Chapter 7

Geospatial Analysis I: Vector Operations

In Chapter 6 "Data Characteristics and Visualization", we discussed different ways to query, classify, and summarize information in attribute tables. These methods are indispensable for understanding the basic quantitative and qualitative trends of a dataset. However, they don't take particular advantage of the greatest strength of a geographic information system (GIS), notably the explicit spatial relationships. Spatial analysis is a fundamental component of a GIS that allows for an in-depth study of the topological and geometric properties of a dataset or datasets. In this chapter, we discuss the basic spatial analysis techniques for vector datasets.

7.1 Single Layer Analysis

LEARNING OBJECTIVE

1. The objective of this section is to become familiar with concepts and terms related to the variety of single overlay analysis techniques available to analyze and manipulate the spatial attributes of a vector feature dataset.

As the name suggests, single layer analyses are those that are undertaken on an individual feature dataset. **Buffering**[1] is the process of creating an output polygon layer containing a zone (or zones) of a specified width around an input point, line, or polygon feature. Buffers are particularly suited for determining the area of influence around features of interest. **Geoprocessing**[2] is a suite of tools provided by many geographic information system (GIS) software packages that allow the user to automate many of the mundane tasks associated with manipulating GIS data. Geoprocessing usually involves the input of one or more feature datasets, followed by a spatially explicit analysis, and resulting in an output feature dataset.

Buffering

Buffers are common vector analysis tools used to address questions of proximity in a GIS and can be used on points, lines, or polygons (<u>Figure 7.1 "Buffers around Red Point, Line, and Polygon Features"</u>). For instance, suppose that a natural resource manager wants to ensure that no areas are disturbed within 1,000 feet of breeding habitat for the federally endangered Delhi Sands flower-loving fly (*Rhaphiomidas terminatus abdominalis*). This species is found only in the few remaining Delhi Sands soil formations of the western United States. To accomplish this task, a 1,000-foot protection zone (buffer) could be created around all the observed point locations of the species. Alternatively, the manager may decide that there is not enough point-specific location information related to this rare species and decide to protect all Delhi Sands soil formations. In this case, he or she could create a 1,000-foot buffer around all polygons labeled as "Delhi Sands" on a soil formations dataset. In either case, the use of buffers provides a quick-and-easy tool for determining which areas are to be maintained as preserved habitat for the endangered fly.

1. Placing a region of specified width around a point, line, or polygon.

2. Any operation used to manipulate spatial data.

*Figure 7.1 Buffers around Red Point, Line, and Polygon Features*
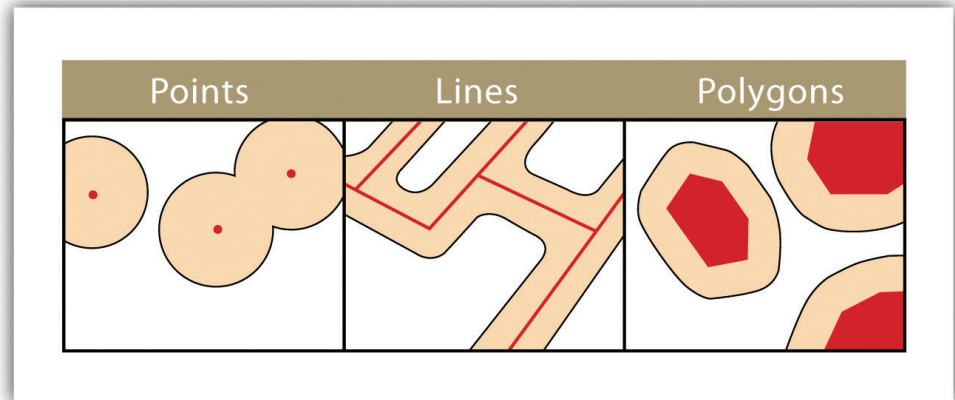
Several buffering options are available to refine the output. For example, the buffer tool will typically buffer only selected features. If no features are selected, all features will be buffered. Two primary types of buffers are available to the GIS users: constant width and variable width. **Constant width buffers**[3] require users to input a value by which features are buffered (Figure 7.1 "Buffers around Red Point, Line, and Polygon Features"), such as is seen in the examples in the preceding paragraph. **Variable width buffers**[4], on the other hand, call on a premade buffer field within the attribute table to determine the buffer width for each specific feature in the dataset (Figure 7.2 "Additional Buffer Options around Red Features: (a) Variable Width Buffers, (b) Multiple Ring Buffers, (c) Doughnut Buffer, (d) Setback Buffer, (e) Nondissolved Buffer, (f) Dissolved Buffer").

