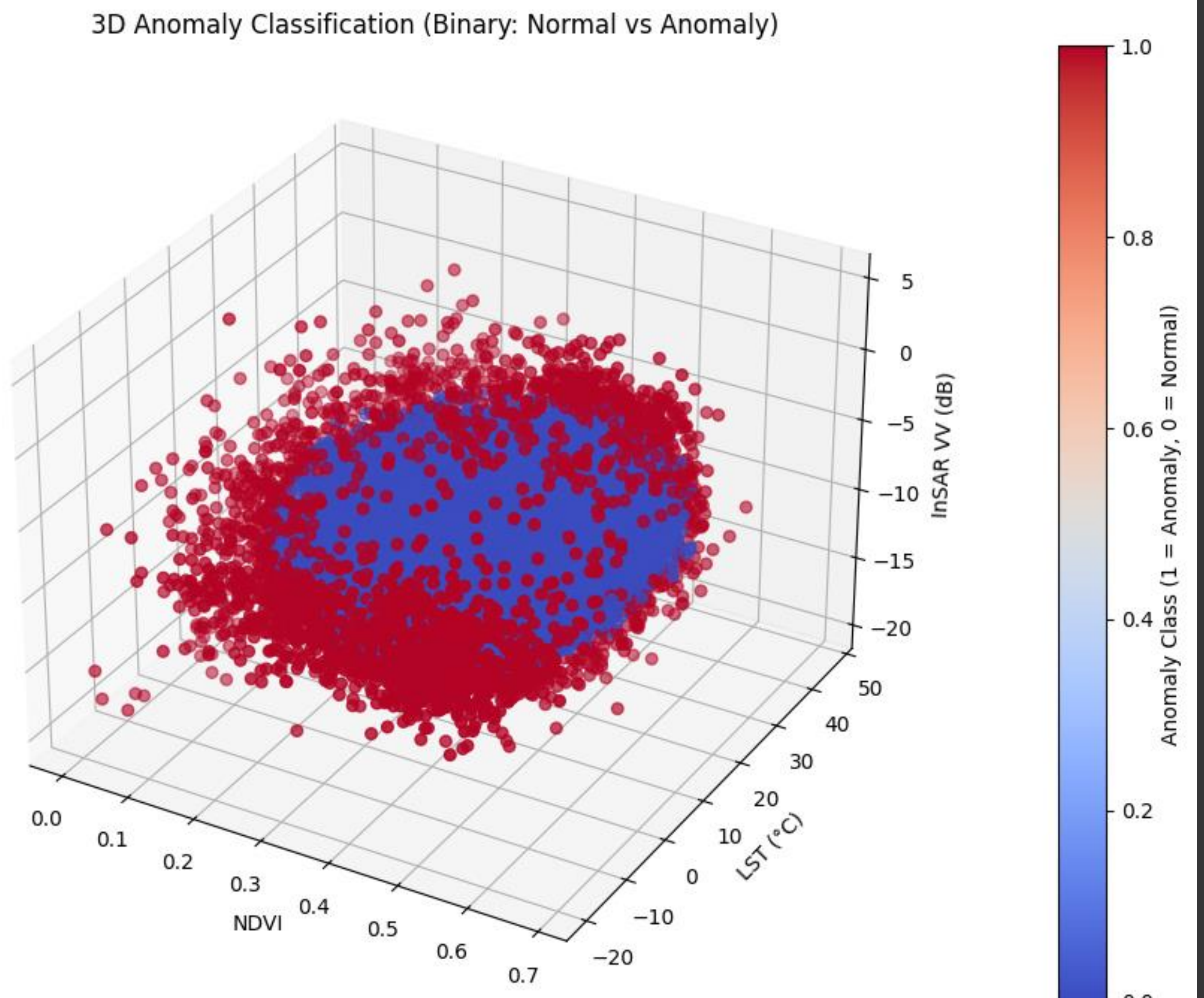
Isolation Forest (iForest)

3D Anomaly Maps

These two 3D plots offer different but complementary perspectives on the **Isolation Forest** anomaly detection applied to geothermal prospecting dataset (NDVI, LST, VV):

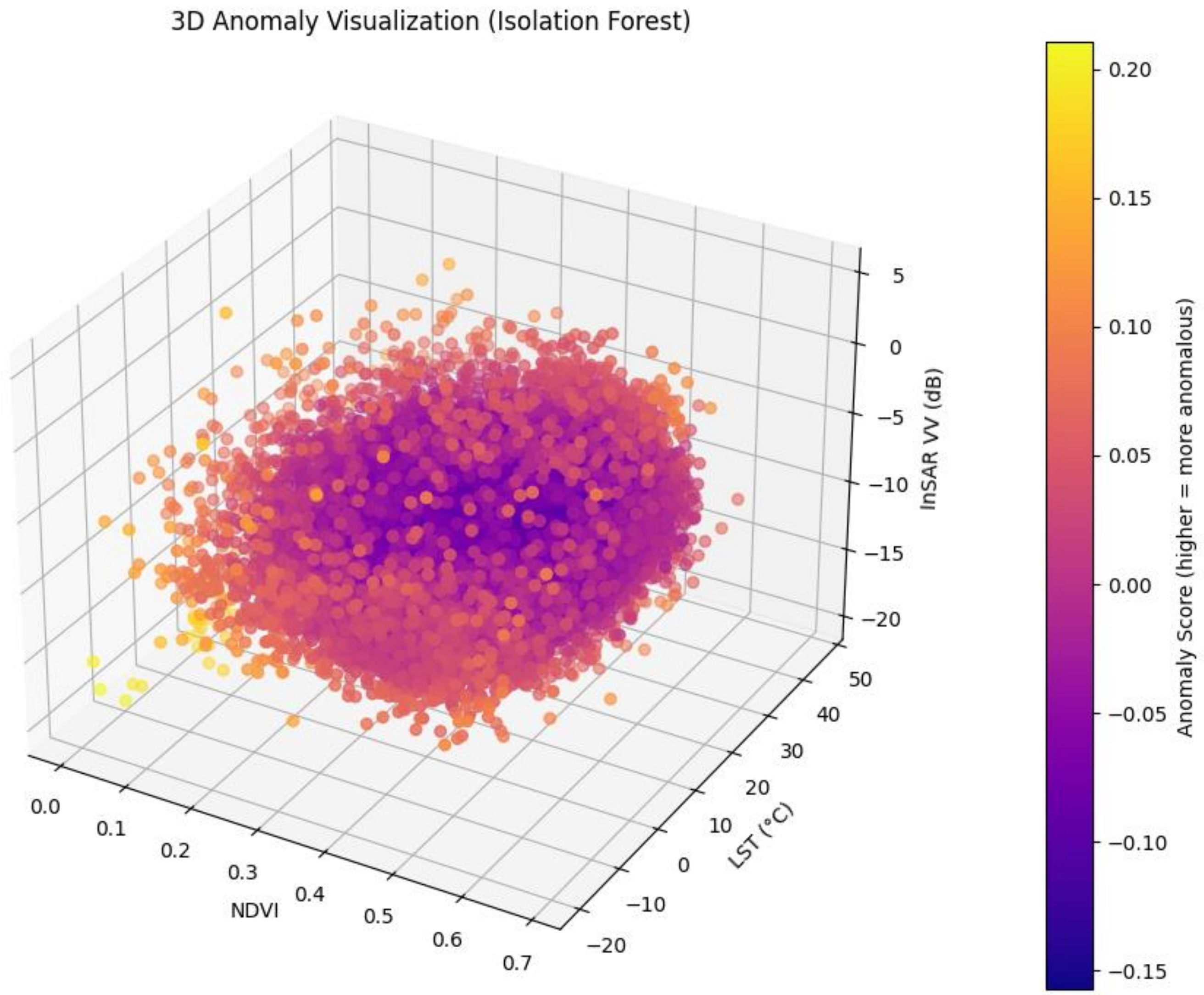


Colored by anomaly score



Colored by anomaly score

3D Anomaly Maps : What does it say?



3D Anomaly Visualization (Anomaly Score)

Plot Insight:

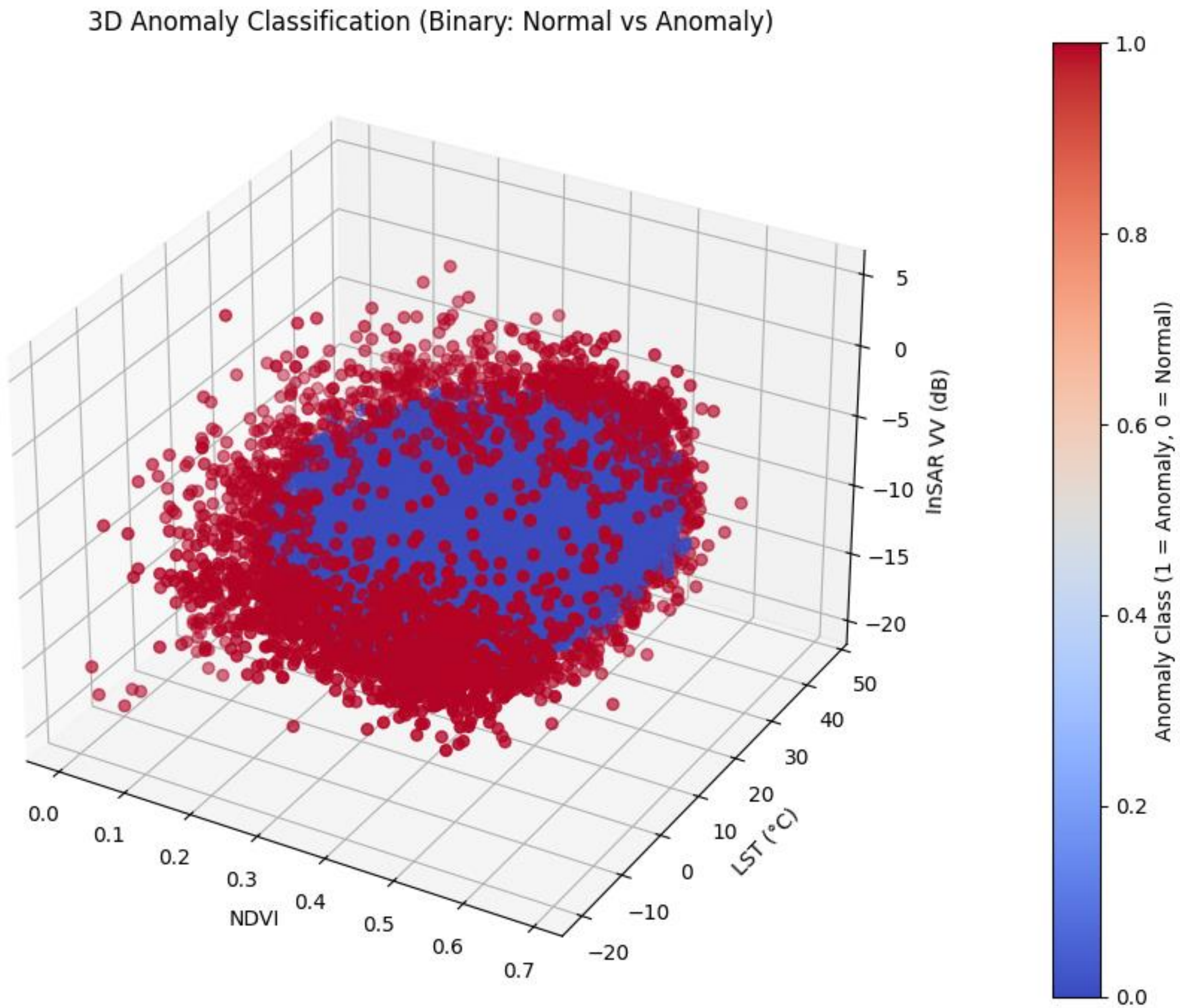
•**Color gradient (plasma)** represents the **anomaly score**, where:

- **Lighter colors (yellow)** = higher anomaly scores → **more anomalous**
- **Darker colors (purple)** = normal values

•**Interpretation:**

- Points farther from the main data cloud (in 3D feature space) tend to have **higher anomaly scores**.
- These points may correspond to geothermal signatures:
  e.g., high LST + low NDVI + ground displacement (VV).

