# Spatial Data Modeling and Analysis

## Contents

**Spatial Data Model**

- Spatial data are often referred to as **layers**, **coverages**, or **layers**. We will use the term **layers** from this point on, since this is the recognized term used in ArcGIS. Layers represent, in a special digital storage format, features on, above, or below the surface of the earth.
- Depending on the type of features they represent, and the purpose to which the data will be applied, layers will be one of 2 major types.
    - **Vector** data represent features as discrete points, lines, and polygons.
    - **Raster** data represent the landscape as a rectangular matrix of square cells.
- Depending on the type of problem that needs to be solved, the type of maps that need to be made, and the data source, either raster or vector, or a combination of the two can be used. Each data model has strengths and weaknesses in terms of functionality and representation.

# VECTOR DATA ANALYSIS

- The vector data model uses points and their *x-, y*-coordinates to construct spatial features of points, lines, and polygons.
- Spatial features are used as inputs in vector data analysis. Therefore, the accuracy of data analysis depends on the accuracy of these features in terms of their location and shape and whether they are topological or not.
- Additionally, it is important to note that an analysis may apply to all, or selected, features in a layer. The following are the types of analysis used with vector data.
    - Buffering
    - Overlay
    - Distance Measurement
    - Pattern Analysis
    - Feature Manipulation

**Buffering**

- Buffering creates two areas:
- one area that is within a specified distance of select features and the other area that is beyond.
- The area within the specified distance is the buffer zone.

- Features for buffering may be points, lines, or polygons.
- Buffering around points creates circular buffer zones.
- Buffering around lines creates a series of elongated buffer zones around each line segment.
- Buffering around polygons creates buffer zones that extend outward from the polygon boundaries.

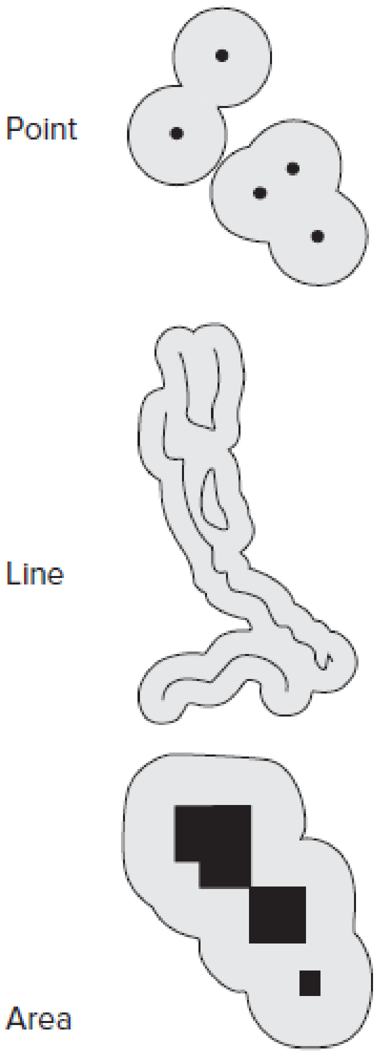



Fig 1: Buffering around points, lines, and polygons.

**Variations in Buffering**

There are several variations in buffering from those of Fig 1.

- The buffer distance or buffer size does not have to be constant; it can vary according to the values of a given field (Fig 2). For example, the width of the riparian buffer can vary depending on its expected function and the intensity of adjacent land use.
- A feature may have more than one buffer zone. As an example, a nuclear power plant may be buffered with distances of 5, 10, 15, and 20 miles, thus forming multiple rings around the plant (Fig 3). Although the interval of each ring is the same at 5 miles, the rings are not equal in area. The second ring from the plant, in fact, covers an area about three times larger than the first ring. One must consider this area difference if the buffer zones are part of an evacuation plan.

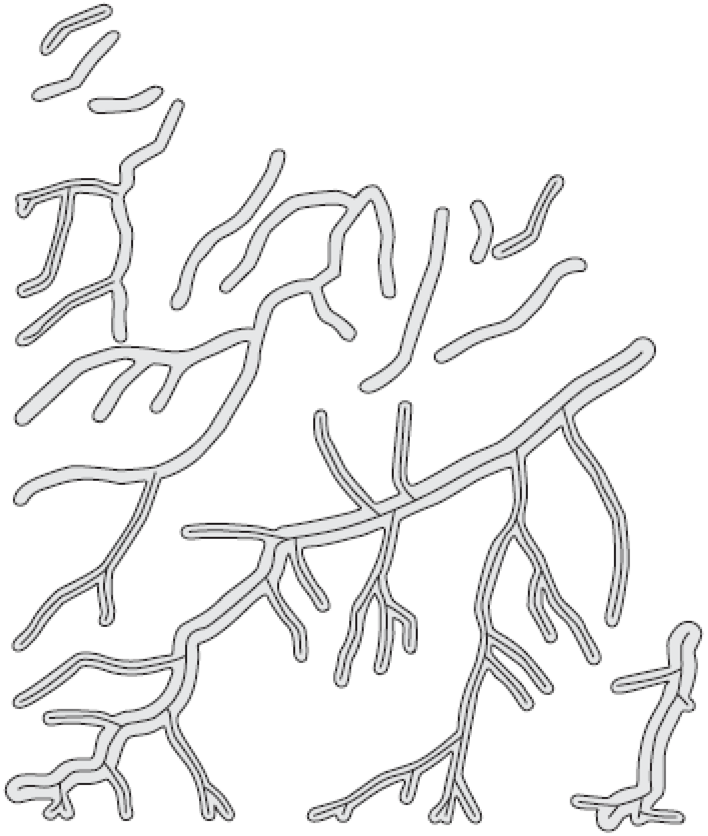


Fig 2: Buffering with different buffer distances.