In addition, users can choose to dissolve or not dissolve the boundaries between overlapping, coincident buffer areas. **Multiple ring buffers**[5] can be made such that a series of concentric buffer zones (much like an archery target) are created around the originating feature at user-specified distances (Figure 7.2 "Additional Buffer Options around Red Features: (a) Variable Width Buffers, (b) Multiple Ring Buffers, (c) Doughnut Buffer, (d) Setback Buffer, (e) Nondissolved Buffer, (f) Dissolved Buffer"). In the case of polygon layers, buffers can be created that include the originating polygon feature as part of the buffer or they be created as a **doughnut buffer**[6] that excludes the input polygon area. **Setback buffers**[7] are similar to doughnut buffers; however, they only buffer the area inside of the polygon boundary. Linear features can be buffered on both sides of the line, only on the left, or only on the right. Linear features can also be buffered so that the end points of the line are rounded (ending in a half-circle) or flat (ending in a rectangle).

3. Regions of constant width around points, lines, or polygons.

4. Regions of variable width around points, lines, or polygons.

5. Mulitple concentric regions of a specified width around points, lines, or polygons.

6. A buffer around a polygon feature that does not include the area inside the buffered polygon.

7. A buffer around a polygon feature that only extends inside of the polygon boundary.

Figure 7.2 *Additional Buffer Options around Red Features: (a) Variable Width Buffers, (b) Multiple Ring Buffers, (c) Doughnut Buffer, (d) Setback Buffer, (e) Nondissolved Buffer, (f) Dissolved Buffer*
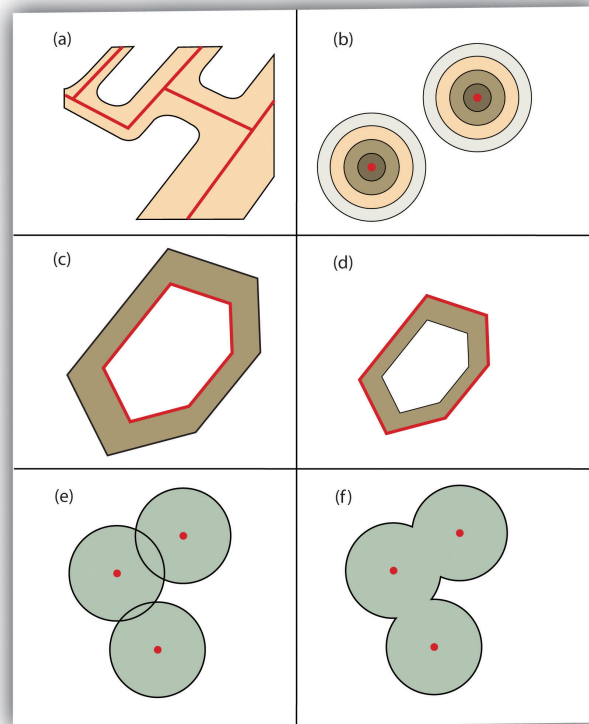
Geoprocessing Operations

"Geoprocessing" is a loaded term in the field of GIS. The term can (and should) be widely applied to any attempt to manipulate GIS data. However, the term came into common usage due to its application to a somewhat arbitrary suite of single layer and multiple layer analytical techniques in the Geoprocessing Wizard of ESRI's ArcView software package in the mid-1990s. Regardless, the suite of geoprocessing tools available in a GIS greatly expand and simplify many of the management and manipulation processes associated with vector feature datasets. The primary use of these tools is to automate the repetitive preprocessing needs of typical spatial analyses and to assemble exact graphical representations for subsequent analysis and/or inclusion in presentations and final mapping products. The union, intersect, symmetrical difference, and identity overlay methods discussed in Section 7.2.2 "Other Multilayer Geoprocessing Options" are often used in conjunction with these geoprocessing tools. The following represents the most common geoprocessing tools.

The **dissolve**[8] operation combines adjacent polygon features in a single feature dataset based on a single predetermined attribute. For example, part (a) of <u>Figure 7.3 "Single Layer Geoprocessing Functions"</u> shows the boundaries of seven different parcels of land, owned by four different families (labeled 1 through 4). The dissolve tool automatically combines all adjacent features with the same attribute values. The result is an output layer with the same extent as the original but without all of the unnecessary, intervening line segments. The dissolved output layer is much easier to visually interpret when the map is classified according to the dissolved field.

The **append**[9] operation creates an output polygon layer by combining the spatial extent of two or more layers (part (d) of <u>Figure 7.3 "Single Layer Geoprocessing Functions"</u>). For use with point, line, and polygon datasets, the output layer will be the same feature type as the input layers (which must each be the same feature type as well). Unlike the dissolve tool, append does not remove the boundary lines between appended layers (in the case of lines and polygons). Therefore, it is often useful to perform a dissolve after the use of the append tool to remove these potentially unnecessary dividing lines. Append is frequently used to mosaic data layers, such as digital US Geological Survey (USGS) 7.5-minute topographic maps, to create a single map for analysis and/or display.

The **select**[10] operation creates an output layer based on a user-defined query that selects particular features from the input layer (part (f) of <u>Figure 7.3 "Single Layer Geoprocessing Functions"</u>). The output layer contains only those features that are selected during the query. For example, a city planner may choose to perform a select on all areas that are zoned "residential" so he or she can quickly assess which areas in town are suitable for a proposed housing development.