3D Anomaly Maps (2) : What does it say?



3D Binary Classification (0 = Normal, 1 = Anomaly)

Plot Insight:

This version uses the final classification output of the Isolation Forest:

Red points = Anomalies (1)

Blue points = Normal (0)

Interpretation:

It visually segments the anomaly boundary, showing:

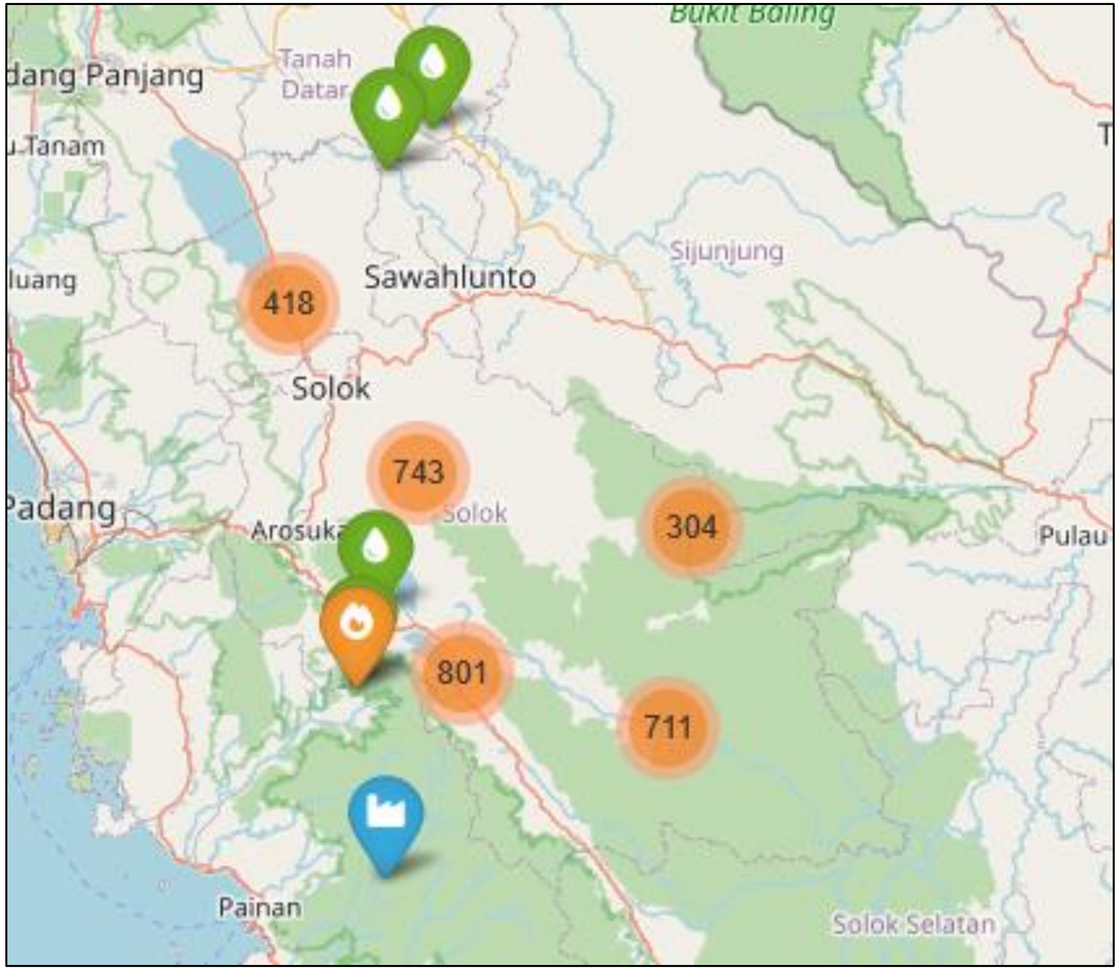
How sparse, edge, or outlying points (in NDVI–LST–VV space) are flagged as anomalies.

These red outliers are the system's prediction of potential geothermal anomaly zones.

What Can be Concluded :

- **Consistent Anomalous Cluster Shape**:
  The anomalies form a discernible boundary region on the edge or margin of the data cloud.
- **Physically plausible anomaly zones**:
  Most red/yellow points tend to cluster in **lower NDVI**, **higher LST**, and **stronger VV deformation** — aligning with typical geothermal surface expressions (sparse vegetation, hot surface, shifting land).
- **Use both plots synergistically**:
  - The **first plot** gives **confidence scores** (ranking).
  - The **second plot** gives a **clear threshold decision** (for binary classification or mapping).

Anomaly maps + known geothermal sites



This map is a **visualization of anomaly points detected by the Isolation Forest algorithm**, overlaid on a geographic map of the Solok region, with added **known geothermal sites** (manifestations and facility)

💧**Geothermal Icon (green drop):**

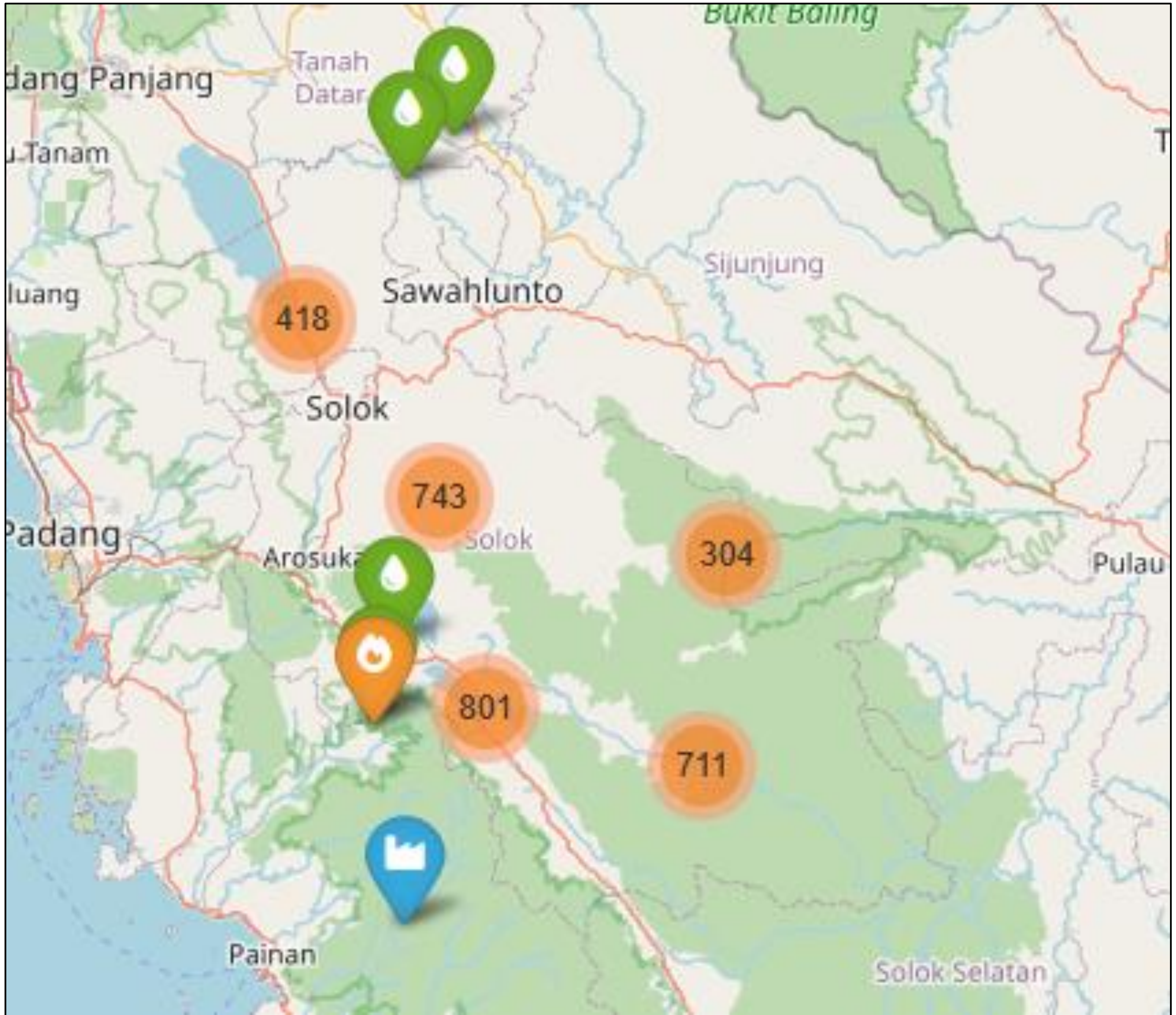
•This marks a **known geothermal manifestation site**—in this case, **Alahan Panjang Hot Spring**.

•Used for spatial comparison between **detected anomalies** and **verified geothermal activity**.

Building icon : Industrial geothermal facility (Muara Laboh)

It serves as a **reference point** to visually compare **anomaly detection results** (e.g., red dots) with real-world known geothermal operations.