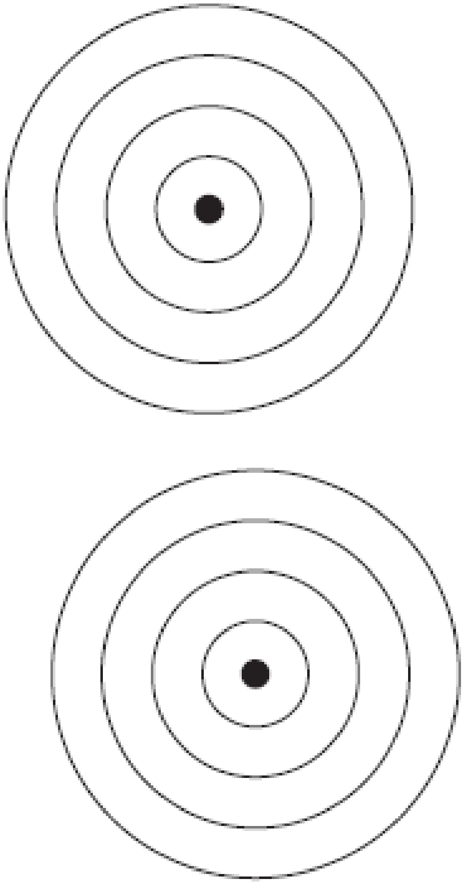


Fig 3: Buffering with four rings.

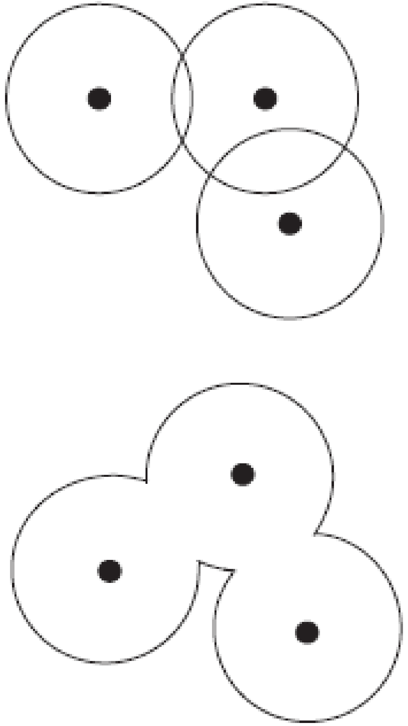


Fig 4: Buffer zones not dissolved (top) or dissolved (bottom).

- Likewise, buffering around line features does not have to be on both sides of the lines; it can be on either the left side or the right side of the line feature.
- Likewise, buffer zones around polygons can be extended either outward or inward from the polygon boundaries.
- Boundaries of buffer zones may remain intact so that each buffer zone is a separate polygon for further analysis. Or these boundaries may be dissolved to create an aggregate zone, leaving no overlapped areas between individual buffer zones (Fig 4). Even the ends of buffer zones can be either rounded or flat.

**Applications of Buffering:** A buffer zone is often treated as a protection zone and is used for planning or regulatory purposes:

- A city ordinance may stipulate that no liquor stores or pornographic shops shall be within 1000 feet of a school or a church.
- Government regulations may set 2-mile buffer zones along streams to minimize sedimentation from logging operations.
- A national forest may restrict oil and gas well drilling within 500 feet of roads or highways.
- A planning agency may set aside land along the edges of streams to reduce the effects of nutrient, sediment, and pesticide runoff; to maintain shade to prevent the rise of stream temperature; and to provide shelter for wildlife and aquatic life.
- A planning agency may create buffer zones around geographic features such as water, wetlands, critical habitats, and wells to be protected and exclude these zones from landfill siting consideration

## Overlay

- An overlay operation combines the geometries and attributes of two feature layers to create the output. (A GIS package may offer overlay operations with more than two layers at a time)
- The geometry of the output represents the geometric intersection of features from the input layers.
- Overlay operation with two simple polygon layers. Each feature on the output contains a combination of attributes from the input layers, and this combination differs from its neighbors (figure below).

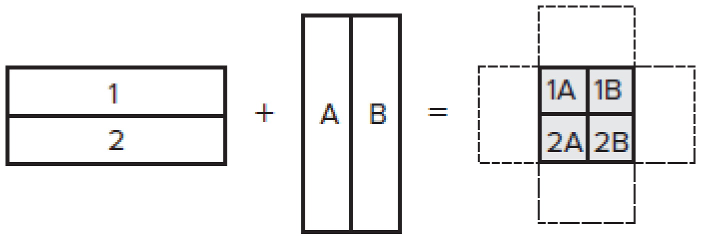



Fig 5: Overlay combines the geometries and attributes from two layers into a single layer. The dashed lines are for illustration only and are not included in the output.

## Feature Type and Overlay

- In practice, the first consideration for overlay is feature type. Overlay operations can take polygon, line, or point layers as the inputs and create an output of a lower-dimension feature type.
- For example, given the inputs of polygon and line layers, the output will be a line layer.
- The layer that may be a point, line, or polygon layer is called the input layer, and the layer that is a polygon layer is called the overlay layer.
- Three common overlay operations:
    - Point-in-polygon,
    - Line-in-polygon, and
    - Polygon-on-polygon.

- In a **point-in-polygon** overlay operation, the same point features in the input layer are included in the output but each point is assigned with attributes of the polygon within which it falls.
- For example, a point-in-polygon overlay can find the association between wildlife locations and vegetation types.

