SOAP

Swiss Airspace Obstacle Profiling

Spatial Clustering for Low-Altitude UAS Corridor Planning

Technical Report

Roman Aebi

7. Mai 2026

1 Introduction

Switzerland's airspace below 150 m AGL contains over 14,000 registered obstacles: transmission lines crossing valleys, cable cars spanning alpine terrain, industrial chimneys in urban zones, and antenna masts near airfields. For any organisation planning low-altitude UAS (drone) corridors, the question is not whether obstacles exist, but how they are spatially distributed and whether distinct risk zones can be identified from their characteristics alone.

This report presents SOAP, a geospatial data science pipeline that answers this question. Starting from raw government data published by the Federal Office of Civil Aviation (BAZL), the pipeline transforms 14,334 point-level obstacle records into a hexagonal risk map of Switzerland with four operationally distinct zones. The methodology is fully reproducible: all data sources are publicly available under the Swiss Open Government Data licence (Opendata BY), and each step from API download to interactive map is captured in executable code.

The report follows the pipeline chronologically: data acquisition and quality assessment (Section 2), spatial aggregation and feature design (Section 3), the search for meaningful cluster structure (Section 4), validation of the result (Section 5), and translation into an operational risk classification (Section 6).

2 Data Acquisition and Quality

2.1 From API to Parquet

The primary dataset is the collection ch.bazl.luftfahrthindernis, accessed via the STAC API (v0.9) on data.geo.admin.ch. The ingestion script queries the API to resolve the current KMZ download URL, with a hardcoded fallback in case the API structure changes. The KMZ (a zipped KML archive) is extracted and parsed with ElementTree. Each Placemark yields an obstacle record with coordinates and attributes.

Point geometries provide direct longitude/latitude. LineString geometries (cable car spans, transmission line segments) are reduced to their geographic midpoint $(\bar{\lambda}, \bar{\varphi})$ by averaging all vertex coordinates:

$$\bar{\lambda} = \frac{1}{n}\sum_{i=1}^{n} \lambda_i, \qquad \bar{\varphi} = \frac{1}{n}\sum_{i=1}^{n} \varphi_i \tag{1}$$

The result is a GeoDataFrame with 14,334 records in WGS 84 (EPSG:4326), saved as Parquet. The core parsing logic handles two KML encoding patterns:

```python
# Pattern 1: <Data name="..."><value>...</value></Data>
for d in extended.findall(f"{KML_NS}Data"):
    name = d.get("name", "")
    value_el = d.find(f"{KML_NS}value")
    if name and value_el is not None:
        data[name] = value_el.text

# Pattern 2: <SchemaData><SimpleData name="...">value</SimpleData>
for sd in extended.findall(f".//{KML_NS}SimpleData"):
    name = sd.get("name", "")
    if name:
        data[name] = sd.text
```

2.2 What the Data Contains

The schema comprises 18 attributes. Three are analytically central: obstacle type (10 categories), height above ground level in metres, and top elevation above mean sea level. Table 1 shows the full picture.

Tabelle 1: Dataset schema. Null rates reflect structural absence, not quality issues.

Column	Dtype	Non-Null	Null %	Notes
uuid	str	14 334	0.0	Unique identifier, zero duplicates
obstacleType	str	14 334	0.0	10 distinct types
maxHeightAGL	float64	14 334	0.0	Height above ground level (m)
topElevationAMSL	float64	14 334	0.0	Top elevation above mean sea level (m)
airport	str	3 233	77.4	ICAO code; null for non-airfield obstacles
lighting	str	7 970	44.4	Null for 6 364 transmission lines
radius	float64	504	96.5	Only populated for area obstacles
small	object	0	100.0	Entirely null; dropped

2.3 Key Observations

Two findings from the quality assessment directly shaped the subsequent analysis. First, both height and elevation distributions are right-skewed (Table 2), with outliers reaching 586 m AGL and 3,870 m AMSL. These represent real structures (alpine cable car stations, tall transmission towers) and are retained.

