

Spatial Association Rule Mining with an easy-to-follow example

1. Concept Recap

Spatial Association Rule Mining finds **relationships** between spatial features and their attributes.
It answers questions like:

2. Simple Real-Life Example

Imagine a city GIS database with:

- **Feature 1:** Locations of *parks*
- **Feature 2:** Locations of *restaurants*
- **Attribute:** *High customer ratings* for restaurants

You might find a rule like:

IF a restaurant is *within 500 meters of a park*

THEN it has a *high customer rating*

Confidence: 85%

This means **85% of restaurants near parks are highly rated.**

3. Numerical Example

Dataset (simplified):

Restaurant ID Near Park? (Y/N) High Rating? (Y/N)

R1	Y	Y
R2	Y	Y
R3	Y	N
R4	N	N
R5	Y	Y
R6	N	N
R7	Y	Y
R8	N	Y
R9	Y	Y
R10	N	N

Step 1 – Support:

Support measures how often both the condition and result occur.

Support=Number of cases where Near Park = Y AND High Rating = Y
Total records\text{Support} = \frac{\text{Number of cases where Near Park = Y AND High Rating = Y}}{\text{Total records}}

Count:

- Near Park = Y **and** High Rating = Y → R1, R2, R5, R7, R9 → **5 records**

Support=510=0.5 (50%)\text{Support} = \frac{5}{10} = 0.5 \ (50\%)

Step 2 – Confidence:

Confidence measures how often the rule is correct when the condition is true.

Confidence=Near Park = Y AND High Rating = Y
Near Park = Y\text{Confidence} = \frac{\text{Near Park = Y AND High Rating = Y}}{\text{Near Park = Y}}

Count:

- Near Park = Y → R1, R2, R3, R5, R7, R9 → **6 records**

Confidence=56≈83.3%\text{Confidence} = \frac{5}{6} \approx 83.3\%

Step 3 – Rule:

Rule: If Near Park = Yes → High Rating = Yes\text{Rule: If Near Park = Yes → High Rating = Yes}

- **Support:** 50%
- **Confidence:** 83.3%

Interpretation:

Half of all restaurants in the dataset are near parks *and* have high ratings.

When a restaurant is near a park, there's an 83.3% chance it has a high rating.

If you like, I can also give you a **spatial map diagram** showing these points so you can *visually* see the clusters and understand why the rule emerges — which is very helpful in spatial mining studies.

Spatial Data Mining (SDM) is the process of discovering interesting, useful, and non-trivial patterns from **spatial datasets** — data that includes location, shape, and spatial relationships in addition to regular attributes.

It's basically **data mining + GIS concepts** — we're not just looking at "what" the data says, but also **where** it occurs and how spatial features interact.

Key Characteristics

1. **Spatial dependency** – Nearby things tend to be more related than distant things (Tobler's First Law).
2. **Spatial heterogeneity** – Patterns may change over different regions.

-
3. **Complex data types** – Includes points, lines, polygons, and raster data with geometry + attributes.
-

Types of Spatial Data Mining Tasks

Task	Description	Example
Spatial Classification	Assigning categories based on spatial and attribute data	Classifying land use types from satellite imagery
Spatial Clustering	Grouping similar spatial objects into clusters	Identifying disease outbreak zones
Spatial Association Rule Mining	Finding relationships between spatial features	“If a house is near a park, it is likely to have higher value”
Spatial Outlier Detection	Finding unusual spatial patterns	A cold spot in a normally warm area
Spatial Trend Detection	Discovering directional patterns	Soil quality decreasing from north to south
Spatial Prediction	Predicting unknown spatial values based on known data	Predicting rainfall in unmeasured areas

Applications

- **Urban planning** → Detecting growth patterns of cities
 - **Agriculture** → Crop health monitoring using satellite imagery
 - **Epidemiology** → Tracking and predicting disease spread
 - **Environmental monitoring** → Detecting deforestation patterns
 - **Retail** → Location-based customer behavior analysis
-

Techniques Used

- **Clustering algorithms** (DBSCAN, k-means with spatial constraints)
 - **Spatial regression models** (Geographically Weighted Regression)
 - **Spatial decision trees & rule mining**
 - **Deep learning for spatial imagery** (CNNs for remote sensing)
-

Alright — let's break down **Spatial Association Rule Mining** with an **easy-to-follow example** and a **numerical illustration** so you can really “see” how it works.

1. Concept Recap

Spatial Association Rule Mining finds **relationships** between spatial features and their attributes. It answers questions like:

"If a spatial feature A occurs near spatial feature B, then C is likely true."

2. Simple Real-Life Example

Imagine a city GIS database with:

- **Feature 1:** Locations of *parks*
- **Feature 2:** Locations of *restaurants*
- **Attribute:** *High customer ratings* for restaurants

You might find a rule like:

IF a restaurant is *within 500 meters of a park*

THEN it has a *high customer rating*

Confidence: 85%

This means **85% of restaurants near parks are highly rated.**

3. Numerical Example

Dataset (simplified):

Restaurant ID Near Park? (Y/N) High Rating? (Y/N)

R1	Y	Y
R2	Y	Y
R3	Y	N
R4	N	N
R5	Y	Y
R6	N	N
R7	Y	Y
R8	N	Y
R9	Y	Y
R10	N	N

Step 1 – Support:

Support measures how often both the condition and result occur.