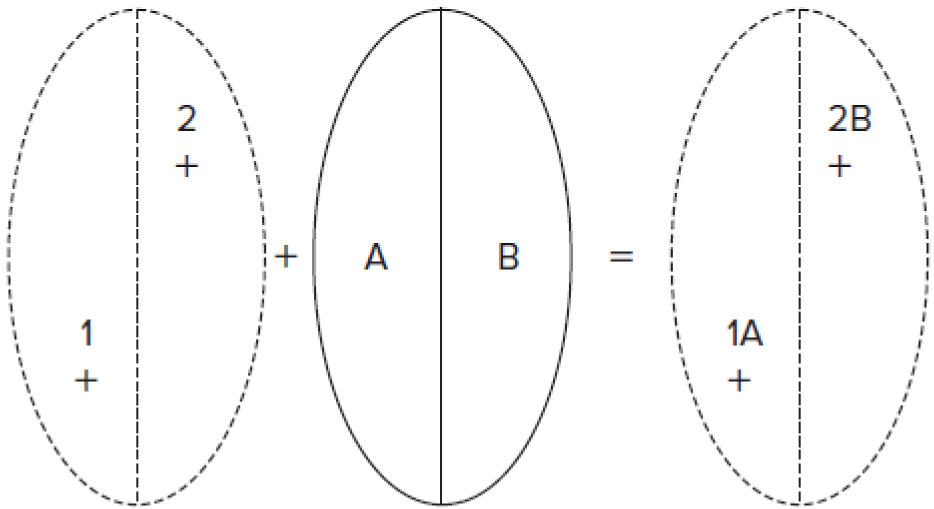



Fig 6: Point-in-polygon overlay. The input is a point layer. The output is also a point layer but has attribute data from the polygon layer.

- In a **line-in-polygon overlay** operation, the output contains the same line features as in the input layer but each line feature is dissected by the polygon boundaries on the overlay layer.
- Thus the output has more line segments than does the input layer. Each line segment on the output combines attributes from the input layer and the underlying polygon.
- For example, a line-in-polygon overlay can find soil data for a proposed road. The input layer includes the proposed road. The overlay layer contains soil polygons.

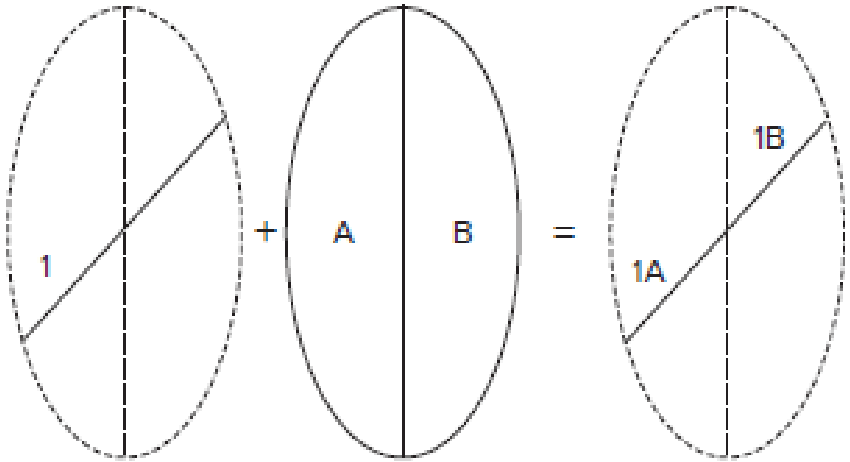



Fig 7: Line-in-polygon overlay. The input is a line layer. The output is also a line layer.

- The most common overlay operation is **polygon-on-polygon**, involving two polygon layers. The output combines the polygon boundaries from the input and overlay layers to create a new set of polygons. Each new polygon carries attributes from both layers, and these attributes differ from those of adjacent polygons.
- For example, a polygon-on-polygon overlay can analyze the association between elevation zones and vegetation types

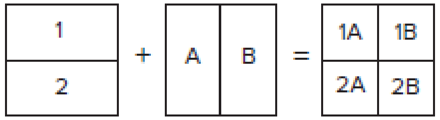



Fig 8: Polygon-on-polygon overlay. In the illustration, the two layers for overlay have the same area extent. The output combines the geometries and attributes from the two layers into a single polygon layer.

**Overlay Methods:**

- The overlay methods are based on the Boolean connectors AND, OR, and XOR. Intersect uses the AND connector. Union uses the OR connector. Symmetrical Difference or Difference uses the XOR connector. Identity or Minus uses the following expression: [(input layer) AND (identity layer)] OR (input layer). The following explains in more detail these four common overlay methods by using two polygon layers as the inputs.
- **Union** preserves all features from the inputs. The area extent of the output combines the area extents of both input layers.
- **Intersect** preserves only those features that falls within the area extent common to the inputs. Intersect is often a preferred method of overlay because any feature on its output has attribute data from both of its inputs.
- For example, a forest management plan may call for an inventory of vegetation types within riparian zones. Intersect will be a more efficient overlay method than Union in this case because the output contains only riparian zones with vegetation types.

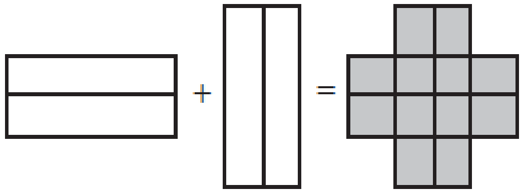



Fig 9: The Union method keeps all the areas of the two input layers in the output.

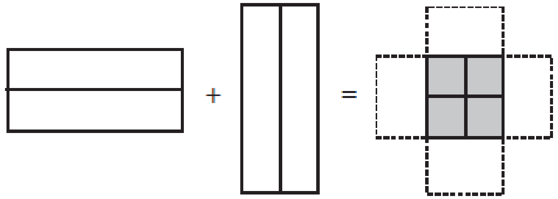



Fig 10: The Intersect method preserves only the area common to the two input layers in the output.

- Symmetrical Difference preserves features that fall within the area extent that is common to only one of the inputs. In other words, Symmetrical Difference is opposite to Intersect in terms of the output's area extent.
- Identity preserves only features that fall within the area extent of the layer defined as the input layer. The other layer is called the identity layer.