Finally, the **merge**[11] operation combines features within a point, line, or polygon layer into a single feature with identical attribute information. Often, the original features will have different values for a given attribute. In this case, the first attribute encountered is carried over into the attribute table, and the remaining attributes are lost. This operation is particularly useful when polygons are found to be unintentionally overlapping. Merge will conveniently combine these features into a single entity.

8. A geoprocessing technique that removes the boundary between adjacent polygons with identical values.

9. A geoprocessing technique that combines adjacent polygon datasets into a single dataset.

10. To define a subset of the larger set of data points or locales.

11. To combine adjacent or overlapping spatial features into a single feature.

Figure 7.3   *Single Layer Geoprocessing Functions*
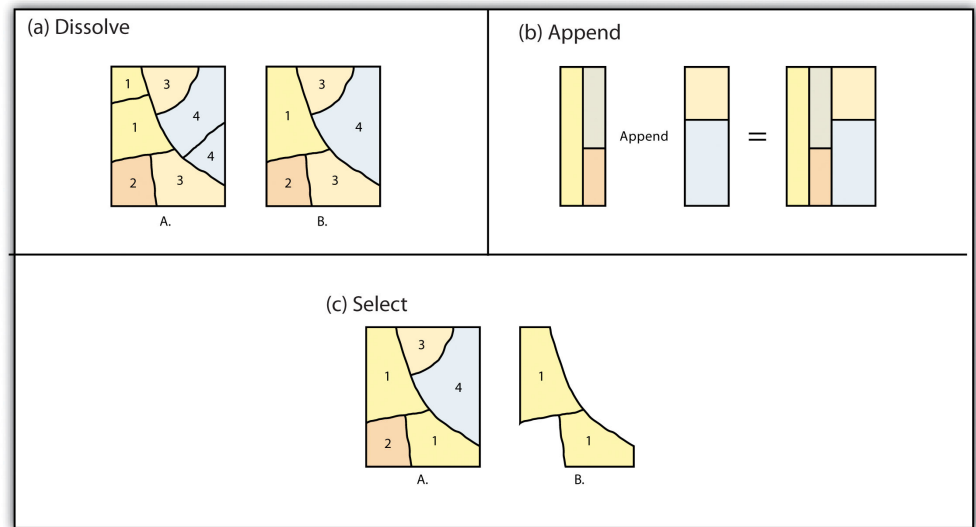
KEY TAKEAWAYS

- Buffers are frequently used to create zones of a specified width around points, lines, and polygons.
- Vector buffering options include constant or variable widths, multiple rings, doughnuts, setbacks, and dissolve.
- Common single layer geoprocessing operations on vector layers include dissolve, merge, append, and select.

EXERCISES

1. List and describe the various buffering options available in a GIS.
2. Why might you use the various geoprocessing operations to answer spatial questions related to your particular field of study?

7.2 Multiple Layer Analysis

> **LEARNING OBJECTIVE**
>
> 1. The objective of this section is to become familiar with concepts and terms related to the implementation of basic multiple layer operations and methodologies used on vector feature datasets.

Among the most powerful and commonly used tools in a geographic information system (GIS) is the overlay of cartographic information. In a GIS, an **overlay**[12] is the process of taking two or more different thematic maps of the same area and placing them on top of one another to form a new map (Figure 7.4 "A Map Overlay Combining Information from Point, Line, and Polygon Vector Layers, as Well as Raster Layers"). Inherent in this process, the overlay function combines not only the spatial features of the dataset but also the attribute information as well.

Figure 7.4 *A Map Overlay Combining Information from Point, Line, and Polygon Vector Layers, as Well as Raster Layers*
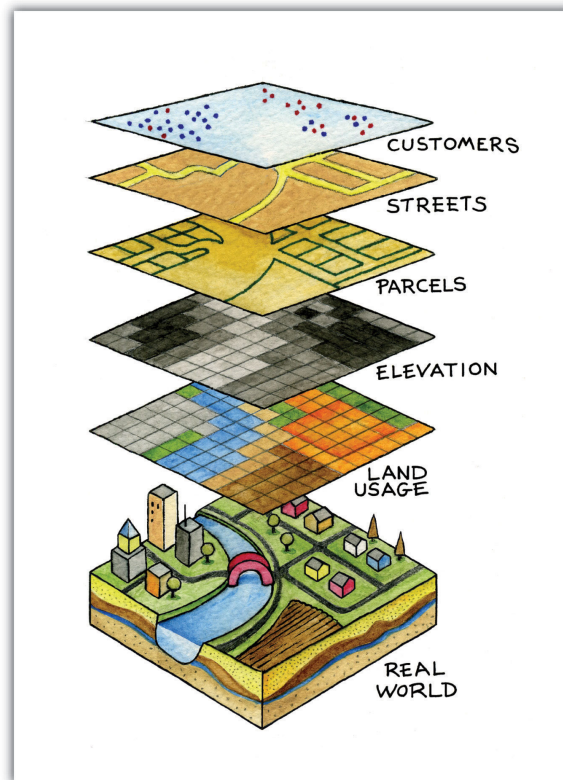
12. The process of taking two or more different thematic maps of the same area and placing them on top of one another to form a new map.