Support=Number of cases where Near Park = Y AND High Rating = Y
Total records\text{Support} = \frac{\text{Number of cases where Near Park = Y AND High Rating = Y}}{\text{Total records}}

Count:

- Near Park = Y **and** High Rating = Y → R1, R2, R5, R7, R9 → **5 records**

Support=5/10=0.5 (50%)\text{Support} = \frac{5}{10} = 0.5 \ (50\%)

Step 2 – Confidence:

Confidence measures how often the rule is correct when the condition is true.

Confidence=Near Park = Y AND High Rating = Y
Near Park = Y \text{Confidence} = \frac{\text{Near Park = Y AND High Rating = Y}}{\text{Near Park = Y}}

Count:

- Near Park = Y → R1, R2, R3, R5, R7, R9 → **6 records**

Confidence=56≈83.3%\text{Confidence} = \frac{5}{6} \approx 83.3\%

Step 3 – Rule:

Rule: If Near Park = Yes → High Rating = Yes \text{Rule: If Near Park = Yes → High Rating = Yes}

- **Support:** 50%
- **Confidence:** 83.3%

Interpretation:

Half of all restaurants in the dataset are near parks *and* have high ratings.
When a restaurant is near a park, there's an 83.3% chance it has a high rating.

Sure — here are a few **more practical examples of Spatial Association Rule Mining** across different domains so you can see how it's applied in real life:

1. Urban Planning

Dataset: City GIS with layers for schools, playgrounds, and traffic accident locations.

Rule Found:

IF a school is within 200 meters of a busy intersection
THEN traffic accidents in that area are high

Support: 40%

Confidence: 78%

Use: City planners might decide to add traffic signals or pedestrian bridges.

2. Agriculture

Dataset: Farmland maps with soil type, irrigation sources, and crop yield.

Rule Found:

IF soil type = "Loamy" AND near river < 1 km
THEN crop yield is high
Support: 55%
Confidence: 87%

Use: Farmers can plan high-value crops near such soil/water conditions.

3. Epidemiology (Disease Spread)

Dataset: Locations of mosquito breeding sites, hospitals, and dengue cases.
Rule Found:

IF a residential area is within 500 meters of >3 mosquito breeding sites
THEN dengue cases are high
Support: 35%
Confidence: 82%

Use: Public health agencies can target vector control in high-risk zones.

4. Retail & Business

Dataset: Store locations, competitor locations, and sales volume.
Rule Found:

IF a retail store is within 1 km of a metro station AND no competitor within 500 m
THEN sales are high
Support: 48%
Confidence: 91%

Use: Helps businesses choose profitable store locations.

5. Environmental Monitoring

Dataset: Deforestation maps, slope data, rainfall patterns.
Rule Found:

IF slope < 10° AND rainfall > 2000 mm/year
THEN deforestation probability is high
Support: 30%
Confidence: 80%

Use: Environmental agencies can prioritize these areas for conservation.

Here's your **Spatial Association Rule Mining Quick Revision Sheet**:

Spatial Association Rule Mining – Examples Summary

Domain	Example Rule	Support	Confidence	Use Case
Urban Planning	IF a school is within 200 m of a busy intersection → THEN traffic accidents are high	40%	78%	Install signals, pedestrian bridges
Agriculture	IF soil type = Loamy AND near river < 1 km → THEN crop yield is high	55%	87%	Choose crop locations for high yield
Epidemiology	IF residential area within 500 m of >3 mosquito breeding sites → THEN dengue cases are high	35%	82%	Target vector control programs
Retail	IF store within 1 km of metro AND no competitor within 500 m → THEN sales are high	48%	91%	Optimal retail location planning
Environment	IF slope < 10° AND rainfall > 2000 mm/year → THEN deforestation probability is high	30%	80%	Prioritize conservation measures

Alright — here's your **Spatial Association Rule Mining Practice Set** with **numeric datasets** for each domain so you can practice calculating **support** and **confidence**.

1. Urban Planning

Dataset (10 intersections):

Intersection ID Near School? High Accident Rate?

I1	Y	Y
I2	N	N
I3	Y	Y
I4	Y	N
I5	Y	Y
I6	N	N
I7	Y	Y
I8	N	Y
I9	Y	Y

Intersection ID Near School? High Accident Rate?

I10 N N

Task:

Rule: *If Near School = Yes → High Accident Rate = Yes*

- **Support** = (Near School = Y & Accident Rate = Y) ÷ Total records
 - **Confidence** = (Near School = Y & Accident Rate = Y) ÷ (Near School = Y)
-

2. Agriculture

Dataset (10 farms):

Farm ID Loamy Soil? Near River (<1 km)? High Yield?

