Mapping Distance and Density

Distance functions allow you to determine the nearest location of something or the least-cost path to a particular destination. Density functions, on the other hand, allow you to see the highest and the lowest concentrations of features in your data.

Distance functions may be directly or indirectly used in an analysis. The Straight Line Distance function, for example, computes the straight-line distance to a particular location, and is often the whole of an analysis (e.g., locating the nearest hospital). The value of each cell in a surface generated with the Straight Line Distance function equals the measured distance of a straight line from the cell’s closest source.

The Cost Weighted Distance function, however, calculates the least-cost path to a particular location, a process that usually represents one part of a larger analysis (e.g., finding the least-cost path for constructing a pipeline). The cell values in a surface created with the Cost Weighted Distance function represent the accumulated cost of travel from the closest source. Distance is measured from each cell to the closest source, such as from a campsite to a store location.

In density surfaces, however, the cell values always indicate the number of features per unit of area, such as the number of people per square mile. Density surfaces are great for visualizing patterns in the data that otherwise might not be apparent.

Distance Surfaces

You may have heard the expression "as the crow flies" used to describe distance or direction, for example, "The campsite is located 8 miles due south of town, as the crow flies." Since a crow can easily fly above obstacles such as forests, rivers, and gated communities, this expression implies a straight line path, or the shortest distance between two points.

For those of us who can’t fly, forests, rivers, and gated communities are obstacles that will impede travel to the campsite, so we must decide the best way to navigate around or through them. You may even choose a campsite that is farther away because getting to it requires less effort. So, as the crow flies, the nearest campsite might in fact be 8 miles due south, but for you it could be a 20-mile hike.

With that in mind, consider two types of raster surfaces: one that measures distance as the crow flies, and the other that measures distance based on the cost of travel. These types of surfaces will help you find the nearest location of something or the least-cost way of getting there. All distance surfaces are calculated based on a source, which is the location of some object of interest, such as your home. If there is more than one source, as in the case of city fire stations, the calculation is based on the nearest one.
The Straight Line Distance Function

A straight line distance surface is used to find the shortest distance, a straight line, from each cell to the closest source. Similar to using a ruler to measure the distance between two points on a piece of paper, the Straight Line Distance function measures the distance between cells from the center of each cell in the surface to the center of a source cell.

A straight line distance surface makes it easy to find the nearest location of something, such as the nearest hospital. While there can be more than one source, each cell value is the straight line distance to only one source, the closest one.

Straight Line Allocation and Direction Functions

The Straight Line Allocation function creates a surface where each cell is assigned to the nearest source based on the straight line distance between them. If there is only one source, then all of the cells in the surface are allocated to that one source.
If there is more than one source, the surface is partitioned into areas of adjacent cells. You can think of the partitioned areas as compartments dedicated to one feature, such as the location of a hospital or a store. The size and shape of the compartments is determined by the proximity of the cells to the sources.

The cell values in a straight line direction surface point the way back to the nearest source, indicating the direction, in degrees azimuth, back to the source. Imagine a compass, like the one below, hovering over each cell in the surface.

If there is more than one source, the cell value is the direction to the nearest source.
Cost Surfaces

A cost surface represents some factor or combination of factors that affect travel across an area. For example, steep terrain can increase road construction costs, so the slope of the terrain is a cost factor. Percentage slope values do not, in themselves, indicate whether costs are high or low. To reflect the cost, or to create the cost surface, you must transform the slope values to cost values, such as dollars, or rank the slope values using a common scale.