A common example used to illustrate the overlay process is, "Where is the best place to put a mall?" Imagine you are a corporate bigwig and are tasked with determining where your company's next shopping mall will be placed. How would you attack this problem? With a GIS at your command, answering such spatial questions begins with amassing and overlaying pertinent spatial data layers. For example, you may first want to determine what areas can support the mall by accumulating information on which land parcels are for sale and which are zoned for commercial development. After collecting and overlaying the baseline information on available development zones, you can begin to determine which areas offer the most economic opportunity by collecting regional information on average household income, population density, location of proximal shopping centers, local buying habits, and more. Next, you may want to collect information on restrictions or roadblocks to development such as the cost of land, cost to develop the land, community response to development, adequacy of transportation corridors to and from the proposed mall, tax rates, and so forth. Indeed, simply collecting and overlaying spatial datasets provides a valuable tool for visualizing and selecting the optimal site for such a business endeavor.
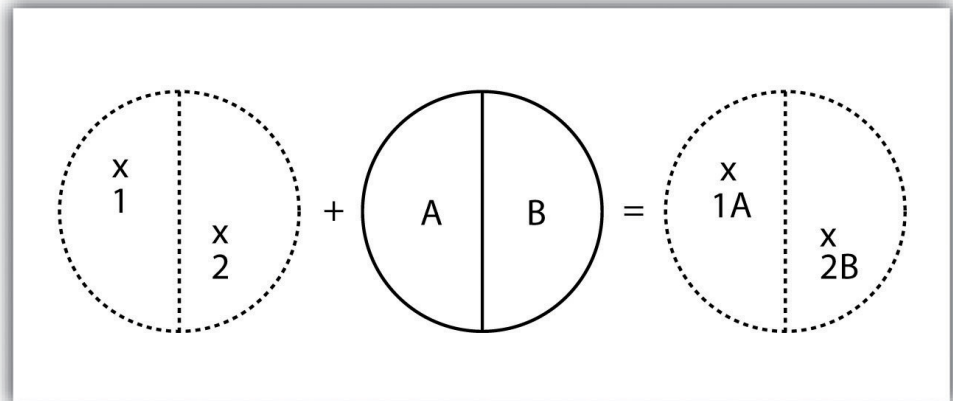
Overlay Operations

Several basic overlay processes are available in a GIS for vector datasets: point-in-polygon, polygon-on-point, line-on-line, line-in-polygon, polygon-on-line, and polygon-on-polygon. As you may be able to divine from the names, one of the overlay dataset must always be a line or polygon layer, while the second may be point, line, or polygon. The new layer produced following the overlay operation is termed the "output" layer.

The **point-in-polygon overlay**[13] operation requires a point input layer and a polygon overlay layer. Upon performing this operation, a new output point layer is returned that includes all the points that occur within the spatial extent of the overlay (Figure 7.4 "A Map Overlay Combining Information from Point, Line, and Polygon Vector Layers, as Well as Raster Layers"). In addition, all the points in the output layer contain their original attribute information as well as the attribute information from the overlay. For example, suppose you were tasked with determining if an endangered species residing in a national park was found primarily in a particular vegetation community. The first step would be to acquire the point occurrence locales for the species in question, plus a polygon overlay layer showing the vegetation communities within the national park boundary. Upon performing the point-in-polygon overlay operation, a new point file is created that contains all the points that occur within the national park. The attribute table of this output point file would also contain information about the vegetation communities being utilized by the species at the time of observation. A quick scan of this output layer and its attribute table would allow you to determine where the

13. An overlay technique that creates an output point layer that includes all the points occurring within the spatial extent of the overlay layer.

species was found in the park and to review the vegetation communities in which it occurred. This process would enable park employees to make informed management decisions regarding which onsite habitats to protect to ensure continued site utilization by the species.

Figure 7.5   Point-in-Polygon Overlay

As its name suggests, the **polygon-on-point overlay**[14] operation is the opposite of the point-in-polygon operation. In this case, the polygon layer is the input, while the point layer is the overlay. The polygon features that overlay these points are selected and subsequently preserved in the output layer. For example, given a point dataset containing the locales of some type of crime and a polygon dataset representing city blocks, a polygon-on-point overlay operation would allow police to select the city blocks in which crimes have been known to occur and hence determine those locations where an increased police presence may be warranted.