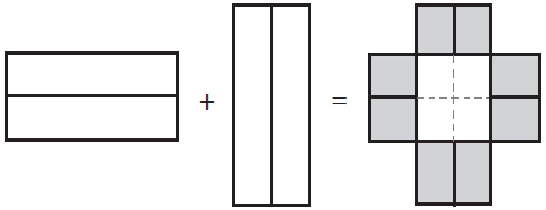


Fig 11: The Symmetrical Difference method preserves areas common to only one of the input layers in the output.

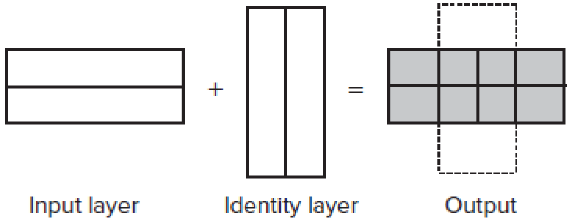



Fig 12: The Identity method produces an output that has the same extent as the input layer. But the output includes the geometry and attributes data from the identity layer

**Applications of Overlay:**

- The overlay methods play a central role in many querying and modeling applications. Suppose an investment company is looking for a land parcel that is zoned commercial, not subject to flooding, and not more than 1 mile from a heavy-duty road. The company can first create the 1-mile road buffer and overlay the buffer zone layer with the zoning and floodplain layers. A subsequent query of the overlay output can select land parcels that satisfy the company's selection criteria.
- A more specific application of overlay is to help solve the areal interpolation problem. Areal interpolation involves transferring known data from one set of polygons (source polygons) to another (target polygons). For example, census tracts may represent source polygons with known populations in each tract from the U.S. Census Bureau, and school districts may represent target polygons with unknown population data. Using overlay the population of the school districts can be calculated using the population given in the census tract.

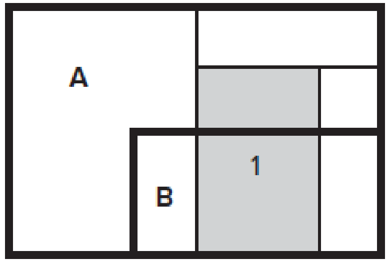



*Fig 13: An example of areal interpolation*

**Distance measurement:**

- Distance measurement refers to measuring straight line (Euclidean) distances between features. Measurements can be made from points in a layer to points in another layer, or from each point in a layer to its nearest point or line in another layer. In both cases, distance measures are stored in a field. Distance measures can be used directly for data analysis.

**Pattern Analysis:**

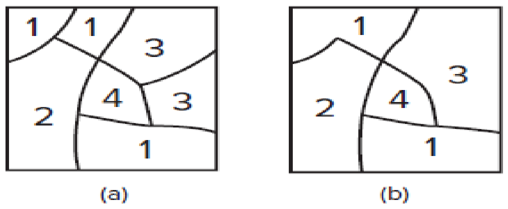
- Pattern analysis is the study of the spatial arrangements of point or polygon features in two dimensional space. Pattern analysis uses distance measurements as inputs and statistics (spatial statistics) for describing the distribution pattern. At the general (global) level, a pattern analysis can reveal if a point distribution pattern is random, dispersed, or clustered
- A classic technique for point pattern analysis, nearest neighbour analysis uses the distance between each point and its closest neighbouring point in a layer to determine if the point pattern is random, regular, or clustered. The nearest neighbour statistic is the ratio (R) of the observed average distance between nearest neighbours (dobs) to the expected average for a hypothetical random distribution (dexp):

$$R = \frac{d_{\text{obs}}}{d_{\text{exp}}}$$

- The R ratio is less than 1 if the point pattern is more clustered than random, and greater than 1 if the point pattern is more dispersed than random.

**Feature Manipulation:**

- Tools are available in a GIS package for manipulating and managing features in one or more feature layers. When a tool involves two layers, the layers must be based on the same coordinate system. Like overlay, these feature tools are often needed for data preprocessing and data analysis; however, unlike overlay, these tools do not combine geometries and attributes from input layers into a single layer.
- Dissolve aggregates features in a feature layer that have the same attribute value or values. For example, we can aggregate roads by highway number or counties by state. An important application of Dissolve is to simplify a classified polygon layer.



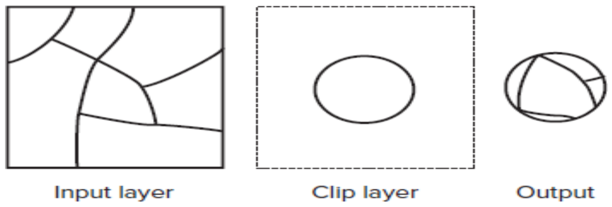




Fig 14: Dissolve removes boundaries of polygons that have the same attribute value in (a) and creates a simplified layer (b)

Fig 15: Clip creates an output that contains only those features of the input layer that fall within the area extent of the clip layer. The output has the same feature type as the input.

- Clip creates a new layer that includes only those features of the input layer, including their attributes that fall within the area extent of the clip layer. Clip is a useful tool, for example, for cutting a map acquired elsewhere to fit a study area.

- Append creates a new layer by piecing together two or more layers, which represent the same feature and have the same attributes.
- For example, Append can put together a layer from four input layers, each corresponding to the area extent of a USGS 7.5-minute quadrangle

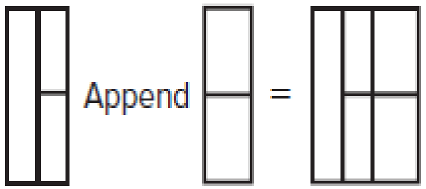



*Fig 16: Append pieces together two adjacent layers into a single layer but does not remove the shared boundary between the layers.*

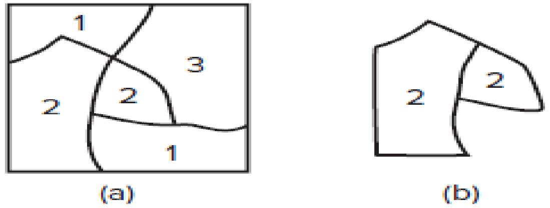



Fig 17: Select creates a new layer (b) with selected features from the input layer (a).