Anomaly maps + known geothermal sites

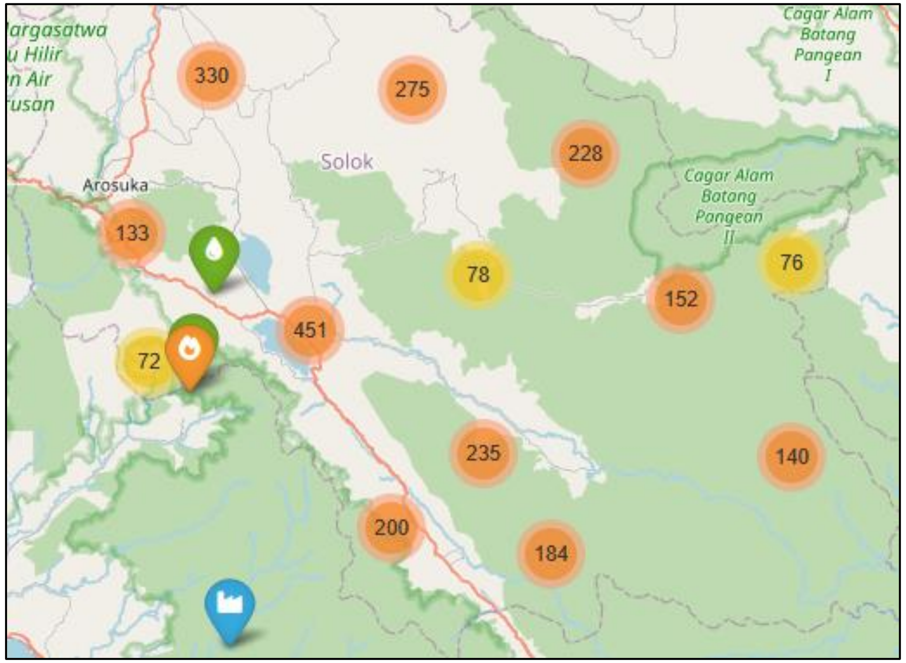


Each orange circle with a number is a cluster of anomaly points grouped by Folium's MarkerCluster plugin.

For example, "801" means there are 801 anomalies clustered in that region.

When zoomed in, they will spread into individual points (red dots or smaller groups).

When zoomed in



Green Circles

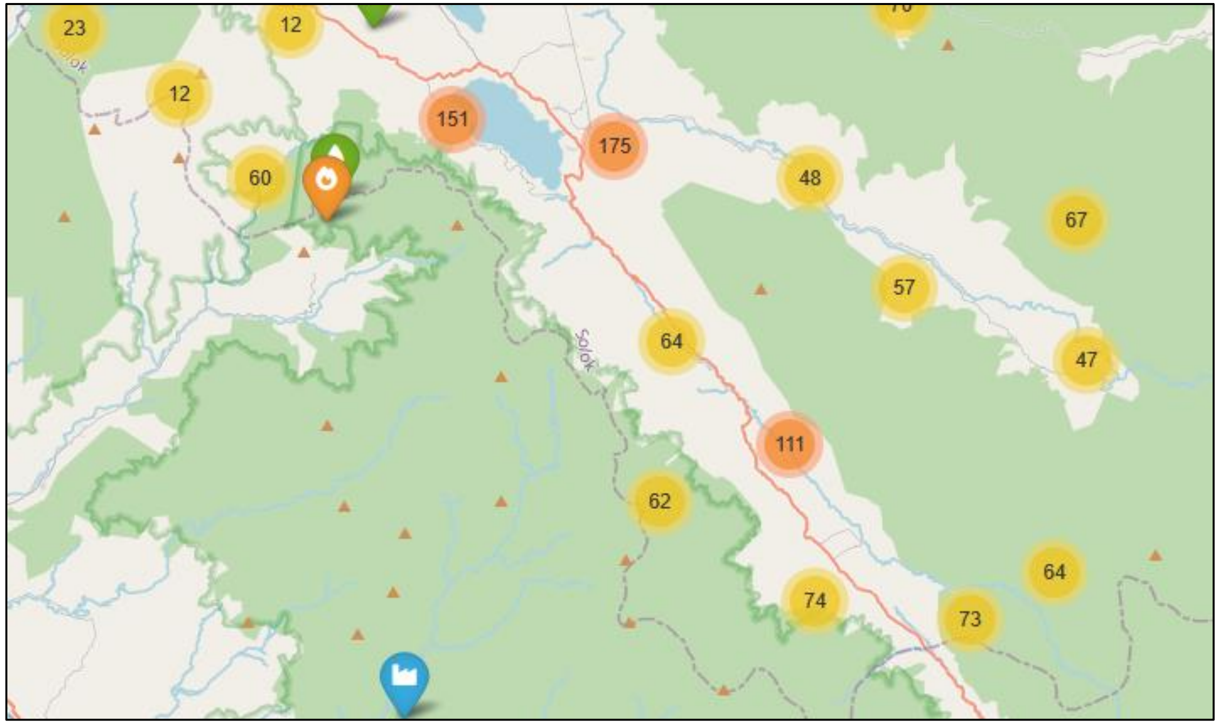
- **Meaning**: These are **low-density clusters (<10)**

Yellow Circles

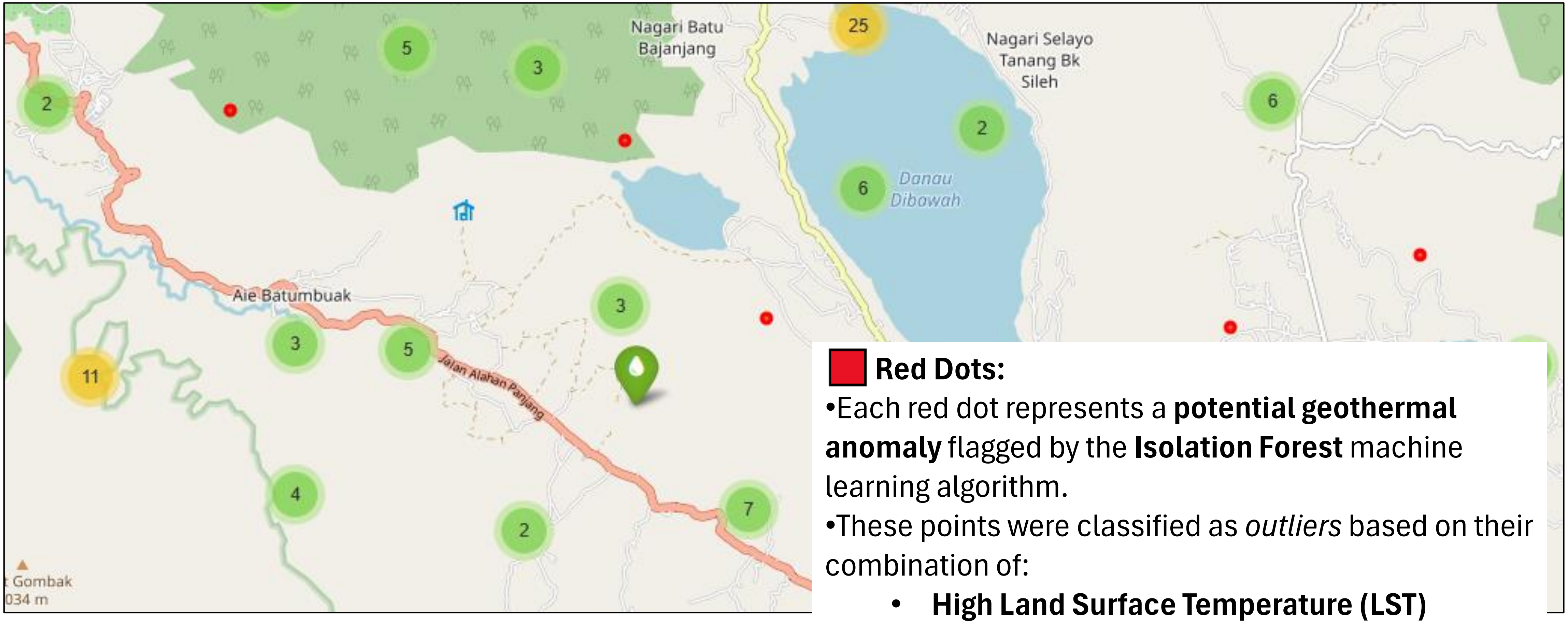
- **Meaning**: These are **moderate-density clusters (10-30)**

Orange Circles

- **Meaning**: These are **high-density clusters (>30)**



Further zoomed in



Red Dots:

•Each red dot represents a **potential geothermal anomaly** flagged by the **Isolation Forest** machine learning algorithm.

•These points were classified as *outliers* based on their combination of:

- **High Land Surface Temperature (LST)**
- **Low NDVI** (sparse vegetation)
- **Anomalous InSAR VV shift** (ground deformation)

Key observations from iForest anomaly map

1. Spatial Overlap with Known Geothermal Sites

1. The **largest cluster (801 anomalies)** is geographically close to the **Muara Laboh Geothermal site**, as well as the **Gunung Talang fumarole** and **Lembah Gumanti hot springs**.
2. This **supports the reliability of the model** in flagging geothermal-related anomalies based on satellite indicators (NDVI, LST, InSAR).

2. Anomalies in Unexplored Areas

1. Anomaly clusters near **Solok**, **Sawahlunto**, and **Padang Panjang** do not correspond to known geothermal manifestations.
2. These may represent **undiscovered or under-studied geothermal zones** worth prioritizing in future exploration campaigns.

3. Consistency Across Features

1. The map suggests a **correlation between dense anomaly zones** and geothermal activity indicators such as:
    1. Low NDVI (sparse vegetation)
    2. High LST (surface temperature)
    3. Surface deformation (from InSAR VV)

📌 Implications

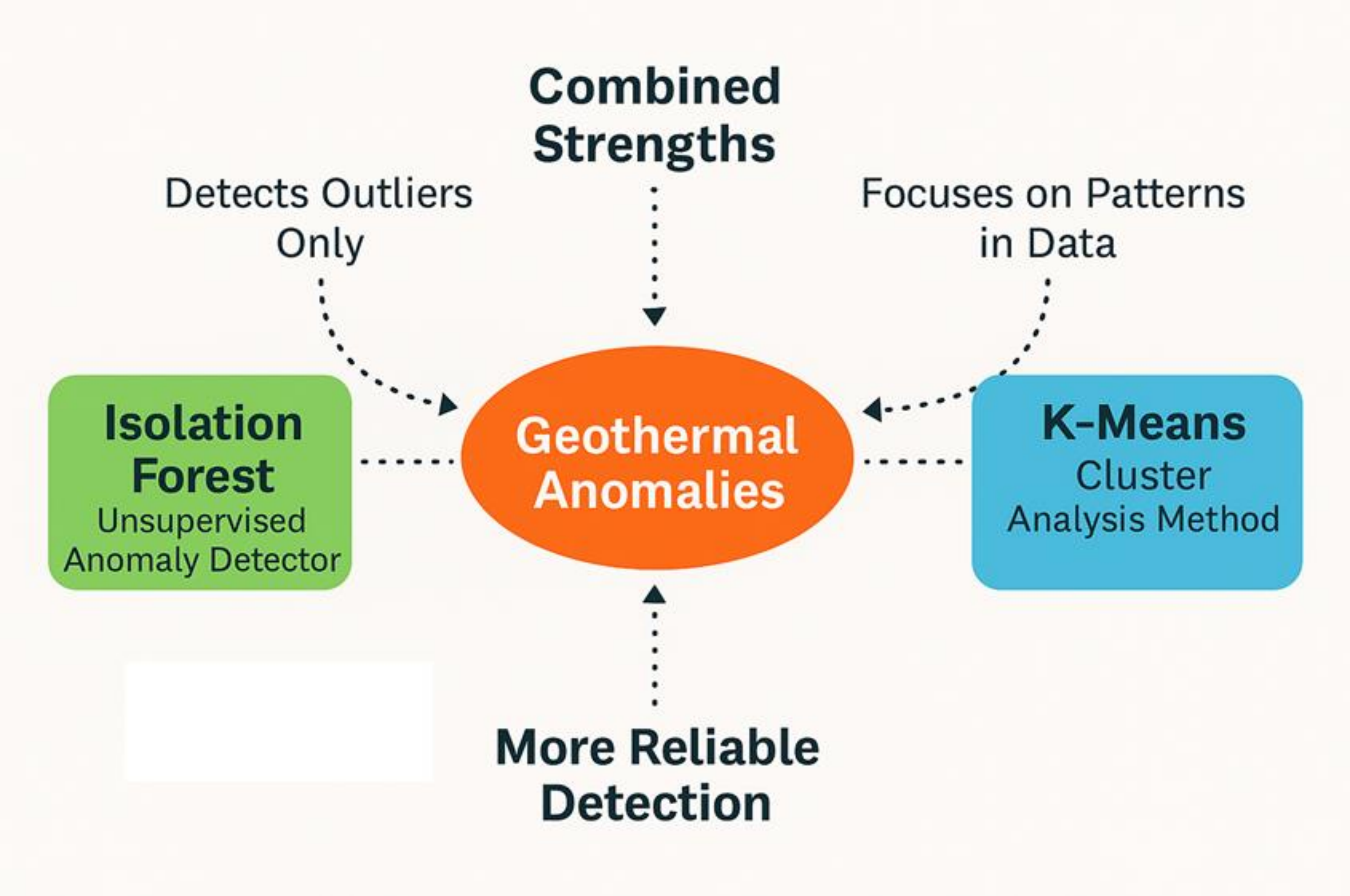
- **Validation**: The model appears effective in locating geothermal anomalies in known regions.
- **Exploration Potential**: High-anomaly zones with **no current geothermal sites** offer candidates for **ground-truthing or pre-drilling surveys**.
- **Scalability**: This method is scalable to other remote volcanic terrains in Indonesia where **field data is scarce**.