Figure 7.6   Polygon-on-Point Overlay
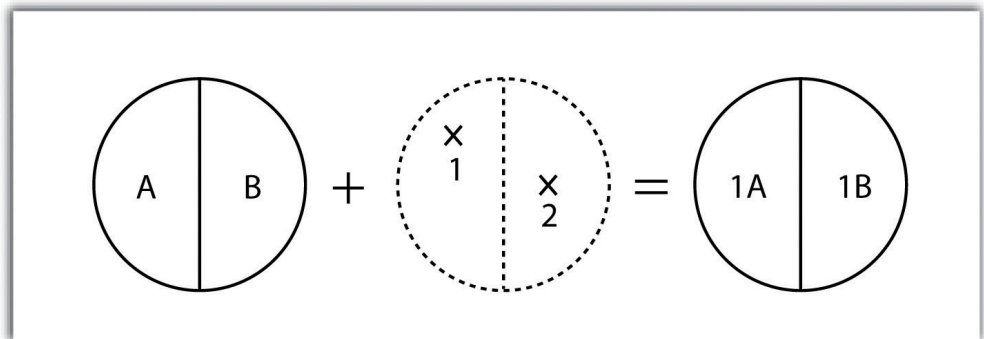
14. An overlay technique that creates a polygon layer from those input polygons that overlay features in a point layer.

A **line-on-line overlay**[15] operation requires line features for both the input and overlay layer. The output from this operation is a point or points located precisely at the intersection(s) of the two linear datasets (Figure 7.7 "Line-on-Line Overlay"). For example, a linear feature dataset containing railroad tracks may be overlain on linear road network. The resulting point dataset contains all the locales of the railroad crossings over a town's road network. The attribute table for this railroad crossing point dataset would contain information on both the railroad and the road over which it passed.

Figure 7.7 Line-on-Line Overlay
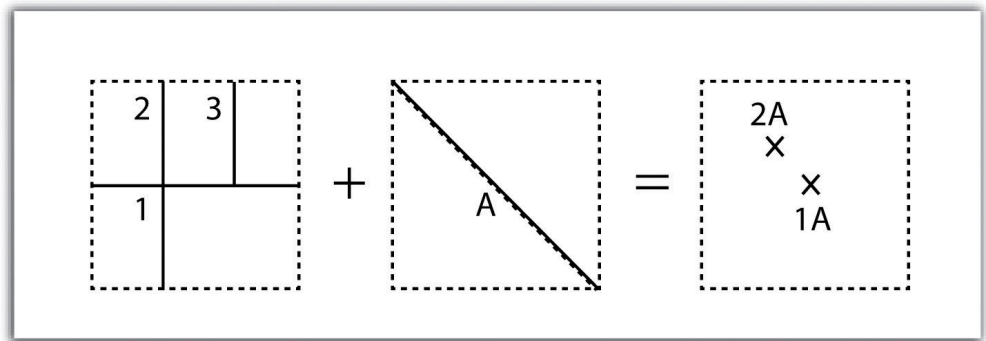
The **line-in-polygon overlay**[16] operation is similar to the point-in-polygon overlay, with that obvious exception that a line input layer is used instead of a point input layer. In this case, each line that has any part of its extent within the overlay polygon layer will be included in the output line layer, although these lines will be truncated at the boundary of the overlay (Figure 7.9 "Polygon-on-Line Overlay"). For example, a line-in-polygon overlay can take an input layer of interstate line segments and a polygon overlay representing city boundaries and produce a linear output layer of highway segments that fall within the city boundary. The attribute table for the output interstate line segment will contain information on the interstate name as well as the city through which they pass.
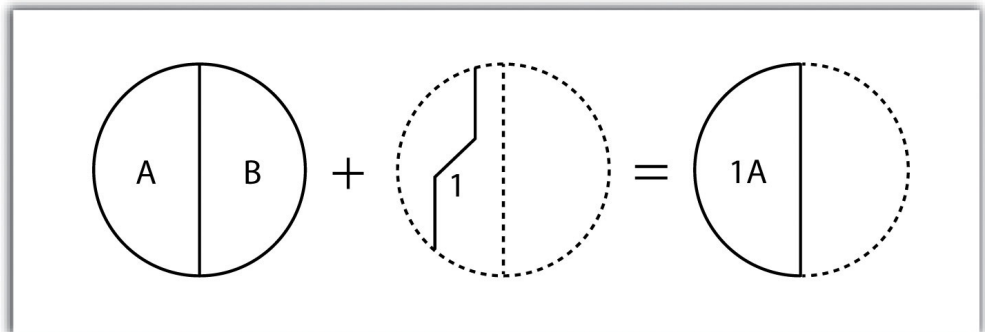
15. An overlay technique in which output from this operation is a point(s) located at the intersection(s) of the two linear datasets.