- Select creates a new layer that contains features selected from a user-defined query expression. For example, we can create a layer showing high-canopy closure by selecting stands that have 60 to 80 percent closure from a stand layer.
- Eliminate creates a new layer by removing features that meet a user-defined query expression. For example, Eliminate can implement the minimum mapping unit concept by removing polygons that are smaller than the defined unit in a layer.

- Update uses a “cut and paste” operation to replace the input layer with the update layer and its features. As the name suggests, Update is useful for updating an existing layer with new features in limited areas.

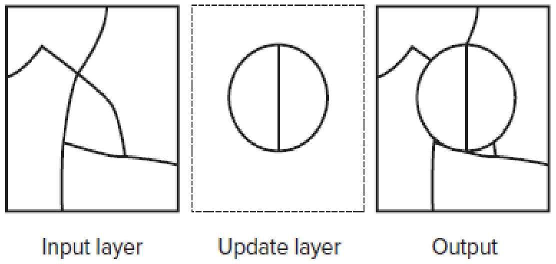



*Fig 18: Update replaces the input layer with the update layer and its features*

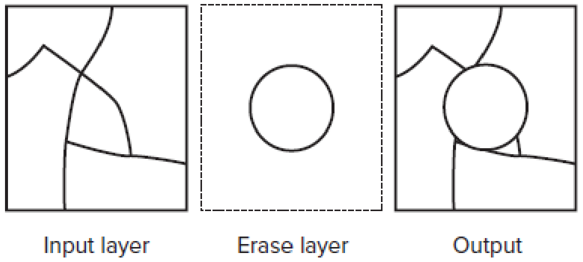



*Fig 19: Erase removes features from the input layer that fall within the area extent of the erase layer*

- Erase removes from the input layer those features that fall within the area extent of the erase layer. Suppose a suitability analysis stipulates that potential sites cannot be within 300 meters of any stream. A stream buffer layer can be used in this case as the erase layer to remove itself from further consideration.
- Split divides the input layer into two or more layers. A split layer, which shows area subunits, is used as the template for dividing the input layer. For example, a national forest can split a stand layer by district so that each district office can have its own layer

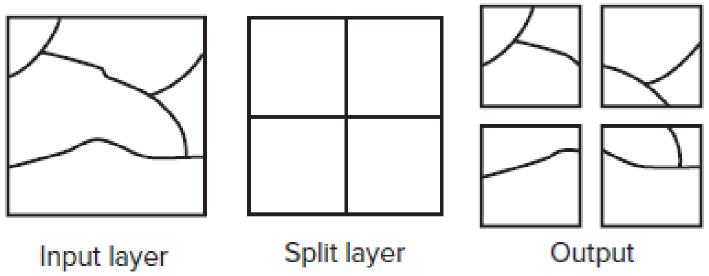



Fig 20: Split uses the geometry of the split layer to divide the input layer into four separate layers.

# RASTER DATA ANALYSIS

- The raster data model uses a regular grid to cover the space and the value in each grid cell to represent the characteristic of a spatial phenomenon at the cell location. This simple data structure of a raster with fixed cell locations not only is computationally efficient, but also facilitates a large variety of data analysis operations. This is why raster data are typically used in geographic information system (GIS), involving heavy computation such as building environmental models.
- In contrast with vector data analysis, which uses points, lines, and polygons, raster data analysis uses cells and rasters. Raster data analysis can be performed at the level of individual cells, or groups of cells, or cells within an entire raster. Some raster data operations use a single raster; others use two or more rasters.

**Data Analysis Environment:**

- Because a raster operation may involve two or more rasters, it is necessary to define the data analysis environment by specifying its area extent and output cell size. The area extent for analysis may correspond to a specific raster, or an area defined by its minimum and maximum x-, y-coordinates, or a combination of rasters. An analysis mask, either a feature layer or a raster, can also determine the area extent for analysis.
- An analysis mask limits analysis to its area coverage. For example, to limit soil erosion analysis to only private lands, we can prepare a mask of either a feature layer showing private lands or a raster separating private lands (e.g., with a cell value of 1) from others (e.g., with a cell value of no data).

**Local Operations:**

- Constituting the core of raster data analysis, local operations are cell-by-cell operations. A local operation can create a new raster from either a single input raster or multiple input rasters. The cell values of the new raster are computed by a function relating the input to the output or are assigned by a classification table.

**Local Operations with a Single Raster:**

- Given a single raster as the input, a local operation computes each cell value in the output raster as a function of the cell value in the input raster at the same location. The function may involve a GIS tool, a mathematical operator, and/or a constant. A large number of mathematical operators are available in a GIS package.

**Reclassification:**

- A local operation, reclassification creates a new raster by classification. Reclassification is also referred to as recoding, or transforming, through lookup tables. Two reclassification methods may be used. The first method is a one-to-one change, meaning that a cell value in the input raster is assigned a new value in the output raster.
- For example, irrigated cropland in a land-use raster is assigned a value of 1 in the output raster.
- The second method assigns a new value to a range of cell values in the input raster.
- For example, cells with population densities between 0 and 25 persons per square mile in a population density raster are assigned a value of 1 in the output raster and so on. An integer raster can be reclassified by either method, but a floating-point raster can only be reclassified by the second method.