Tabelle 2: Descriptive statistics for height and elevation ($N = 14{,}334$).

	Mean	Std	Median	Max	Outlier %
maxHeightAGL (m)	48.72	44.24	35.00	586.40	< 8.2
topElevationAMSL (m)	916.64	539.76	728.75	3 869.60	< 10

Second, linear infrastructure (transmission lines, poles, catenaries) accounts for roughly 80% of all records. This class imbalance means that any naive clustering of individual obstacles would

produce one massive "transmission line" cluster and miss subtler spatial patterns. This observation motivated the decision to aggregate obstacles onto a spatial grid and engineer features that capture the composition of each area, rather than clustering individual records.

3 From Points to a Feature Grid

3.1 Choosing H3 Resolution 7

Rather than clustering 14,334 individual obstacle points, the pipeline aggregates them onto a hexagonal grid. Hexagons have uniform adjacency (every cell has exactly six neighbours) and avoid the edge-distortion of rectangular grids, making them well suited for spatial analysis. The grid system is Uber's H3 at resolution $r = 7$:

Tabelle 3: H3 resolution trade-off. Resolution 7 balances coverage and statistical robustness.

Resolution	Cell Area	Edge Length	Trade-off
6	36.13 km^2	3.23 km	Too coarse, few cells
7	5.16 km^2	1.22 km	Selected: 3,976 cells, median 2 obs.
8	0.74 km^2	0.46 km	Many single-obstacle cells
9	0.11 km^2	0.17 km	Used for cantonal zoom only

3.2 Ten Features per Cell

For each cell h with n_h obstacles, ten features are computed. The height statistics capture the vertical risk profile of each cell:

$$\bar{x}_h = \frac{1}{n_h} \sum_{i \in \mathcal{O}_h} x_i, \qquad x_h^{\max} = \max_{i \in \mathcal{O}_h} x_i, \qquad s_h = \sqrt{\frac{1}{n_h - 1} \sum_{i \in \mathcal{O}_h} (x_i - \bar{x}_h)^2} \tag{2}$$

where x_i is the height above ground of obstacle i, and $s_h = 0$ for single-obstacle cells.

Type composition captures what kind of infrastructure dominates a cell. The following excerpt shows the full aggregation function:

```
LINEAR  = {'TRANSMISSION_LINE', 'POLE', 'CATENARY'}
VERTICAL = {'BUILDING', 'STACK', 'WINDMILL', 'CRANE'}

def engineer_features(group):
    n = len(group)
    return pd.Series({
        'obstacle_count': n,
        'height_mean':    group['maxHeightAGL'].mean(),
        'height_max':     group['maxHeightAGL'].max(),
        'height_std':     group['maxHeightAGL'].std() if n > 1 else 0,
        'elevation_mean': group['topElevationAMSL'].mean(),
        'pct_linear':     group['obstacleType'].isin(LINEAR).sum() / n * 100,
        'pct_vertical':   group['obstacleType'].isin(VERTICAL).sum() / n * 100,
        'pct_diversity':  len(group['obstacleType'].unique()),
        'pct_lighted':    group['lighting'].isin(
                              ['LIGHTED','LOW','MEDIUM','HIGH']).sum() / n * 100,
        'has_airport':    int(group['airport'].notna().any()),
    })
```

```
grid = gdf.groupby('h3_index').apply(engineer_features).reset_index()
```

The resulting feature distributions are summarised in Table 4.

Tabelle 4: Engineered grid-level features ($|\mathcal{H}| = 3{,}976$ cells).