16. An overlay technique in which each line that has any part of its extent within the overlay polygon layer will be included in an output line layer.
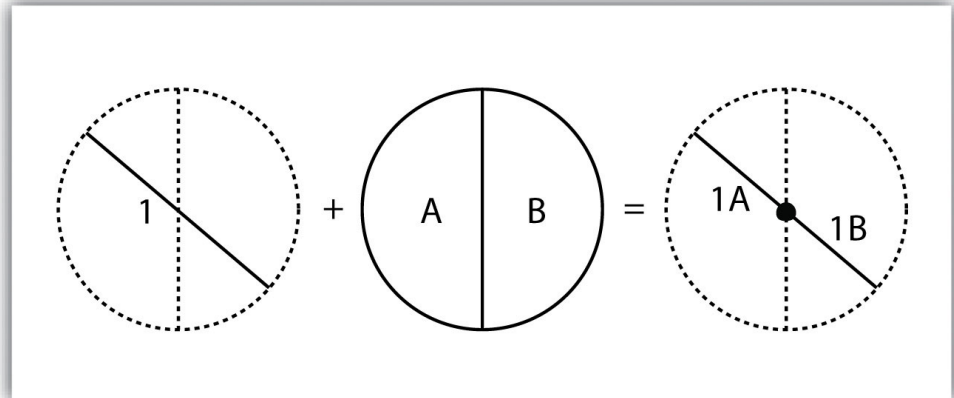
The **polygon-on-line overlay**[17] operation is the opposite of the line-in-polygon operation. In this case, the polygon layer is the input, while the line layer is the overlay. The polygon features that overlay these lines are selected and subsequently preserved in the output layer. For example, given a layer containing the path of a series of telephone poles/wires and a polygon map contain city parcels, a polygon-on-line overlay operation would allow a land assessor to select those parcels containing overhead telephone wires.

Figure 7.9 Polygon-on-Line Overlay

17. An overlay technique in which polygon features that overlay lines are selected and subsequently preserved in an output layer.

18. An overlay technique in which a polygon input and overlay layers are combined to create an output polygon layer with the extent of the overlay.

Finally, the **polygon-in-polygon overlay**[18] operation employs a polygon input and a polygon overlay. This is the most commonly used overlay operation. Using this method, the polygon input and overlay layers are combined to create an output polygon layer with the extent of the overlay. The attribute table will contain spatial data and attribute information from both the input and overlay layers (Figure 7.10 "Polygon-in-Polygon Overlay"). For example, you may choose an input polygon layer of soil types with an overlay of agricultural fields within a given county. The

output polygon layer would contain information on both the location of agricultural fields and soil types throughout the county.

Figure 7.10 *Polygon-in-Polygon Overlay*
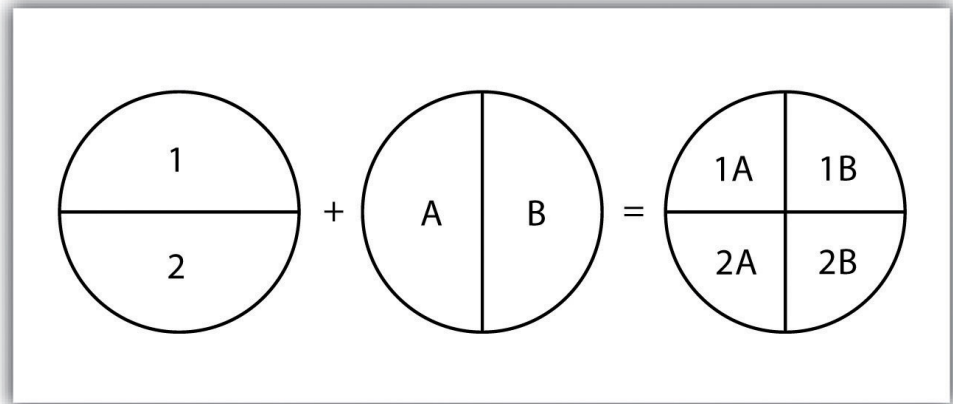
The overlay operations discussed previously assume that the user desires the overlain layers to be combined. This is not always the case. Overlay methods can be more complex than that and therefore employ the basic Boolean operators: AND, OR, and XOR (see Section 6.1.2 "Measures of Central Tendency"). Depending on which operator(s) are utilized, the overlay method employed will result in an intersection, union, symmetrical difference, or identity.

Specifically, the **union**[19] overlay method employs the OR operator. A union can be used only in the case of two polygon input layers. It preserves all features, attribute information, and spatial extents from both input layers (part (a) of Figure 7.11 "Vector Overlay Methods "). This overlay method is based on the polygon-in-polygon operation described in Section 7.1.1 "Buffering".

Alternatively, the **intersection**[20] overlay method employs the AND operator. An intersection requires a polygon overlay, but can accept a point, line, or polygon input. The output layer covers the spatial extent of the overlay and contains features and attributes from both the input and overlay (part (b) of Figure 7.11 "Vector Overlay Methods ").

The **symmetrical difference**[21] overlay method employs the XOR operator, which results in the opposite output as an intersection. This method requires both input layers to be polygons. The output polygon layer produced by the symmetrical difference method represents those areas common to only one of the feature datasets (part (c) of Figure 7.11 "Vector Overlay Methods ").

19. An overlay method that preserves all features, attribute information, and spatial extents from an input layer.

20. An overlay method that contains common features and attributes from both the input and overlay layers.

21. An overlay method that contains those areas common to only one of the feature datasets.

In addition to these simple operations, the **identity**[22] (also referred to as "minus") overlay method creates an output layer with the spatial extent of the input layer (part (d) of Figure 7.11 "Vector Overlay Methods ") but includes attribute information from the overlay (referred to as the "identity" layer, in this case). The input layer can be points, lines, or polygons. The identity layer must be a polygon dataset.