.... But is it enough?? NO (see next slide)

Why iForest alone is not enough?

While iForest is **excellent at spotting potential anomalies**, using it **alone** risks:

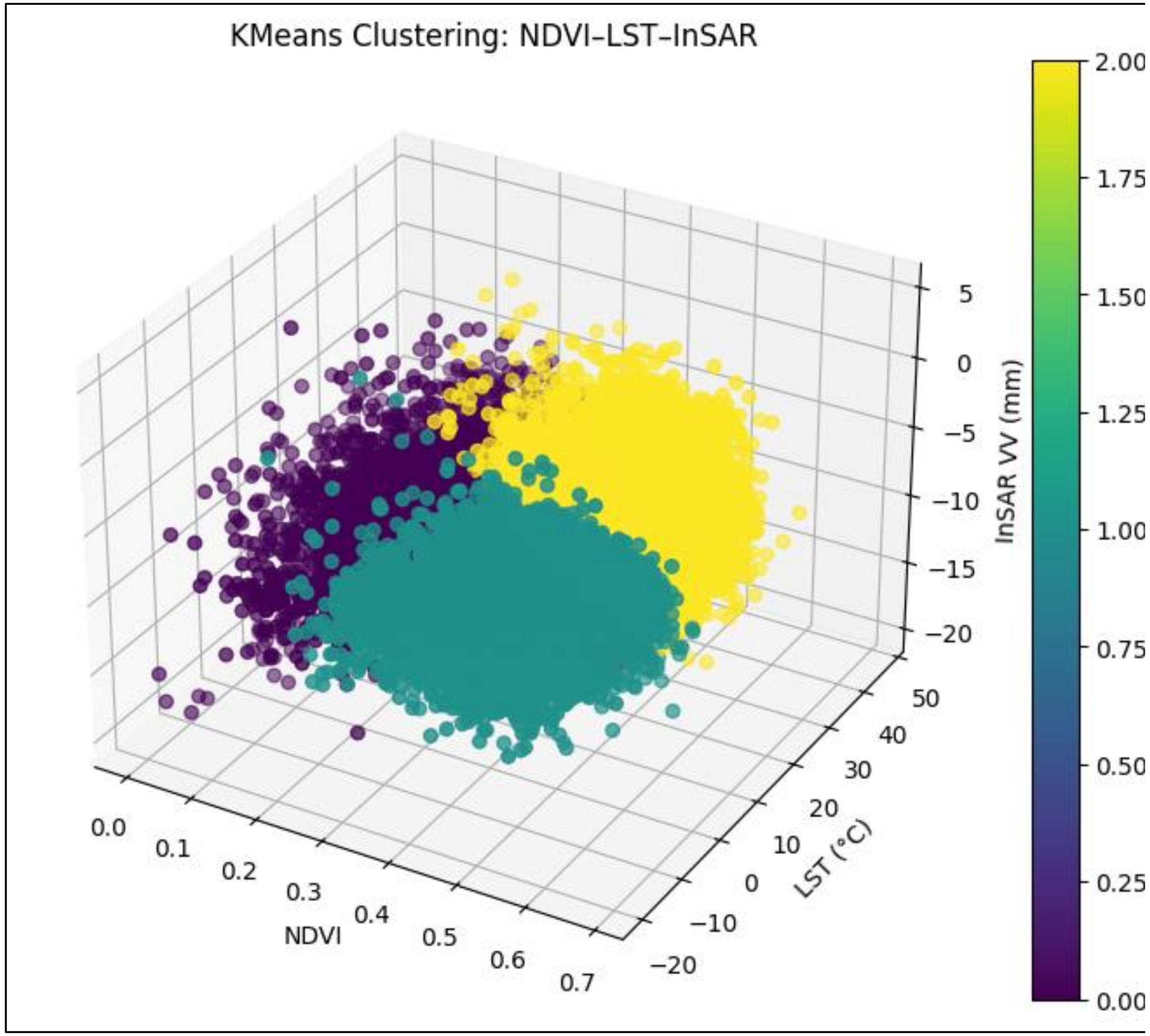
- Missing subtle geothermal patterns,
- Misidentifying irrelevant outliers,
- And lacking interpretability.
- 👉 **Best practice**: Use **iForest + clustering (e.g. Kmeans) + domain filters** to improve confidence and precision in geothermal exploration.


Combined Strengths
Detects Outliers Only
Focuses on Patterns in Data
Isolation Forest
Unsupervised Anomaly Detector
Geothermal Anomalies
K-Means
Cluster Analysis Method
More Reliable Detection

Results

Kmean Model

Kmeans Clustering



What it shows:

•Each point in the 3D space represents a data sample characterized by:

- **NDVI (x-axis):** Vegetation health
- **LST (y-axis):** Land surface temperature
- **InSAR VV (z-axis):** Vertical displacement

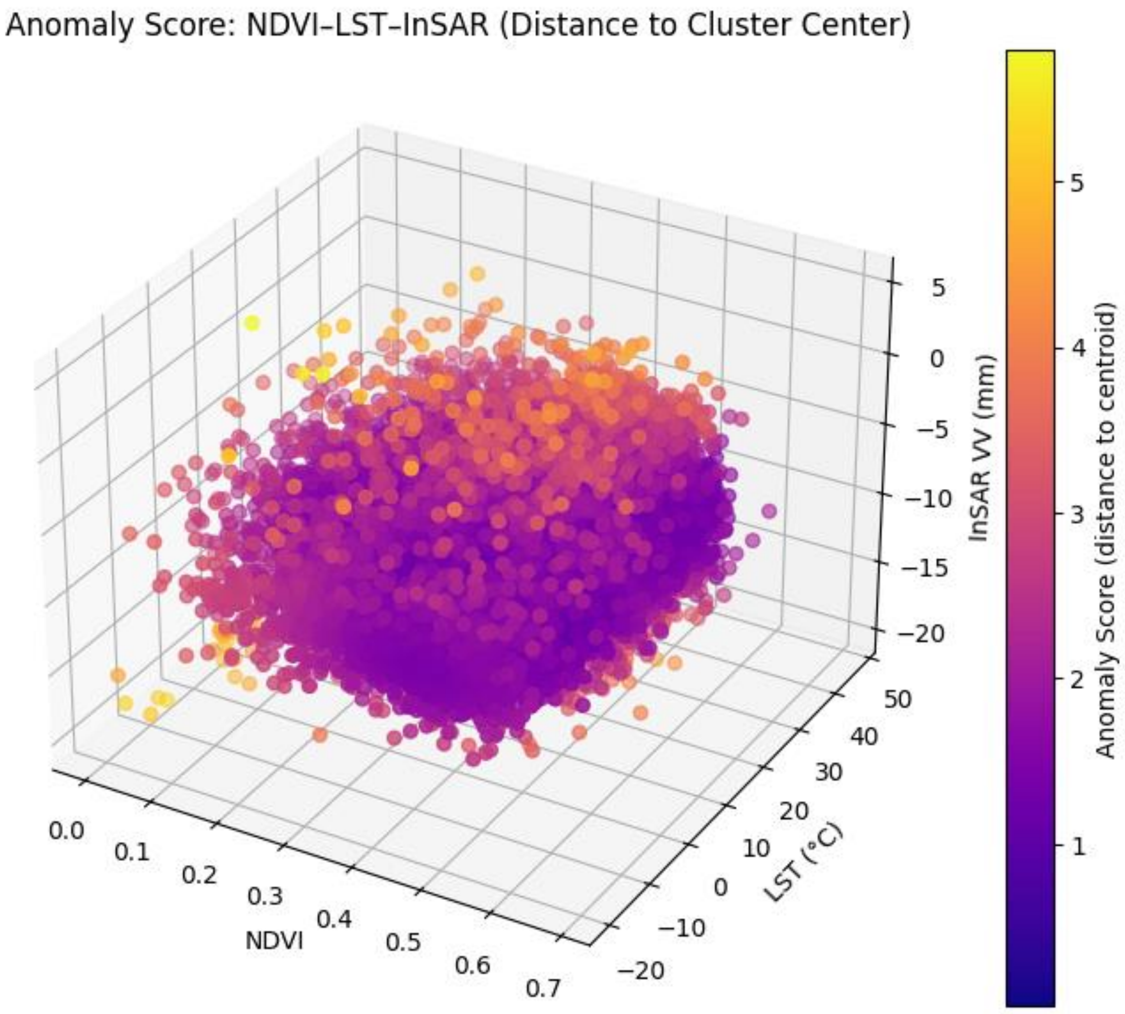
What to observe:

•**Three distinct clusters** are visually separable and colored differently.

•**Cluster interpretation**:

- **Cluster 0 (purple):** Low NDVI, moderate LST, and most negative VV → indicates **potential geothermal anomalies**.
- **Cluster 1 (teal):** high NDVI, lower LST, moderate VV → likely background terrain.
- **Cluster 2 (yellow):** moderate to high NDVI, high LST, less ground displacement → possibly vegetated areas affected by sun exposure or agriculture.