Category	Feature	Mean	Std	Description
Density	obstacle count	3.61	4.61	n_h
Height	height mean	50.73	35.47	$\bar{x}_h$ (m AGL)
	height max	70.34	56.53	$x_h^{\max}$ (m AGL)
	height std	15.32	23.55	s_h
Terrain	elevation mean	1 041.56	587.29	Mean AMSL (m)
Type	% linear	82.49	32.34	p_h^{lin}
	% vertical	8.68	23.64	p_h^{vert}
	type diversity	1.56	0.83	Distinct types per cell
Risk	% lighted	7.27	21.92	Lighting coverage
	has airport	0.15	0.36	Binary airfield flag

A correlation check confirms that linear and vertical percentages are inversely related (Pearson $r = -0.66$): cells are either transmission corridors or built-up zones, rarely both. This pattern suggests that a clustering algorithm should be able to separate Swiss airspace into distinct profiles.

4 Finding Cluster Structure

4.1 Standardisation and Feature Selection

All 10 features are z-score normalised to zero mean and unit variance before clustering. An important design decision is that latitude and longitude are excluded from the feature matrix. This forces the algorithm to group cells by their obstacle profile, not their location. A cell in the Jura and a cell in the Engadin can end up in the same cluster if their obstacles share similar characteristics.

4.2 Three Algorithms, One Winner

HDBSCAN was tried first because it does not require a pre-specified k and can label ambiguous points as noise. A sweep over five minimum cluster size values revealed a trade-off between granularity and noise (Table 5).

Tabelle 5: HDBSCAN parameter sweep (min_samples = 5 fixed).

min_cluster_size	Clusters	Noise %	$\bar{s}$
15	17	23.2	0.220
25	13	23.4	0.229
50	10	42.6	0.099
75	5	53.9	0.018
100	3	59.8	−0.030

The best configuration (mcs = 25) still assigns nearly a quarter of all cells to noise, making it difficult to produce a complete risk map. DBSCAN was tried next, but its best result ($\varepsilon = 1.5$) placed 77% of cells into a single cluster, which is operationally meaningless.
K-Means requires a pre-specified k but assigns every cell to a cluster. The elbow method on WCSS and silhouette analysis converge on $k = 4$:

```
wcss, sil_scores = [], []
for k in range(2, 11):
    km = KMeans(n_clusters=k, init='k-means++', random_state=42, n_init=10)
    labels = km.fit_predict(X_scaled)
    wcss.append(km.inertia_)
    sil_scores.append(silhouette_score(X_scaled, labels))
```

Tabelle 6: K-Means evaluation across $k = 2 \ldots 10$. Peak silhouette at $k = 4$.

k	WCSS	Silhouette $\bar{s}$
2	32 489	0.331
3	25 978	0.360
4	22 862	0.363
5	20 570	0.268
6	18 722	0.291
7	17 088	0.306
8	15 750	0.289
9	14 504	0.292
10	13 518	0.290

The comparison across all three methods (Table 7) confirms K-Means with $k = 4$ as the best choice: highest silhouette, complete coverage, and the most interpretable segmentation.

Tabelle 7: Final method comparison.

Method	Config	K	Noise %	$\bar{s}$	Assessment
HDBSCAN	mcs = 25, ms = 5	13	23.4	0.229	Fine-grained, high noise
K-Means	$k = 4$, k-means++	4	0.0	0.363	Best score, interpretable
DBSCAN	$\varepsilon = 1.5$, ms = 5	11	5.0	0.310	77% in one cluster

4.3 Four Zones of Swiss Airspace

The four clusters correspond to operationally distinct airspace profiles (Table 8).

Tabelle 8: Cluster profiles (per-cluster means). $\bar{n}_h$ = obstacle count, $\bar{x}_h$ = height AGL, $\bar{e}_h$ = elevation AMSL.