Figure 7.11   *Vector Overlay Methods*
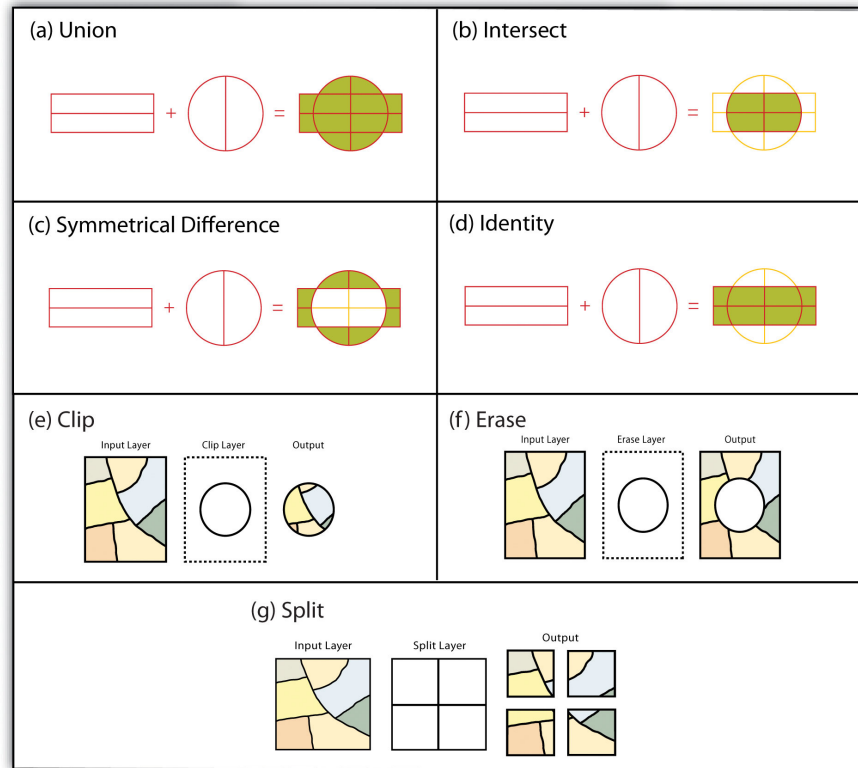
Other Multilayer Geoprocessing Options

In addition to the aforementioned vector overlay methods, other common multiple layer geoprocessing options are available to the user. These included the clip, erase, and split tools. The **clip**[23] geoprocessing operation is used to extract those features from an input point, line, or polygon layer that falls within the spatial extent of the clip layer (part (e) of Figure 7.11 "Vector Overlay Methods "). Following the clip, all attributes from the preserved portion of the input layer are included in the output. If any features are selected during this process, only those selected features within the clip boundary will be included in the output. For example, the clip tool could be used to clip the extent of a river floodplain by the extent of a county boundary. This would provide county managers with insight into which portions of the floodplain

22. An overlay method that creates an output layer with the spatial extent of the input layer but includes attribute information from an overlay.

23. A geoprocessing operation that extracts those features from an input point, line, or polygon layer that falls within the spatial extent of a clip layer.

they are responsible to maintain. This is similar to the intersect overlay method; however, the attribute information associated with the clip layer is not carried into the output layer following the overlay.

The **erase**[24] geoprocessing operation is essentially the opposite of a clip. Whereas the clip tool preserves areas within an input layer, the erase tool preserves only those areas outside the extent of the analogous erase layer (part (f) of Figure 7.11 "Vector Overlay Methods "). While the input layer can be a point, line, or polygon dataset, the erase layer must be a polygon dataset. Continuing with our clip example, county managers could then use the erase tool to erase the areas of private ownership within the county floodplain area. Officials could then focus specifically on public reaches of the countywide floodplain for their upkeep and maintenance responsibilities.

The **split**[25] geoprocessing operation is used to divide an input layer into two or more layers based on a split layer (part (g) of Figure 7.11 "Vector Overlay Methods "). The split layer must be a polygon, while the input layers can be point, line, or polygon. For example, a homeowner's association may choose to split up a countywide soil series map by parcel boundaries so each homeowner has a specific soil map for their own parcel.

Spatial Join

A spatial join is a hybrid between an attribute operation and a vector overlay operation. Like the "join" attribute operation described in Section 5.2.2 "Joins and Relates", a spatial join results in the combination of two feature dataset tables by a common attribute field. Unlike the attribute operation, a spatial join determines which fields from a source layer's attribute table are appended to the destination layer's attribute table based on the relative locations of selected features. This relationship is explicitly based on the property of proximity or containment between the source and destination layers, rather than the primary or secondary keys. The proximity option is used when the source layer is a point or line feature dataset, while the containment option is used when the source layer is a polygon feature dataset.