Anomaly Score



What it shows:

•**Color-coded anomaly scores** based on each point's distance from its cluster centroid.

•Warmer colors (yellow, orange) = **higher anomaly score**

•Cooler colors (purple) = **points close to cluster centers**, i.e., typical or normal.

What to observe:

•The high-scoring points appear more scattered and isolated on the outer edges of clusters.

•**Outliers in all three features** (NDVI, LST, VV) are identified here.

•These points represent **spatial-spectral anomalies** that might reflect:

- Heat spots with low vegetation and high displacement.
- Unexpected thermal uplift with stable surface.
- Any sensor-based or process-based data abnormalities.

Combined Interpretation for Geothermal Targeting

Cluster 0 (purple in first plot) + high anomaly score points (yellow in second lot) are prime geothermal exploration targets.

These areas show:

- Sparse vegetation (low NDVI)
- Elevated temperature (LST)
- Notable ground deformation (InSAR VV)

Such a combination of physical indicators aligns with known geothermal surface expressions like fumaroles, hot springs, and uplifted zones.

Anomaly cluster analysis

```
---------------- Group by cluster and describe key indicators
              NDVI        LST        VV
cluster
2         0.485703  20.613280 -7.160132
0         0.284928  14.811353 -8.914127
1         0.472786  -1.395935 -8.035957
```

Feature	Ideal Anomaly Indicator	Cluster 0	Cluster 2	Cluster 1
NDVI	Low (vegetation stress)	✅ Lowest (0.285)	Moderate (0.486)	Moderate (0.473)
LST (°C)	High (surface thermal anomaly)	Moderate (14.81)	✅ Highest (20.61)	❌ Extremely Low (-1.40)
VV (InSAR)	Strong deformation (more negative)	✅ Strongest (-8.91)	Less deformation (-7.16)	Moderate (-8.04)

From result above we can see that cluster 0 **is the best for geothermal anomaly candidate**.

- Lowest NDVI (0.285) → sparse vegetation
- Moderate LST (14.81 °C)
- Strongest InSAR shift (-8.91 mm)

Conclusion : These result indicates low vegetation, strong ground movement, and elevated temperature. **This is the best geothermal anomaly candidate.**

But… we need to confirm this also by anomaly score analysis (next slide)

Anomaly score analysis

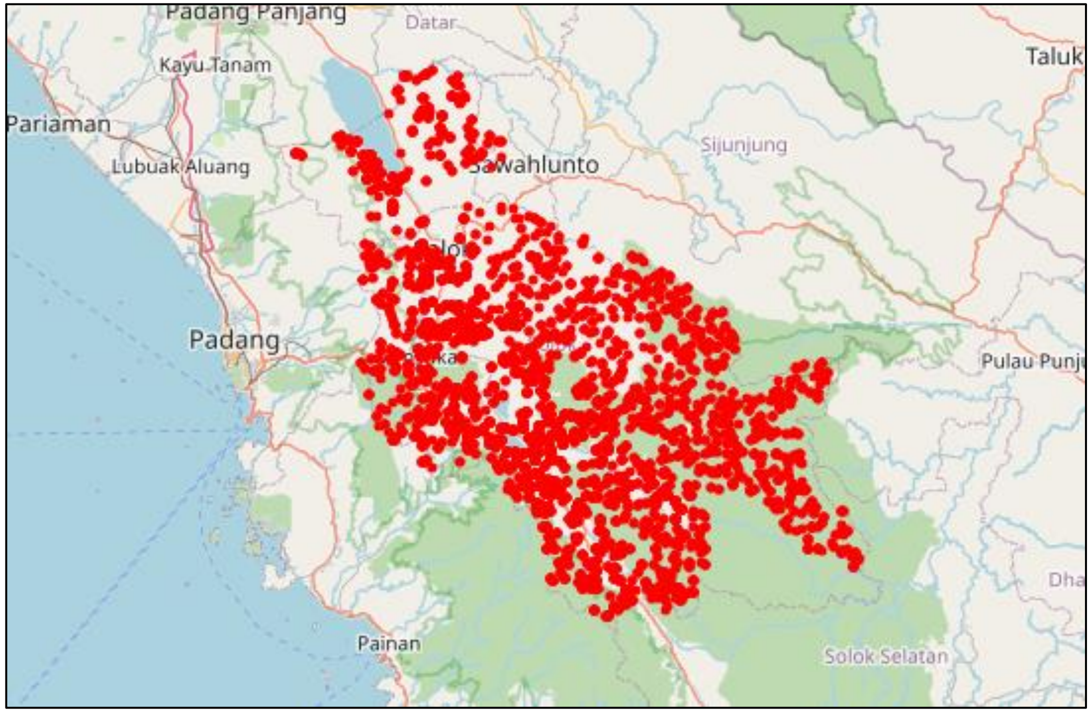
cluster	mean	max	std
0	1.381221	5.794033	0.698196
2	1.134353	4.871692	0.614688
1	1.100095	5.594797	0.509490

Cluster	Mean Anomaly Score	Max Score	Std Dev	Insight
0	**1.38**	**5.79**	0.70	Most anomalous cluster overall — high average + highest peak. Likely to contain geothermal-related outliers.
2	1.13	4.87	0.61	Moderate anomalies. Worth investigating, but less extreme.
1	1.10	5.59	0.51	Lowest anomaly average. Likely to represent "normal" baseline conditions.

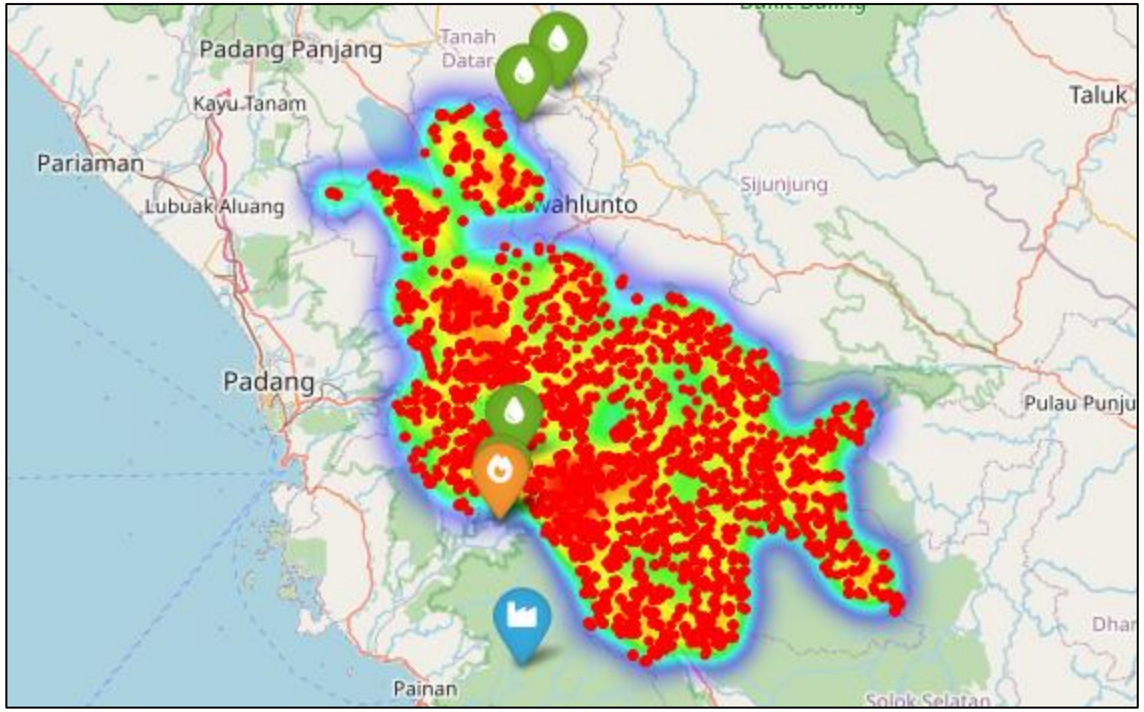
Conclusion:

Cluster 0 is the **most anomalous**, making it a **top candidate for geothermal anomaly sites**.

Map of Anomaly Cluster “0” (The best anomaly candidate)



Cluster “zero anomalies



Cluster “zero” anomalies with heatmap and known geothermal manifestations overlays

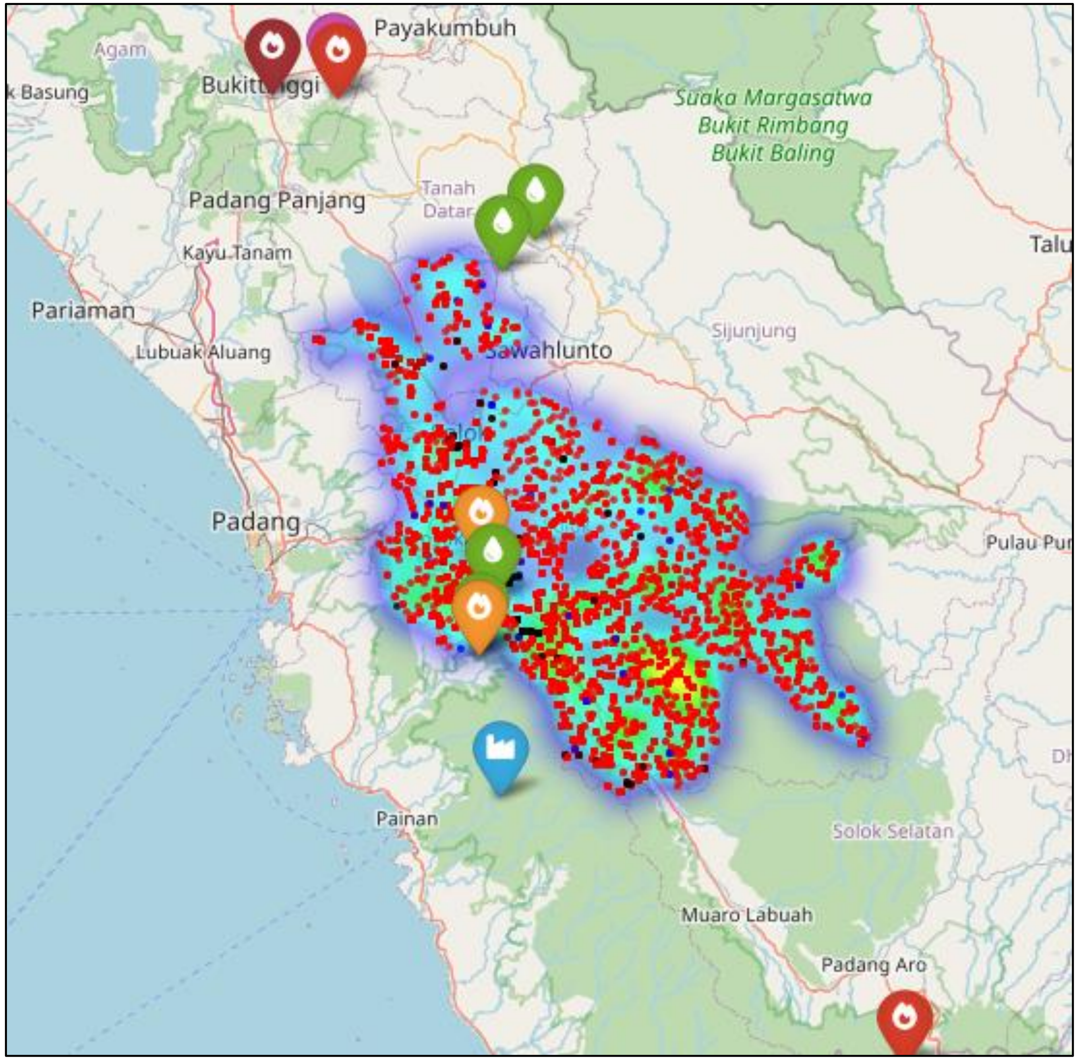
Results

Combined ML Models (Kmean & Isoforest)

Combined ML Models Overlay

Map below is an overlay map of geothermal anomalies using KMeans & Isolation Forest ML algorithms + Anomaly density heatmap + Geothermal Markers (Surface Manifestations and Volcanos).