F1	Y	Y	Y
F2	N	Y	N
F3	Y	Y	Y
F4	Y	N	N
F5	Y	Y	Y
F6	N	N	N
F7	Y	Y	Y
F8	Y	N	N
F9	Y	Y	Y
F10	N	Y	N

Task:

Rule: *If Loamy Soil = Y AND Near River = Y → High Yield = Y*

3. Epidemiology

Dataset (10 residential zones):

Zone ID >3 Mosquito Sites? Dengue Cases High?

Z1	Y	Y
Z2	Y	Y
Z3	Y	N
Z4	N	N
Z5	Y	Y
Z6	N	N
Z7	Y	Y
Z8	N	Y
Z9	Y	Y
Z10	N	N

Task:

Rule: *If Mosquito Sites >3 → Dengue Cases High*

4. Retail

Dataset (10 stores):

Store ID Near Metro (<1 km)? Competitor within 500m? Sales High?

S1	Y	N	Y
S2	Y	N	Y
S3	Y	Y	N
S4	N	N	N
S5	Y	N	Y
S6	N	Y	N
S7	Y	N	Y
S8	Y	N	Y
S9	N	N	Y
S10	Y	N	Y

Task:

Rule: *If Near Metro = Y AND Competitor = N → Sales High = Y*

5. Environment

Dataset (10 land plots):

Plot ID Slope <10°? Rainfall >2000 mm? Deforestation?

P1	Y	Y	Y
P2	N	Y	N
P3	Y	Y	Y
P4	Y	N	N
P5	Y	Y	Y
P6	N	N	N
P7	Y	Y	Y
P8	Y	N	N
P9	Y	Y	Y
P10	N	Y	N

Task:

Rule: *If Slope <10° AND Rainfall >2000 mm → Deforestation = Yes*

 Your Practice Steps for Each:

1. Count records where **condition AND result** are true → gives **numerator** for both support & confidence.
 2. Support = (condition & result) ÷ total records.
 3. Confidence = (condition & result) ÷ (condition).
-

Nice — let's solve the **Urban Planning** example step-by-step (with all arithmetic shown) so it's exam-ready.

Dataset recap (10 intersections)

Rows with Near School = Y and High Accident = Y are: **I1, I3, I5, I7, I9**.

Rows with Near School = Y (regardless of accidents): **I1, I3, I4, I5, I7, I9** (6 rows).

Rows with High Accident = Y (regardless of school): **I1, I3, I5, I7, I8, I9** (6 rows).

Total records = **10**.

1) Support of rule

Rule: **If Near School = Y → High Accident = Y**

Support = (count of NearSchool=Y **and** HighAccident=Y) ÷ (total records)

Numerator = 5 (I1,I3,I5,I7,I9)

Denominator = 10

Compute $5 \div 10$:

- $5/10 = 1/2 = 0.5$

Support = 0.5 (50%)

2) Confidence of rule

Confidence = (count of NearSchool=Y **and** HighAccident=Y) ÷ (count of NearSchool=Y)

Numerator = 5

Denominator = 6

Compute $5 \div 6$ by long division:

- 6 goes into 5 → 0 times, remainder 5 → write 0 . and continue with decimal
- 6 goes into 50 → 8 times ($8 \times 6 = 48$), remainder 2 → first decimal digit 8
- bring down 0 → 20; 6 goes into 20 → 3 times ($3 \times 6 = 18$), remainder 2 → next digit 3
- remainder 2 repeats so digits 3 repeat

So $5 \div 6 = 0.833333\dots \approx 0.8333 (83.33\%)$

Confidence ≈ 83.33%

3) Lift (optional but useful)

Lift = Confidence / Support(consequent)

Support(consequent) = Support(HighAccident = Y) = 6/10 = 0.6

Lift = $(5/6) \div (6/10)$

Compute algebraically: $(5/6) \times (10/6) = 50/36 = 25/18$

Compute $25 \div 18$ by long division:

- 18 into 25 → 1 ($1 \times 18 = 18$), remainder 7 → 1 .
- bring 0 → 70; 18→3 ($3 \times 18=54$), remainder 16 → next digit 3
- bring 0 → 160; 18→8 ($8 \times 18=144$), remainder 16 → digits 8 repeat

So $25/18 = 1.388888\dots \approx \mathbf{1.3889}$

Lift ≈ 1.39 (> 1) → positive association (restaurants near schools are more likely to have high accident areas than average).

4) Interpretation (straightforward)

- **Support 50%:** half of all intersections are both near a school and have high accidents.
 - **Confidence ~83%:** when an intersection is near a school, there is an ~83% chance it has a high accident rate.
 - **Lift ~1.39 (>1):** presence of a school increases the likelihood of high accidents by ~39% compared to a random intersection.
-

5) What to do with this rule

- Treat it as a **strong indicator** for intervention (traffic calming, signals, pedestrian crossings) around schools.
 - Before action, check for confounders (traffic volume, road design) and spatial autocorrelation (nearby intersections influencing each other).
-

Assistant Professor,
Computer Engineering Department
K.J Somaiya school of Engineering