- Reclassification serves three main purposes. First, reclassification can create a simplified raster. For example, instead of having continuous slope values, a raster can have 1 for slopes of 0 to 10 percent, 2 for 10 to 20 percent, and so on. Second, reclassification can create a new raster that contains a unique category or value such as slopes of 10 to 20 percent. Third, reclassification can create a new raster that shows the ranking of cell values in the input raster. For example, a reclassified raster can show the ranking of 1 to 5, with 1 being least suitable and 5 being most suitable.

**Local Operations with Multiple Rasters:**

- Local operations with multiple rasters are also referred to as compositing, overlaying, or superimposing maps. Because local operations can work with multiple rasters, they are the equivalent of vector-based overlay operations. A greater variety of local operations have multiple input rasters than have a single input raster.
- Besides mathematical operators that can be used on individual rasters, other measures that are based on the cell values or their frequencies in the input rasters can also be derived and stored in the output raster. Some of these measures are, however, limited to rasters with numeric data.
- Summary statistics, including maximum, minimum, range, sum, mean, median, and standard deviation, are measures that apply to rasters with numeric data

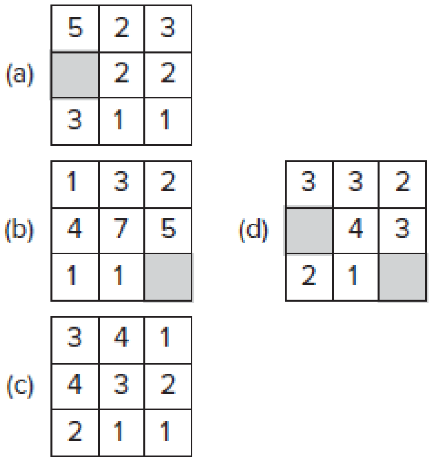



Fig 21: The cell value in (d) is the mean calculated from three input rasters (a, b, and c) in a local operation. The shaded cells have no data.

- For example, above figure shows a local operation that calculates the mean from three input rasters. If a cell contains no data in one of the input rasters, the cell also carries no data in the output raster by default.

- Some local operations do not involve statistics or computation. A local operation called Combine assigns a unique output value to each unique combination of input values. Suppose a slope raster has three cell values (0 to 20 percent, 20 to 40 percent, and greater than 40 percent slope), and an aspect raster has four cell values (north, east, south, and west aspects).

- The Combine operation creates an output raster with a value for each unique combination of slope and aspect, such as 1 for greater than 40 percent slope and the south aspect, 2 for 20 to 40 percent slope and the south aspect, and so on

(a)

| | | |
|---|---|---|
| 3 | 2 | 1 |
| 2 | 1 | 2 |
| 1 | 2 | 3 |

(b)

| | | |
|---|---|---|
| 3 | 2 | 4 |
| 3 | 2 | 4 |
| 2 | 4 | 1 |

(c)

| | | |
|---|---|---|
| 1 | 3 | 6 |
| 2 | 4 | 5 |
| 4 | 5 | 7 |

(d)

| Combine code | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| (slope, aspect) | (3,3) | (2,3) | (2,2) | (1,2) | (2,4) | (1,4) | (3,1) |

Fig 22: Each cell value in (c) represents a unique combination of cell values in (a) and (b). The combination codes and their representations are shown in (d).

**Applications of Local Operations:**

- A change detection study of land cover, for example, can use the unique combinations produced by the Combine operation to trace the change of the land cover type. The land cover databases (2001, 2006, and 2011) from the U.S. Geological Survey (USGS)) are ideal for such a change detection study.
- The Revised Universal Soil Loss Equation (RUSLE) uses six environmental factors in the equation: A = R K L S C P,

where A is the predicted soil loss, R is the rainfall–runoff erosivity factor, K is the soil erodibility factor, L is the slope length factor, S is the slope steepness factor, C is the crop management factor, and P is the support practice factor.

- With each factor prepared as an input raster, we can multiply the rasters in a local operation to produce the output raster of predicted soil loss.

**Neighborhood Operations:**

- A neighborhood operation, also called a focal operation, involves a focal cell and a set of its surrounding cells. The surrounding cells are chosen for their distance and/or directional relationship to the focal cell. A required parameter for neighborhood operations is the type of neighborhood. Common neighbourhoods include rectangles, circles, annuluses, and wedges.

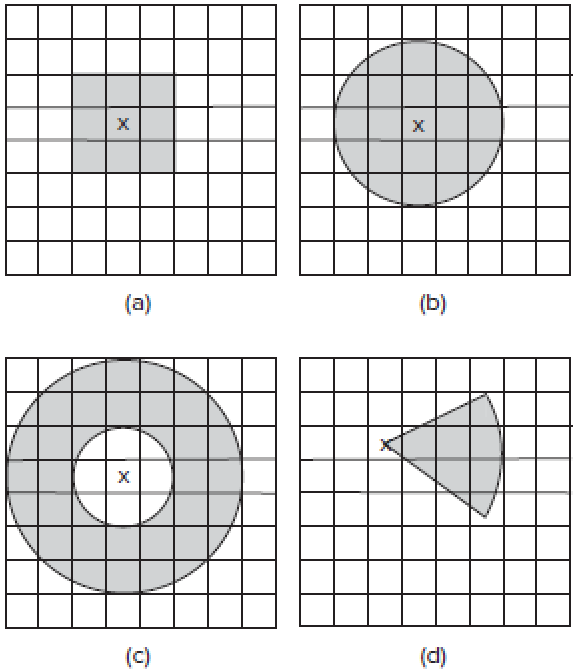



Fig 23: Four common neighborhood types: rectangle (a),circle (b), annulus (c), and wedge (d).
The cell marked with an x is the focal cell.

- A rectangle is defined by its width and height in cells, such as a 3-by-3 area centered at the focal cell. A circle extends from the focal cell with a specified radius. An annulus or doughnut-shaped neighborhood consists of the ring area between a smaller circle and a larger circle centered at the focal cell. And a wedge consists of a piece of a circle centered at the focal cell.