Click the link to view interactive map :



Red markers: Locations identified as anomalous by KMeans only (Cluster 0).

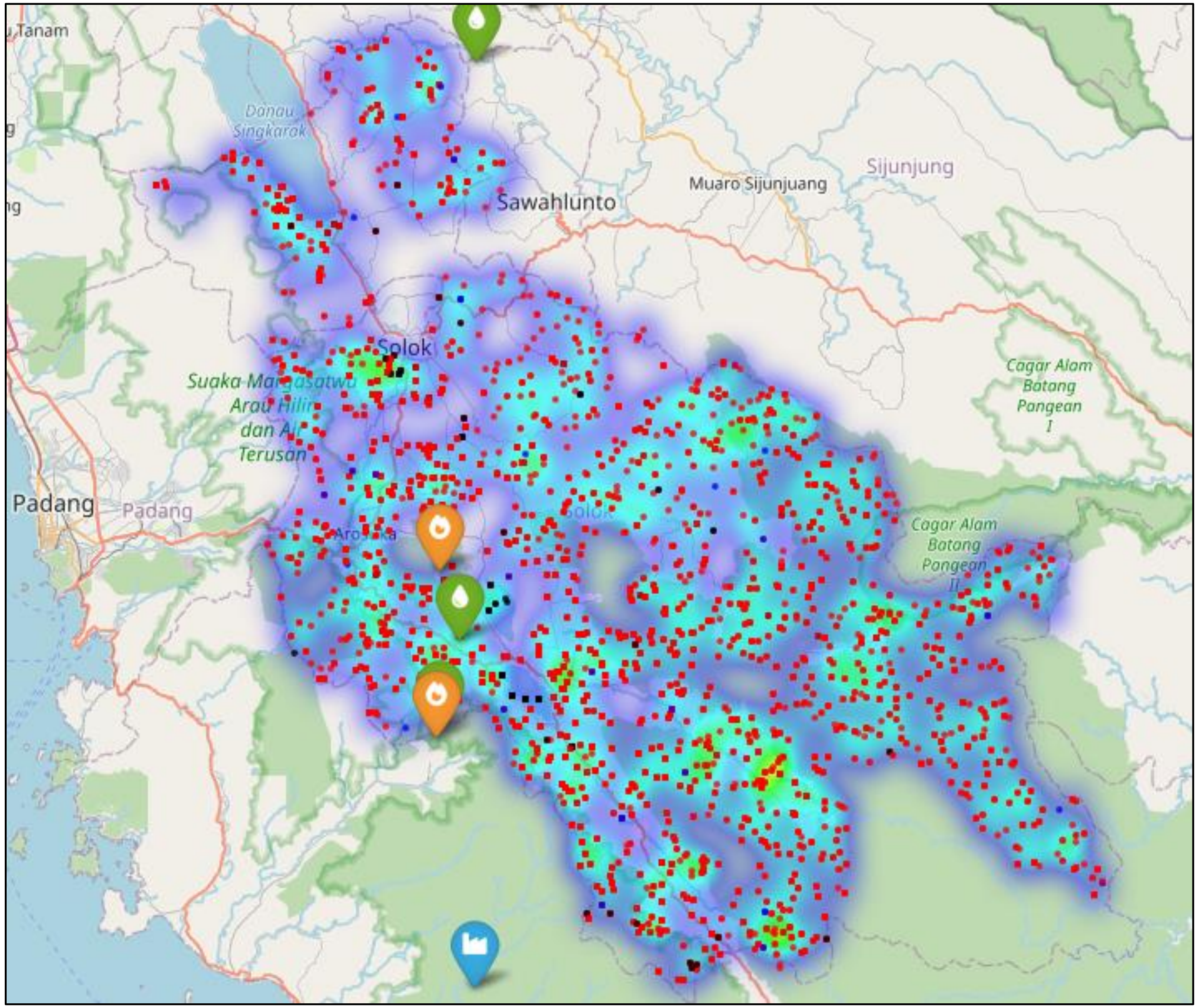
Blue markers: Locations flagged as outliers by Isolation Forest only.

Black markers: Where both methods agree—these are your most statistically significant candidates.

Heatmap : density of geospatial anomalies (red:very high, yellow:high, green:moderate, blue: low)

✅ **Insight**: *The black dots are high-confidence geothermal anomaly zones*

Heatmap Intensity



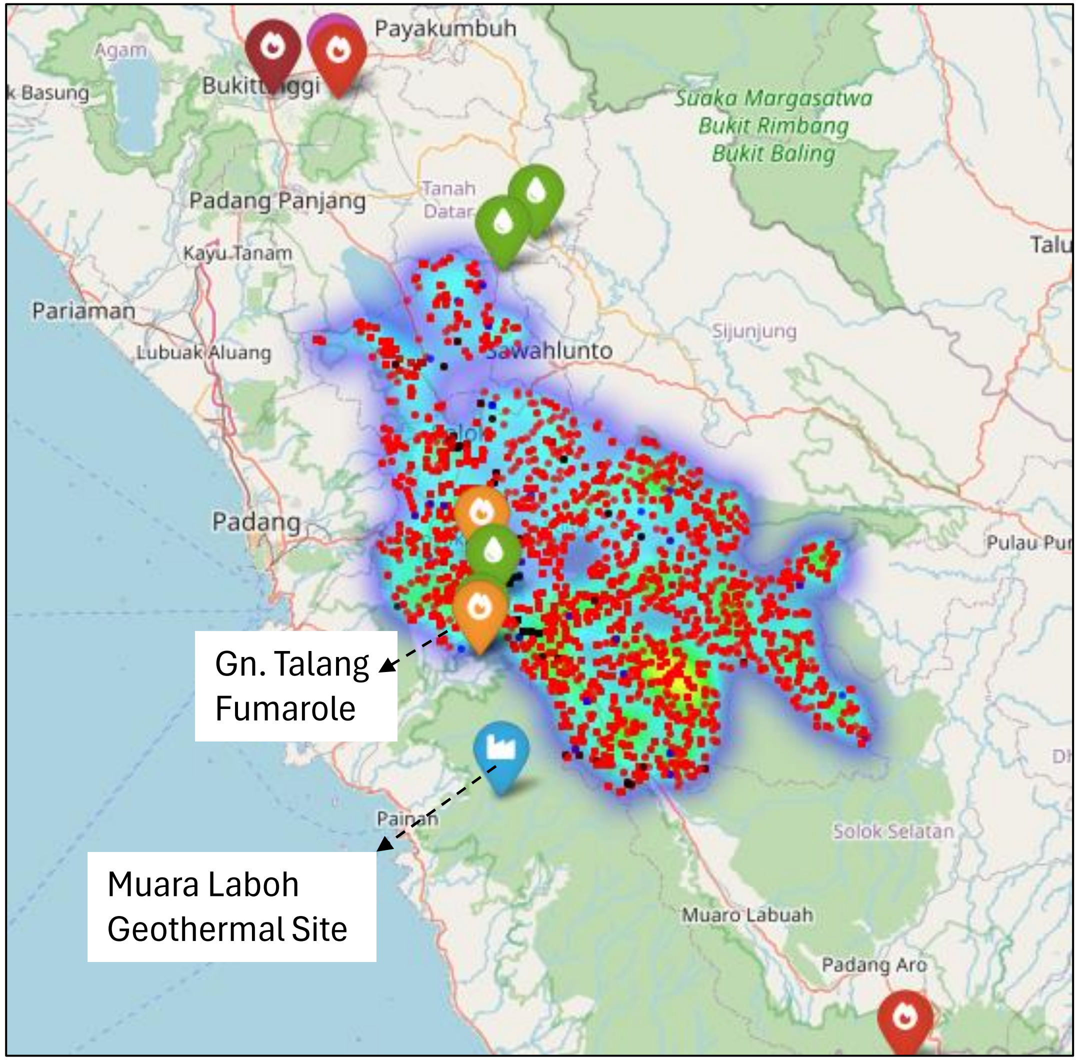
Heatmap Intensity

•The **heatmap layer** shows a gradient of point density and geospatial clustering.

•The **hottest areas** (brightest zones e.g. red,yellow,green) likely indicate:

- Dense surface feature anomalies (from NDVI, LST, VV).
- Spatially consistent geophysical variation.
- Possibly anthropogenic or vegetative effects—but co-validation with geological context (e.g., volcanoes) improves reliability.
- ✅ **Insight**: A heat-concentrated zone that also has black (both-model) anomalies suggests **strong geothermal potential**.