Zone	Cells	Share	$\bar{n}_h$	$\bar{x}_h$	$\bar{e}_h$	p^{lin}	p^{vert}	p^{lit}
Plateau Transmission	2 539	63.9%	2.4	40 m	1 057 m	91%	1%	4%
Airport & Urban	551	13.9%	8.1	38 m	693 m	76%	12%	9%
Alpine High-Rise	602	15.1%	4.7	107 m	1 481 m	82%	1%	4%
Urban Vertical	284	7.1%	3.2	51 m	651 m	16%	82%	45%

Plateau Transmission (63.9%) represents the Swiss Plateau's power grid: low-lying cells with sparse, low-height transmission lines. Airport & Urban (13.9%) captures high-density zones near airfields such as LSZH and LSGG, where 94% of cells have airport associations. Alpine High-Rise (15.1%) identifies alpine infrastructure: cable car lines and valley crossings at a mean elevation of 1,481 m with extreme heights averaging 107 m. Urban Vertical (7.1%) marks industrial and built-up zones where 82% of obstacles are buildings, stacks, or cranes, and 45% carry lighting.

Although coordinates were excluded from the clustering features, the geographic overlay reveals strong spatial coherence: the four zones align with recognisable Swiss landscapes.

5 Validation

The cluster assignment needs to be verified in two spaces: the abstract feature space (do the clusters actually separate?) and geographic space (do the zones make operational sense?).

5.1 PCA and t-SNE

PCA on the standardised matrix shows that 8 of 10 components are needed to capture 95% of variance. PC1 (27.3%) and PC2 (23.4%) together explain 50.7%, providing a partial but informative 2D view. In this projection, Plateau Transmission forms a compact group while Alpine High-Rise fans out along the height-driven PC1 axis. A biplot overlay confirms that elevation, maximum height, and height standard deviation drive the Alpine separation, while the vertical percentage and lighting coverage drive Urban Vertical.

t-SNE at three perplexity values (10, 30, 60) consistently produces four separated regions, confirming the structure is robust across parameter choices. A hybrid approach (PCA to 8 components, then t-SNE at perplexity 60) yields the cleanest result. A side-by-side comparison coloured by K-Means versus HDBSCAN labels shows that K-Means produces tighter, more clearly separated islands.

5.2 Silhouette Analysis

The per-cluster silhouette plot reveals soft boundaries between Airport & Urban and Alpine High-Rise, consistent with gradual geographic transitions. No cluster shows systematically negative values. The moderate global score of $\bar{s} = 0.363$ is expected: geographic data rarely produces hard cluster boundaries, and the score exceeds both HDBSCAN (0.229) and DBSCAN (0.310).

6 Operational Risk Profiling

6.1 Risk Classification

With the four zones validated, the final step translates them into actionable risk levels for UAS planning (Table 9).

Tabelle 9: Operational risk classification per zone.

Zone	Risk	Rationale
Plateau Transmission	Moderate	Predictable linear geometry, low height variance
Airport & Urban	High	Airfield proximity, ATC coordination required
Alpine High-Rise	High	Extreme heights, cable car crossings, limited lighting
Urban Vertical	Moderate-High	Dense vertical structures, dynamic obstacles (cranes)

6.2 Corridor Identification

Low-risk transit corridors are cells satisfying all four constraints simultaneously:

```
corridor_mask = (
    (grid['obstacle_count'] <= 2) &
    (grid['height_max'] <= 50) &
    (grid['has_airport'] == 0) &
    (grid['pct_lighted'] == 0)
)
```

Cells with no registered obstacles at all are visualised as "clear airspace" on the final map. The maps use official swisstopo boundary geometries from the geo.admin.ch REST API and are exported as standalone HTML files with interactive popups per cell (Figure 1).