**Applications of Neighborhood Operations:**

- An important application of neighborhood operations is data simplification. The moving average method, for instance, reduces the level of cell value fluctuation in the input raster. The method typically uses a 3-by-3 or a 5-by-5 rectangle as the neighborhood. As the neighborhood is moved from one focal cell to another, the average of cell values within the neighborhood is computed and assigned to the focal cell. The output raster of moving averages represents a generalization of the original cell values.

(a)

| | | | | |
|---|---|---|---|---|
| 1 | 2 | 2 | 2 | 2 |
| 1 | 2 | 2 | 2 | 3 |
| 1 | 2 | 1 | 3 | 3 |
| 2 | 2 | 2 | 3 | 3 |
| 2 | 2 | 2 | 2 | 3 |

(b)

| | | |
|---|---|---|
| 1.56 | 2.00 | 2.22 |
| 1.67 | 2.11 | 2.44 |
| 1.67 | 2.11 | 2.44 |

- Neighborhood operations are common in image processing. These operations are variously called filtering, convolution, or moving window operations for spatial feature manipulation

Fig 24: The cell values in (*b*) are the neighborhood means of the shaded cells in (*a*) using a 3-by-3 neighborhood.

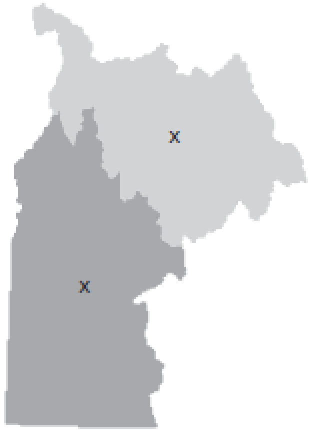
For example, 1.56 in the output raster is calculated from (1 + 2 + 2 + 1 + 2 + 2 + 1 + 2 + 1)/9.

**Zonal Operations:**

- A zonal operation works with groups of cells of same values or like features. These groups are called zones. Zones may be contiguous or noncontiguous.
- A contiguous zone includes cells that are spatially connected, whereas a non-contiguous zone includes separate regions of cells. A watershed raster is an example of a contiguous zone, in which cells that belong to the same watershed are spatially connected.
- A land use raster is an example of a non-contiguous zone, in which one type of land use may appear in different parts of the raster.

**Zonal Statistics**

- A zonal operation may work with a single raster or two rasters. Given a single input raster, zonal operations measure the geometry of each zone in the raster, such as area, perimeter, thickness, and centroid (Figure below) The area is the sum of the cells that fall within the zone times the cell size.



| Zone | Area | Perimeter | Thickness |
|---|---|---|---|
| 1 | 36,224 | 1708 | 77.6 |
| 2 | 48,268 | 1464 | 77.4 |

Fig 25: Thickness and centroid for two large watersheds (zones). Area is measured in square kilometers, and perimeter and thickness are measured in kilometers.The centroid of each zone is marked with an x.

- Given two rasters in a zonal operation, one input raster and one zonal raster, a zonal operation produces an output raster, which summarize the cell values in the input raster for each zone in the zonal raster.
- The summary statistics and measures include area, minimum, maximum, sum, range, mean, standard deviation, median, majority, minority, and variety.

- Figure below shows a zonal operation of computing the mean by zone. Figure *b* is the zonal raster with three zones, Figure *a* is the input raster, and Figure *c* is the output raster.

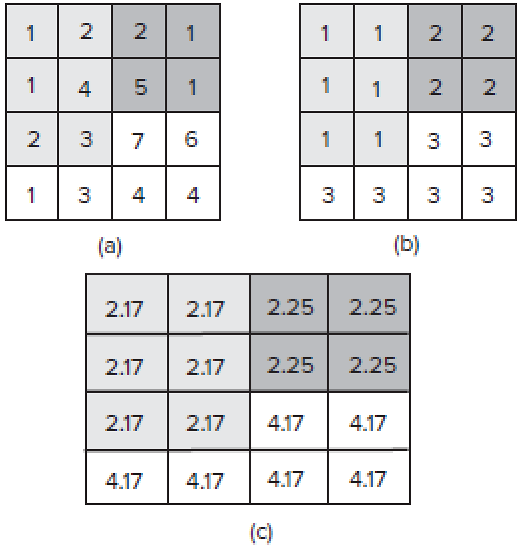



Fig 26: The cell values in (c) are the zonal means derived from an input raster (a) and a zonal raster (b). For example, 2.17 is the mean of {1, 1, 2, 2, 4, 3} for zone 1.

**Applications of Zonal Operations**

- Measures of zonal geometry such as area, perimeter, thickness, and centroid are particularly useful for studies of landscape ecology (Forman and Godron 1986; McGarigal and Marks 1994). Many other geometric measures can be derived from area and perimeter.
- For example, the perimeter-area ratio (i.e., zonalperimeter/zonalarea) is a simple measure of shape complexity used in landscape ecology.
- Zonal operations with two rasters can generate useful descriptive statistics for comparison purposes. For example, to compare topographic characteristics of different soil textures, we can use a soil raster that contains the categories of sand, loam, and clay as the zonal raster and slope, aspect, and elevation as the input rasters. By running a series of zonal operations, we can summarize the slope, aspect, and elevation characteristics associated with the three soil textures.

**Physical distance measure operations:**

- In a GIS project, distances may be expressed as physical distances or cost distances. The **physical distance** measures the straight-line or Euclidean distance, whereas the **cost distance** measures the cost for traversing the physical distance. The distinction between the two types of distance measures is important in real-world applications.
- A truck driver, for example, is more interested in the time or the fuel cost for covering a route than in its physical distance. The cost distance in this case is based on not only the physical distance but also the speed limit and road condition.
- Physical distance measure operations calculate straight-line distances away from cells designated as the source cells. For example, to get the distance between cells (1, 1) and (3, 3) in Figure below, we can use the following formula:

$$\text{cell size} \times \sqrt{(3-1)^2 + (3-1)^2}$$

or cell size × 2.828. If the cell size were 30 meters, the distance would be 84.84 meters.