Geothermal & Volcanic Site Correlation

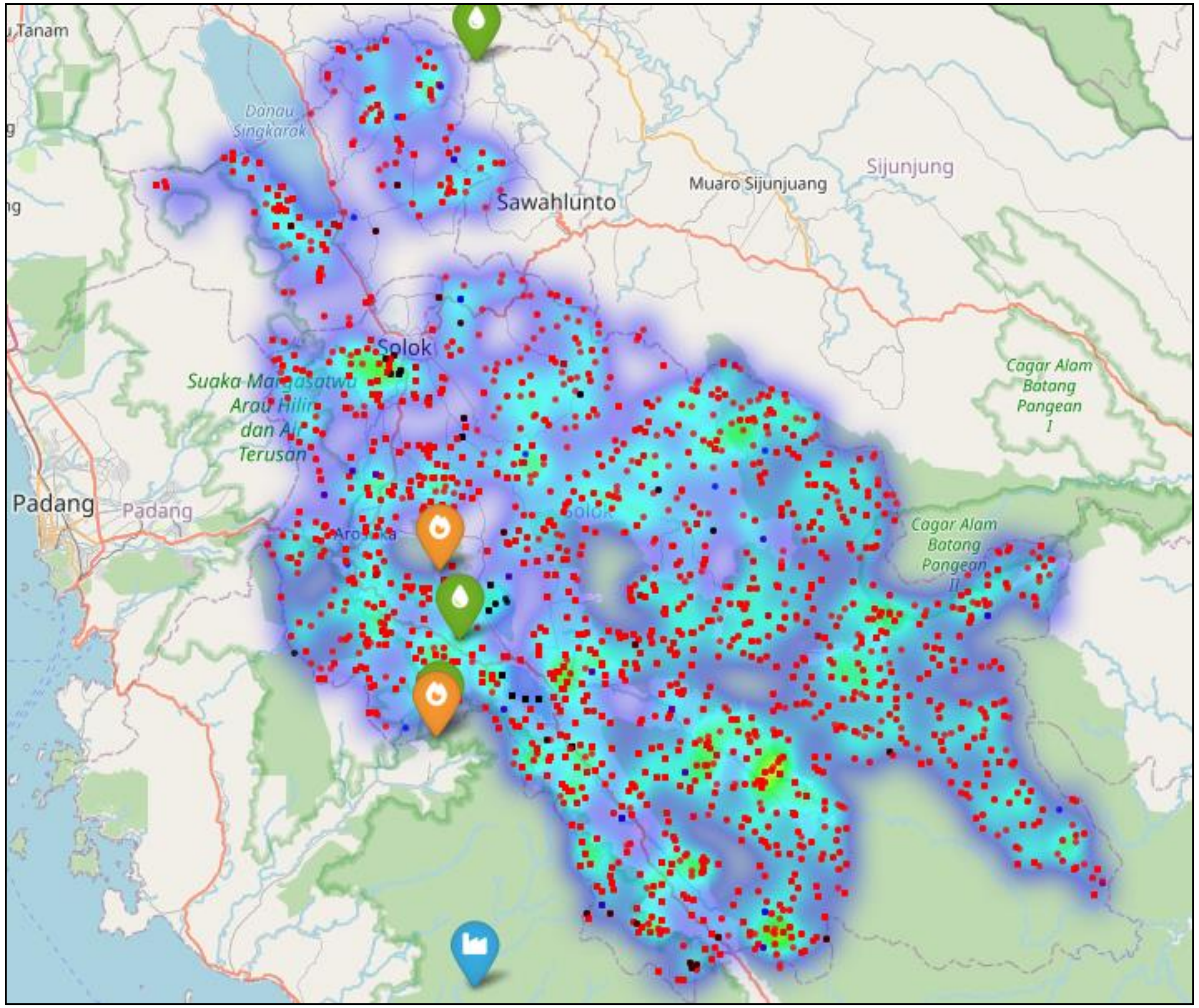


The Muara Laboh and Gunung Talang fumarole field appear close to multiple anomaly markers (especially black).

Padang Ganting and Alahan Panjang hot springs are also near both-model consensus points.

✅ **Insight**: The anomaly model is validated by its proximity to known geothermal sites. This confirms that the feature selection (NDVI, LST, InSAR/VV) and ML pipeline are functioning effectively.

Isolated Blue/Red Anomalies



Some blue markers (iForest only) exist outside the main geothermal areas. These may be:

Statistical outliers unrelated to geothermal activity.

Or emerging targets not yet explored geologically.

🔍 **Recommendation**: These areas warrant cautious review—look at their elevation, fault lines, and vegetation patterns.

High-priority geothermal anomaly coordinates

High-priority geothermal anomaly coordinates ARE locations where both KMeans and Isolation Forest models agree there is an anomaly. These points are statistically significant and spatially distinct based on NDVI, LST, and InSAR (VV) features. This coordinates can be used as prime candidates for further geothermal investigation or field validation.

From this model, There are **56 high-priority anomaly coordinates** — locations where both **KMeans (Cluster 0)** and **Isolation Forest** models agree there's an anomaly.

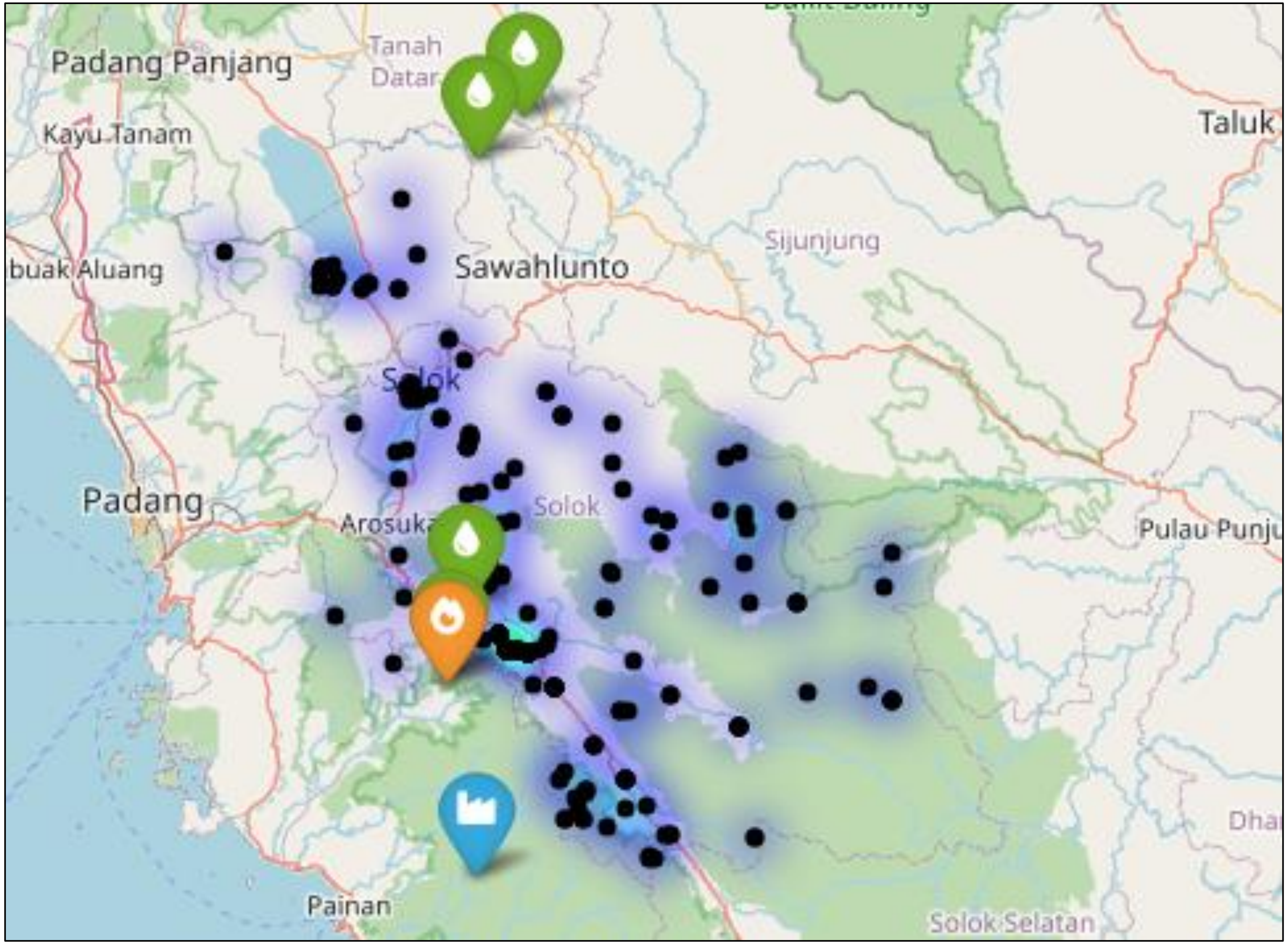
Highest-priority geothermal anomaly coordinates

Top 10 highest-ranked geothermal anomaly sites based on a composite anomaly score (favoring high LST, low NDVI, and strong deformation/low VV):

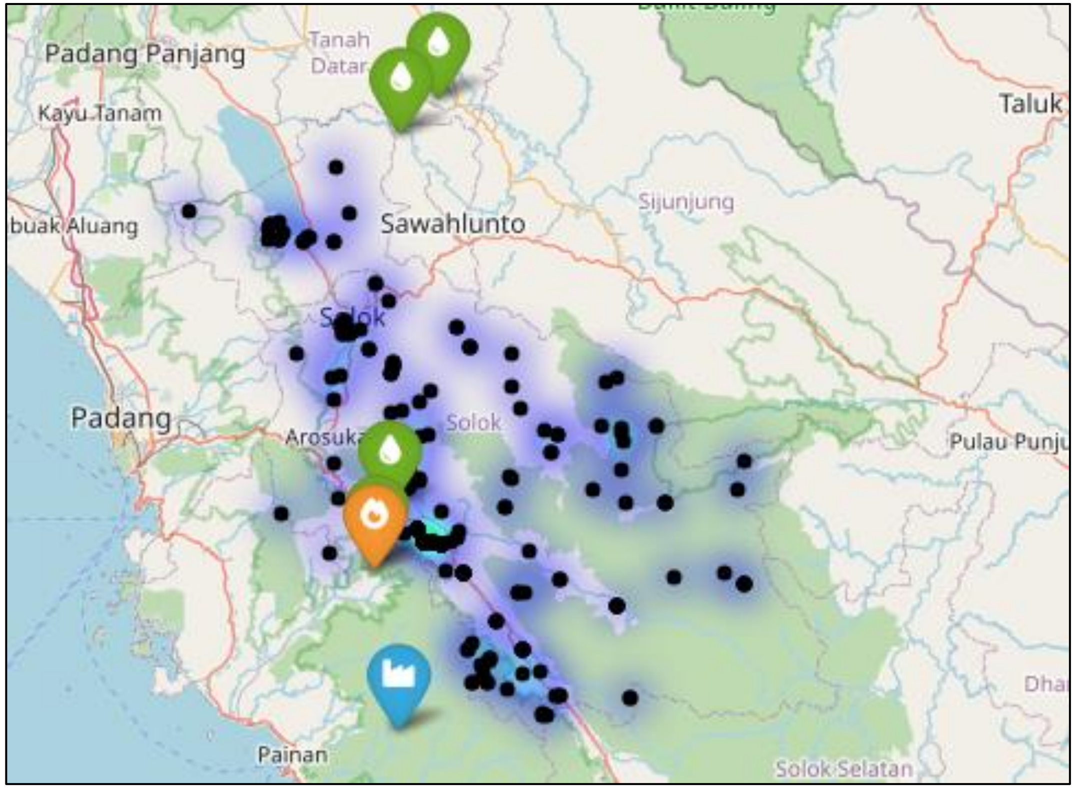
#	Latitude	Longitude	NDVI	LST	VV
0	-1.00763	100.72942	0.331	30.746	-21.147
1	-1.00763	100.72447	0.351	30.326	-20.947
2	-1.03521	100.71594	0.392	30.631	-20.011
3	-1.01097	100.71594	0.371	30.541	-20.161
4	-1.02854	100.71959	0.371	30.512	-20.159
5	-1.03782	100.73015	0.360	30.407	-20.060
6	-1.03521	100.73015	0.367	30.383	-19.930
7	-1.01471	100.71326	0.375	30.245	-20.042
8	-1.02388	100.72198	0.359	30.228	-20.107
9	-1.02388	100.71326	0.353	30.183	-20.191

These are the **highest-priority candidates** for geothermal exploration based on remote sensing anomalies.