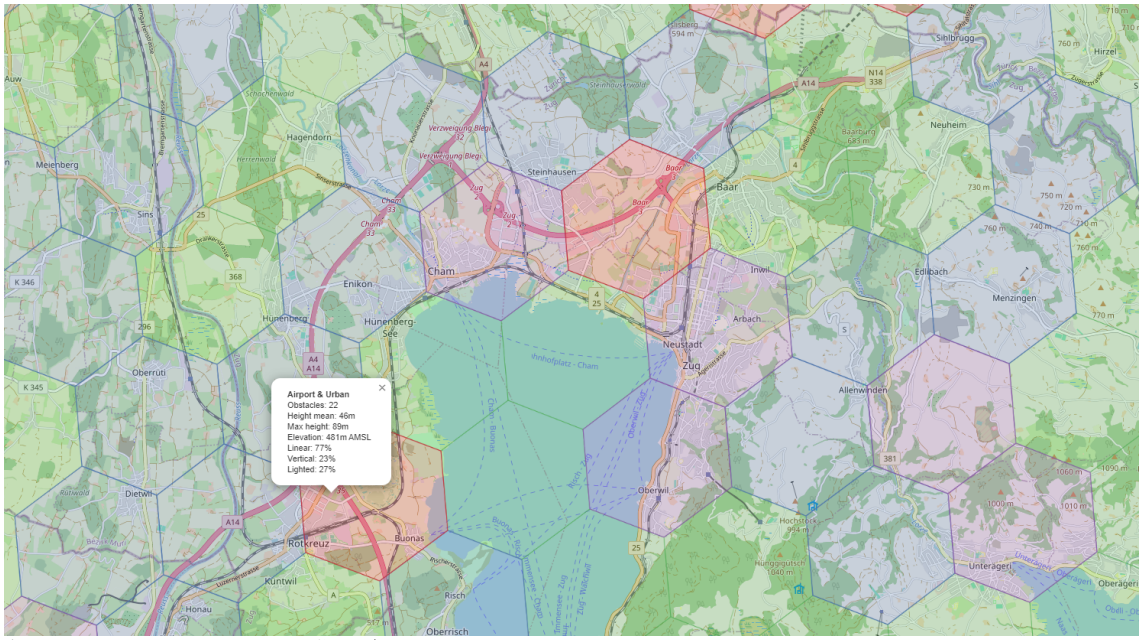


Abbildung 1: Zoomed view near Cham/Zug. Popup shows per-cell statistics: zone label, obstacle count, height, elevation, and type composition.