24. A geoprocessing operation that preserves only those areas outside the extent of an erase layer.

25. A geoprocessing operation that divides an input layer into two or more layers based on a split layer.

When employing the proximity (or "nearest") option, a record for each feature in the source layer's attribute table is appended to the closest given feature in the destination layer's attribute table. The proximity option will typically add a numerical field to the destination layer attribute table, called "Distance," within which the measured distance between the source and destination feature is placed. For example, suppose a city agency had a point dataset showing all known polluters in town and a line dataset of all the river segments within the municipal boundary.

This agency could then perform a proximity-based spatial join to determine the nearest river segment that would most likely be affected by each polluter.

When using the containment (or "inside") option, a record for each feature in the polygon source layer's attribute table is appended to the record in the destination layer's attribute table that it contains. If a destination layer feature (point, line, or polygon) is not completely contained within a source polygon, no value will be appended. For example, suppose a pool cleaning business wanted to hone its marketing services by providing flyers only to homes that owned a pool. They could obtain a point dataset containing the location of every pool in the county and a polygon parcel map for that same area. That business could then conduct a spatial join to append the parcel information to the pool locales. This would provide them with information on the each land parcel that contained a pool and they could subsequently send their mailers only to those homes.
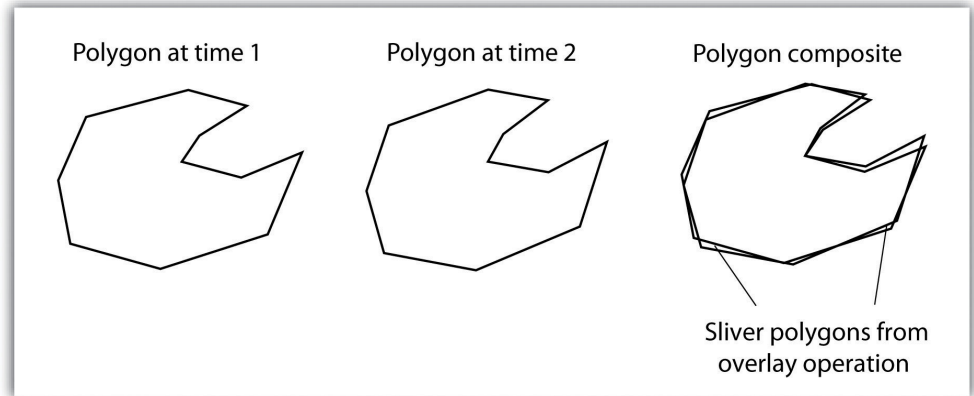
Overlay Errors

Although overlays are one of the most important tools in a GIS analyst's toolbox, there are some problems that can arise when using this methodology. In particular, **slivers**[26] are a common error produced when two slightly misaligned vector layers are overlain (Figure 7.12 "Slivers"). This misalignment can come from several sources including digitization errors, interpretation errors, or source map errors (Chang 2008).Chang, K. 2008. *Introduction to Geographic Information Systems*. New York: McGraw-Hill. For example, most vegetation and soil maps are created from field survey data, satellite images, and aerial photography. While you can imagine that the boundaries of soils and vegetation frequently coincide, the fact that they were most likely created by different researchers at different times suggests that their boundaries will not perfectly overlap. To ameliorate this problem, GIS software incorporates a **cluster tolerance**[27] option that forces nearby lines to be snapped together if they fall within a user-specified distance. Care must be taken when assigning cluster tolerance. Too strict a setting will not snap shared boundaries, while too lenient a setting will snap unintended, neighboring boundaries together (Wang and Donaghy 1995).Wang, F., and P. Donaghy. 1995. "A Study of the Impact of Automated Editing on Polygon Overlay Analysis Accuracy." *Computers and Geosciences* 21: 1177–85.

26. A narrow gap formed when the shared boundary of two polygons do not meet exactly.

27. A geoprocessing setting that forces nearby vertices to be snapped together if they fall within a user-specified distance.

Figure 7.12 *Slivers*

A second potential source of error associated with the overlay process is error propagation. **Error propagation**[28] arises when inaccuracies are present in the original input and overlay layers and are propagated through to the output layer (MacDougall 1975).MacDougall, E. 1975. "The Accuracy of Map Overlays." *Landscape Planning* 2: 23–30. These errors can be related to positional inaccuracies of the points, lines, or polygons. Alternatively, they can arise from attribute errors in the original data table(s). Regardless of the source, error propagation represents a common problem in overlay analysis, the impact of which depends largely on the accuracy and precision requirements of the project at hand.

KEY TAKEAWAYS

- Overlay processes place two or more thematic maps on top of one another to form a new map.
- Overlay operations available for use with vector data include the point-in-polygon, polygon-on-point, line-on-line, line-in-polygon, polygon-on-line, and polygon-in-polygon models.
- Union, intersection, symmetrical difference, and identity are common operations used to combine information from various overlain datasets.

28. When inaccuracies are present in the original input and overlay layers and are carried through to an output layer.

EXERCISES

1. From your own field of study, describe three theoretical data layers that could be overlain to create a new, output map that answers a complex spatial question such as, "Where is the best place to put a mall?"
2. Go online and find the vector datasets related to the question you just proposed.