Map of highest-priority geothermal anomaly coord. (black dots) (56 anomaly points)



Map Analysis (1)



1. Black Dots = High-Priority Anomalies

These are sites that satisfy both:

•**Cluster 0** in KMeans → indicative of geothermal patterns (e.g., elevated LST, lower NDVI, active deformation)

•**Flagged by Isolation Forest** → statistically rare/outlier in multivariate feature space.

This dual filtering increases confidence that these are **true geothermal anomaly candidates**.

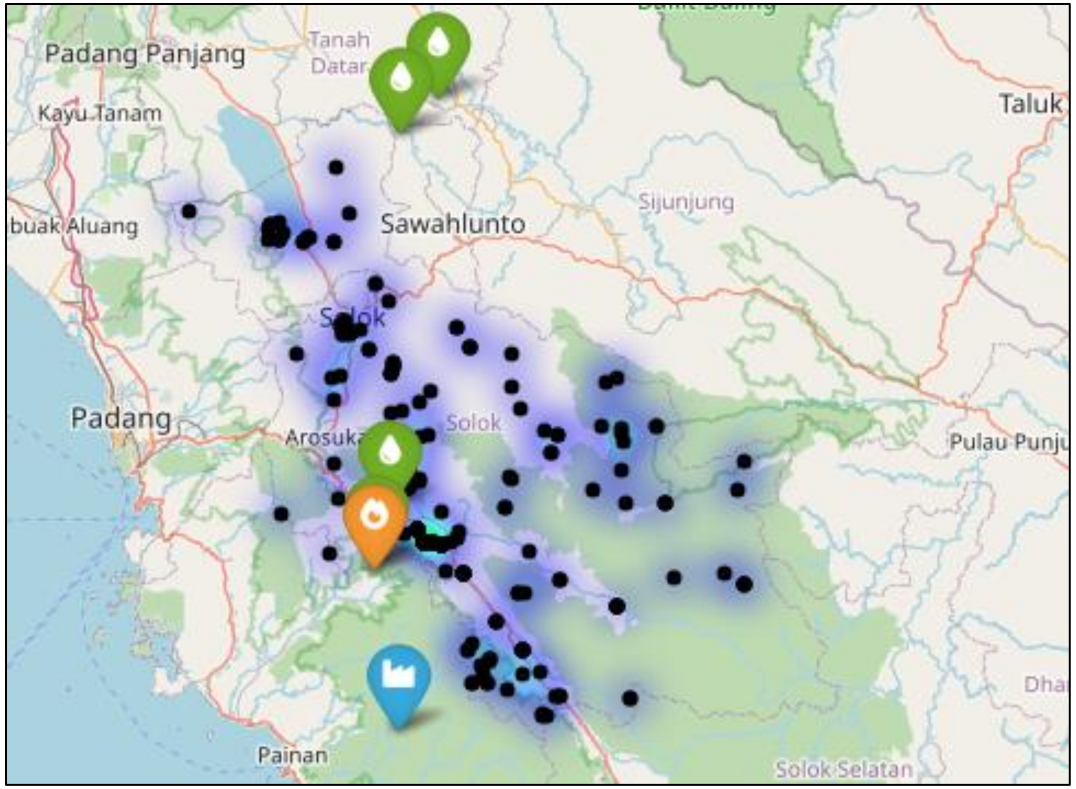
2. Spatial Distribution

•Anomalies cluster densely **south of Mount Talang** and **along the ridge line** toward **Lembah Gumanti** and **Muara Laboh**.

•Scattered anomalies extend to **eastern Solok**, suggesting the geothermal system may stretch beyond previously known sites.

……. ***cont next slides***

Map Analysis (2)



3. Correlation with Known Manifestations

•Several anomalies **overlap or lie close to**:

- Muara Laboh geothermal field
- Lembah Gumanti hot spring
- Gunung Talang fumarole

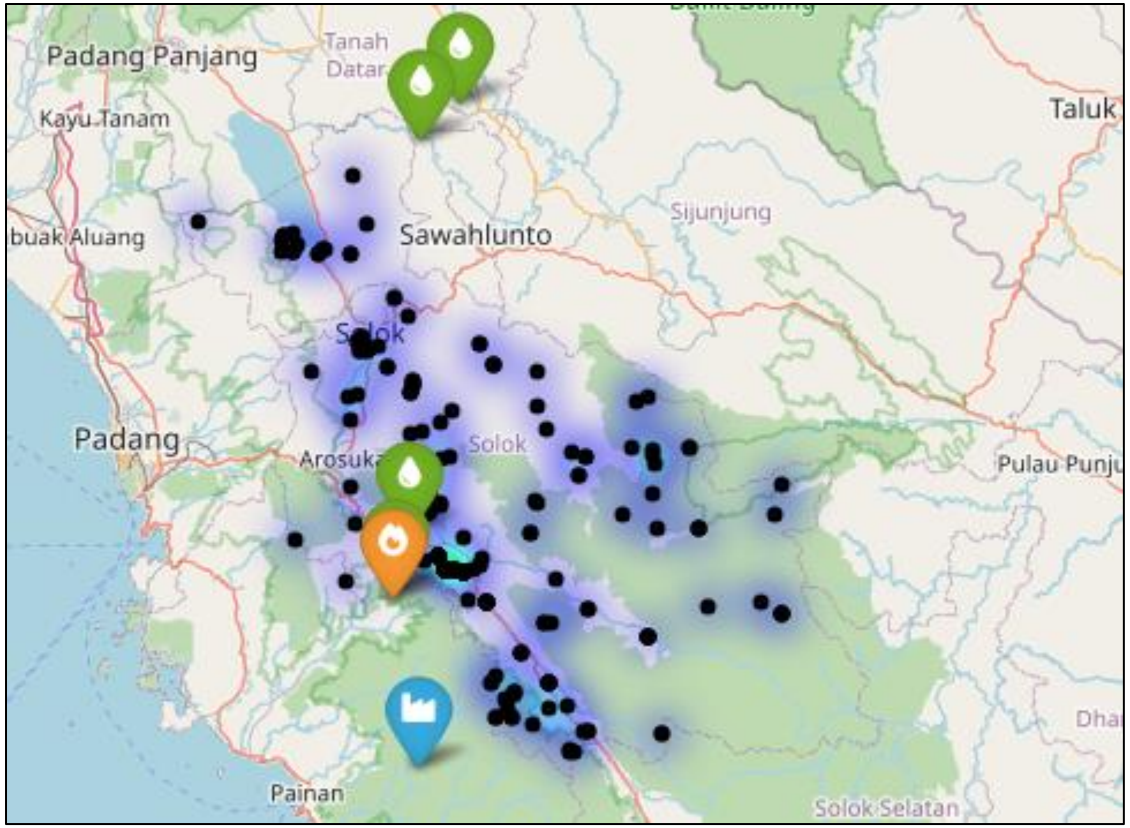
•This spatial agreement **validates the unsupervised model output** using real-world references.

4. Heatmap vs. Black Dots

•The **heatmap** shows survey density—not geothermal potential.

•Black dots are more **discrete and targeted**, helping **prioritize field verification** and reduce exploration costs.

Map Analysis (3)



Implications for Geothermal Prospecting :

•These black-dot sites are **prime targets** for further geophysical or geochemical validation.

•The combined ML-driven anomaly detection allows for:

- **Cost-efficient pre-screening**
- **Wide area coverage**
- **Bias-free detection** (data-driven, unsupervised)

Summary of Strength

Element	Contribution
NDVI, LST, VV features	Surface vegetation, thermal, and deformation indicators
KMeans	Captures cluster-based geothermal "signatures"
Isolation Forest	Flags statistically rare behavior in feature space
Overlay method	Visualizes agreement/disagreement between models
Heatmap	Adds intuitive intensity reading
Geothermal site markers	Validate model accuracy and interpretability

Challenges Faced and Learnings

1. Challenge: Integrating data with different spatial and temporal resolutions (e.g., MODIS vs Sentinel) . ***Solution***: Used monthly temporal aggregation and spatial grid averaging

2. Challenge: InSAR noise and processing complexity. ***Solution***: Averaged deformation over time to stabilize signal

3. ***Challenge***: No labeled geothermal anomalies for supervised training. ***Solution***: Used unsupervised algorithms and domain knowledge for anomaly interpretation

Learning: Preprocessing and data curation were more critical than model complexity; meaningful patterns depend on clean, well-aligned inputs.

Suggestions for Future Improvements (1)

1. Field Verification & Supervised Learning

Collect field-verified geothermal anomaly data (e.g., temperature probes, fumaroles, hot springs) to enable supervised classification models (e.g., Random Forest, SVM). Use this labeled data to train and validate a predictive model, improving accuracy and real-world applicability.

2. Incorporate Terrain and Surface Features

Integrate elevation, slope, aspect, and topographic shading (DEM) into the model. These features can help detect surface deformation patterns or geothermal gradients associated with mountainous regions.

Suggestions for Future Improvements (2)

3. Temporal Dynamics & Time-Series Monitoring

Extend analysis with monthly NDVI, LST, and VV time series to capture seasonal patterns, persistent anomalies, or transient events. Consider trend-based anomaly detection for more dynamic geothermal activity insights.

4. Higher Spatial Resolution & Advanced Sensing

Fuse with higher-resolution satellite data (e.g., ASTER, commercial imagery) or drone-based thermal mapping to enhance spatial granularity. This would allow zoomed-in prioritization for field deployment.

References

1. **Cumming, W. (2009).**
   Cumming, W. (2009). *Geothermal resource conceptual models using surface exploration data*. Proceedings of the 34th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California.
   https://pangea.stanford.edu/ERE/db/IGAstandard/record_detail.php?id=7119

2. **Fialko, Y. (2004).**
   Fialko, Y. (2004). Probing the mechanical properties of seismically active crust with InSAR. *Nature, 428*(6981), 31–36. https://doi.org/10.1038/nature03197

3. **Bignall, G., & Browne, P. R. L. (2000).**
   Bignall, G., & Browne, P. R. L. (2000). *Geology of the Rotokawa geothermal field, Taupo Volcanic Zone, New Zealand*. Proceedings of the World Geothermal Congress, 2000, Kyushu–Tohoku, Japan, 781–786.
   https://www.geothermal-energy.org/pdf/IGAstandard/WGC/2000/R0741.PDF

4. **Wan, Z. (2006).**
   Wan, Z. (2006). *MODIS Land Surface Temperature Products User's Guide (Version 1.0)*. University of California, Santa Barbara, Institute for Computational Earth System Science (ICESS).
   https://modis.gsfc.nasa.gov/data/dataprod/mod11.php

5. **European Space Agency (ESA). (2022).**
   European Space Agency. (2022). *Sentinel-1 SAR User Guide*. Retrieved from https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar

6. **Jensen, J. R. (2007).**
   Jensen, J. R. (2007). *Remote sensing of the environment: An earth resource perspective* (2nd ed.). Upper Saddle River, NJ: Pearson Prentice Hall.