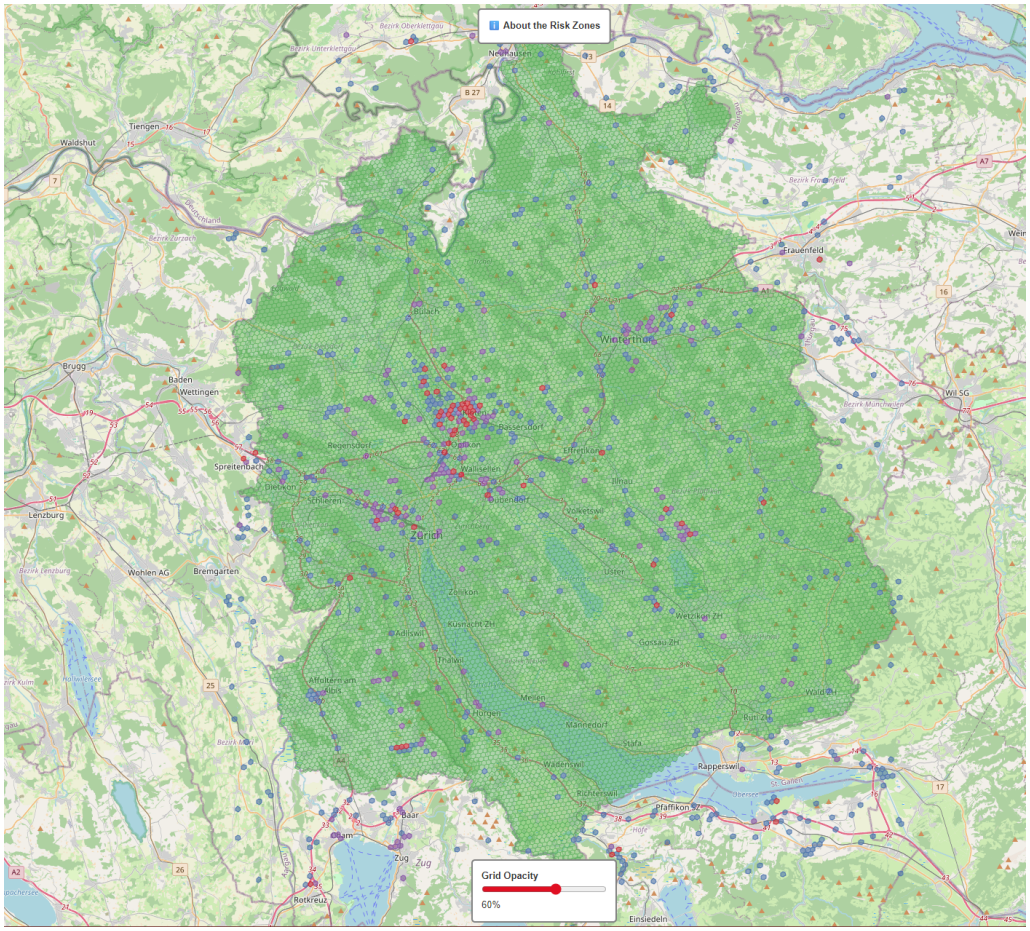
6.3 Scaling to Operational Granularity

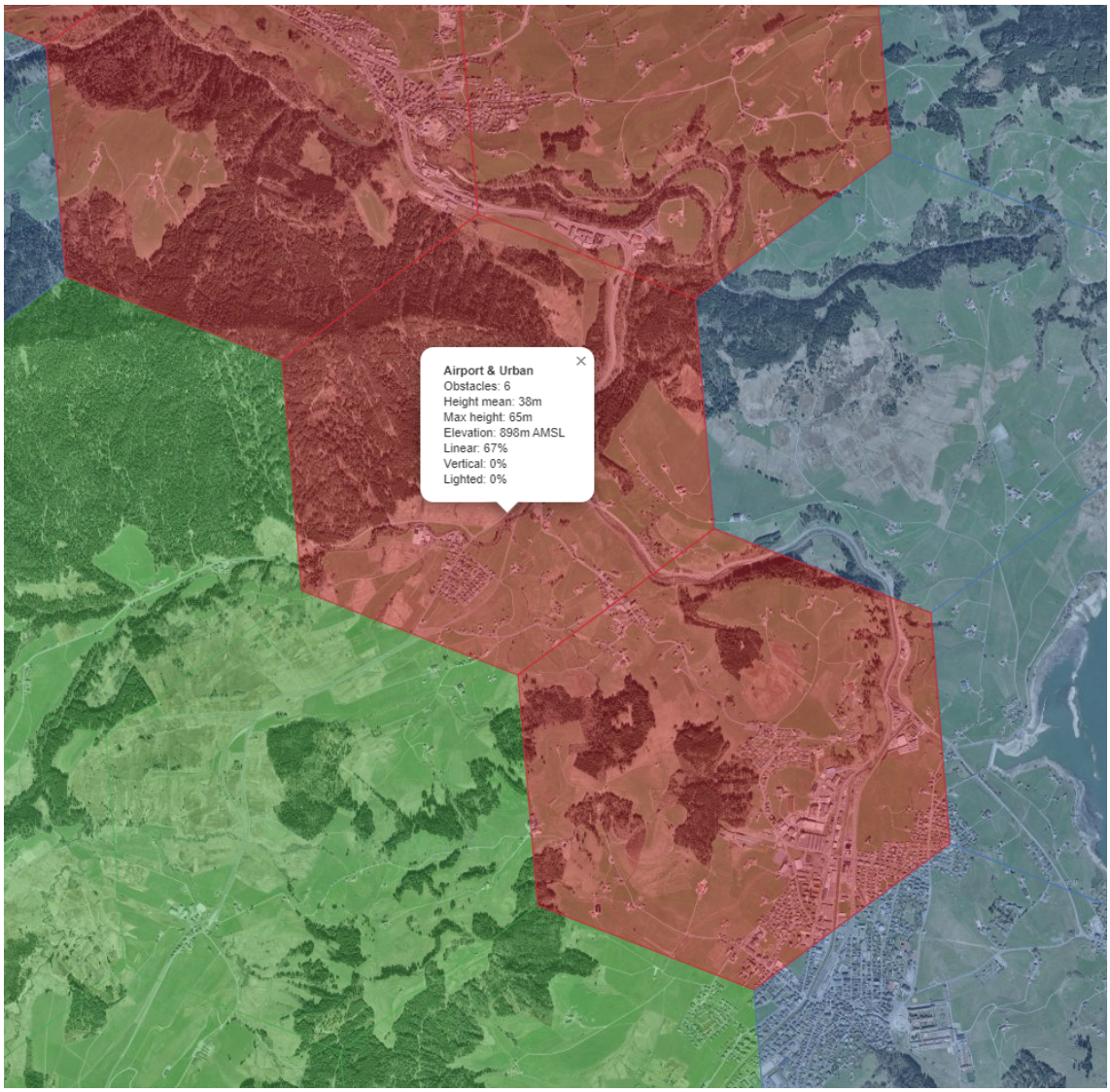
A resolution 7 cell covers $\approx 5\,\mathrm{km}^2$: useful for strategic overview, too coarse for route planning. To demonstrate scalability, the entire pipeline is re-executed for Canton Zürich at resolution 9 ($\approx 0.1\,\mathrm{km}^2$ per cell). This requires no architectural changes:

```
# Only parameter change: resolution constant
H3_RESOLUTION = 9

zh_gdf['h3_index'] = zh_gdf.apply(
    lambda row: h3.latlng_to_cell(row['latitude'], row['longitude'],
                                  H3_RESOLUTION), axis=1)
```