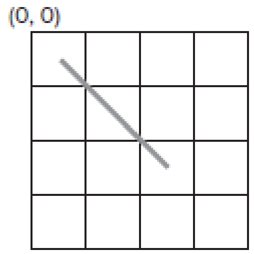



Fig 27: A straight-line distance is measured from a cell center to another cell center. This illustration shows the straight-line distance between cell (1,1) and cell (3,3).

**Applications of Physical Distance:**

- Measure Operations Like buffering around vector-based features, physical distance measure operations have many applications.
- For example, we can create equal-interval distance zones from a stream network or regional fault lines.
- Another example is the use of distance measure operations as tools for implementing a model, such as the potential nesting habitat model of greater sandhill cranes in northwestern Minnesota, developed by Herr and Queen (1993). Based on continuous distance zones measured from undisturbed vegetation, roads, buildings, and agricultural land, the model categorizes potentially suitable nesting vegetation as optimal, suboptimal, marginal, and unsuitable.

**OTHER RASTER DATA OPERATIONS:**

- Local, neighborhood, zonal, and distance measure operations cover the majority of raster data operations. Some operations, however, do not fit well into the classification.

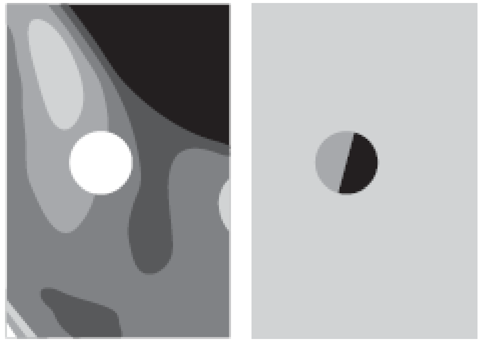
**Raster Data Management**

- In a GIS project, we often need to clip, or combine, raster data downloaded from the Internet to fit the study area. To clip a raster, we can specify an analysis mask or the minimum and maximum *x*-, *y*-coordinates of a rectangular area for the analysis environment and then use the larger raster as the input (Figure below). Mosaic is a tool that can combine multiple input rasters into a single raster. If the input rasters overlap, a GIS package typically provides options for filling in the cell values in the overlapping areas.


Fig 28: An analysis mask (*b*) is used to clip an input raster (*a*). The output raster is (*c*), which has the same area extent as the analysis mask.

**Raster Data Extraction**

- Raster data extraction creates a new raster by extracting data from an existing raster. The operation is similar to raster data query. A data set, a graphic object, or a query expression can be used to define the area to be extracted. If the data set is a point feature layer, the extraction tool extracts values at the point locations (e.g., using bilinear interpolation).
- A graphic object for raster data extraction can be a rectangle, a set of points, a circle, or a polygon. The object is input in *x*-, *y*-coordinates. For example, a circle can be entered with a pair of *x*-, *y*-coordinates for its center and a length for its radius.



(a) (b)

Fig 29: A circle, shown in white, is used to extract cell values from the input raster (*a*). The output (*b*) has the same area extent as the input raster but has no data outside the circular area. (To highlight the contrast, *b* uses a different symbology than *a*.)

## Raster Data Generalization

- Several operations can generalize or simplify raster data. One such operation is resampling, which can build different pyramid levels (different resolutions) for a large raster data set.
- Aggregate is similar to a resampling technique in that it also creates an output raster that has a larger cell size (i.e.,a lower resolution) than the input. But, instead of using nearest neighbor, bilinear interpolation, or cubic convolution, Aggregate calculates each output cell value as the mean, median, sum, minimum, or maximum of the input cells that fall within the output cell.

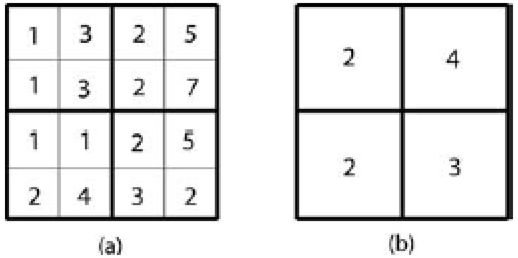



Fig 30: An Aggregate operation creates a lower-resolution raster (*b*) from the input (*a*). The operation uses the mean statistic and a factor of 2 (i.e., a cell in *b* covers 2-by-2 cells in *a*). For example, the cell value of 4 in (*b*) is the mean of {2, 2, 5, 7} in (*a*).

- Some data generalization operations are based on zones, or groups of cells of the same value. For example, ArcGIS has a tool called RegionGroup, which identifies for each cell in the output raster the zone that the cell is connected to (Figure below).
- Generalizing or simplifying the cell values of a raster can be useful for certain applications. For example, a raster derived from a satellite image or LiDAR (light detection and ranging) data usually has a high degree of local variations. These local variations can become unnecessary noises. To remove them, we can use Aggregate or a resampling technique.

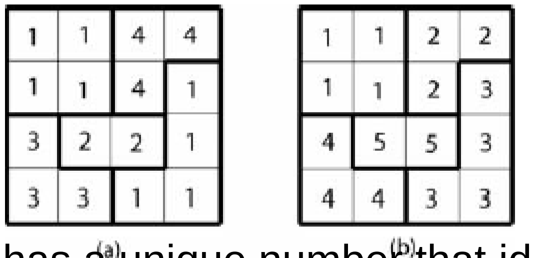



Fig 31: Each cell in the output (*b*) has a unique number that identifies the connected region to which it belongs in the input (*a*). For example, the connected region that has the same cell value of 3 in (*a*) has a unique number of 4 in (*b*).