Cluster labels at the cantonal level are assigned by matching each cluster's feature profile to the national-level zones. The figure shows the result: intra-zone variation invisible at the national scale becomes apparent, with neighbourhood-level transitions between risk categories.





7 Implementation

Table 10 shows the pipeline data flow. Each notebook reads a single Parquet file and writes one, enabling independent re-execution.

Tabelle 10: Pipeline structure and data flow.

Component	Input	Output
Fetch script	STAC API / KMZ	obstacles.parquet
NB 01: Data Ingestion	obstacles.parquet	obstacles_clean.parquet
NB 02: Feature Engineering	obstacles_clean.parquet	grid_features.parquet
NB 03: Clustering	grid_features.parquet	grid_clustered.parquet
NB 04: Dim. Reduction	grid_clustered.parquet	Visualisations only
NB 05: Corridor Profiling	grid_clustered.parquet	Interactive HTML maps

The stack: Python 3.10+, GeoPandas, H3, scikit-learn (StandardScaler, K-Means), HDBSCAN, Folium. Raw data is excluded from version control and reproduced by the fetch script.

8 Limitations and Future Work

This analysis covers registered air navigation obstacles only. A production-grade flight planning system would additionally require terrain elevation models, land-use classification, SORA ground risk data, restricted airspace zones, and real-time NOTAMs. These are planned as separate follow-up projects.

The silhouette score of $\bar{s} = 0.363$ reflects moderate separation, consistent with the continuous nature of geographic data. Fuzzy c-means or Gaussian Mixture Models could model the soft inter-cluster transitions more explicitly.

9 Conclusion

SOAP demonstrates a complete geospatial data science pipeline from raw government API data to an operational airspace risk classification. The four identified zones provide a coherent segmentation of Switzerland based exclusively on obstacle characteristics, and the hexagonal grid enables seamless resolution scaling (validated from $5\,\mathrm{km}^2$ to $0.1\,\mathrm{km}^2$) without pipeline modification. All data is publicly available, the pipeline is fully reproducible, and the interactive HTML maps are self-contained for stakeholder communication.