<table>
<thead>
<tr>
<th>Percentage slope</th>
<th>Ranked slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.14 10.57 16.03 35.36 45.63</td>
<td>1 2 2 9 9</td>
</tr>
<tr>
<td>8.33 9.79 18.11 40.29 45.01</td>
<td>1 1 6 9 9</td>
</tr>
<tr>
<td>8.68 13.23 25.29 37.45 38.56</td>
<td>1 4 9 9 9</td>
</tr>
<tr>
<td>9.05 12.37 15.67 23.34 19.04</td>
<td>1 5 4 9 6</td>
</tr>
<tr>
<td>8.67 10.18 12.96 15.49 17.21</td>
<td>1 2 3 4 5</td>
</tr>
</tbody>
</table>

A particular area can have many cost surfaces: one for each factor. For example, if an area has slope and snow depth as factors affecting travel across it, then the area has two cost surfaces. In order to combine two cost surfaces, the values of both cost surfaces need to be ranked based on a common scale.

Determining which factors affect costs and ranking cost surface values can be a very time-consuming process. You may need to consult experts about what the costs might be, or form a consensus among your peers. Keep in mind that a single, final cost surface is required input for the Cost Weighted Distance function.
The Cost Weighted Distance Function

Distance and direction surfaces created with the Cost Weighted Distance function are used in a least-cost path analysis. With the exception of the cost weighted allocation surfaces, cost weighted surfaces are normally not used directly in analysis; whereas straight line distance surfaces often are.

Cost weighted surfaces are generally not used directly in analysis because cell values in a cost weighted distance surface are not distance values, but accumulated costs.

To calculate the cell values in a cost weighted surface, the Cost Weighted Distance function evaluates the neighbors of each cell, beginning with the source, multiplies the average cost between each pair by the distance between them, and assigns each of the neighboring cells a cost weighted value. The process moves to the cell with the lowest value, evaluates its neighbors with unknown values, and so on.

Instead of a compass direction, each cell value in a cost weighted direction surface is given a code number that indicates the direction of the next least-cost cell. The decoder diagram below shows the direction each number code points.

In an allocation surface created with the Cost Weighted Distance function, the shape and size of its compartments are determined by the cost of travel from the cells within it to the nearest source. This type of analysis would be useful if you had to consider the size of service areas based on the cost of travel over land.
Least Cost Path Analysis

If the shortest path between any two points is a straight line, then the least-cost path is the path of least resistance.

Least-cost path analyses use the cost weighted distance and direction surfaces for an area to determine a cost-effective route between a source and a destination. For example, you can use least-cost path analysis to find the cheapest route for building a pipeline or the quickest way to a set of observation points.

In a least-cost path analysis, the eight neighbors of a cell are evaluated and the path moves to the cell with the smallest accumulated value. The process repeats itself until the source and destination are connected. The completed path represents the smallest sum of cell values between the two points.

Any combination of sources and destinations can be part of a least-cost path analysis. For example, you can find the least-cost path from one source to many destinations, or from many sources to a single destination.
Density Surfaces

Density surfaces are good for illustrating concentrations of point or line locations. For example, if on an annual basis lightening strikes a hilltop more often than the valley below, then the density of lightning strikes will be concentrated near the hilltop.

Density is a calculation of the quantity of something per unit of area, such as the number of annual lightning strikes per acre. You can calculate the density of countable things: discrete objects or events (e.g., businesses, trees, car thefts, earthquake epicenters) and their attributes (e.g., employees of businesses, types of trees, property value of car thefts, magnitude of earthquakes).

The Density function has two methods for calculating density: simple and kernel. Both methods employ a circular neighborhood or search area to make the density calculation.

Simple Density Calculations

The simple method for creating a density surface uses a circular search area, or neighborhood, to calculate cell values. In a density surface, individual cell values are calculated by dividing the number of features that fall within the search area (e.g., observations) by the size of the area (e.g., 2.88 acres). The resulting value is then assigned to the cell. Every cell in the surface is processed in the same way.

\[
\frac{7 \text{ observations}}{2.88 \text{ acres}} = 2.4 \text{ observations per acre}
\]

The size of the search radius affects the result of a density calculation. A larger search radius produces a smoother surface because a larger search radius is more likely to encompass data points.
Kernel Density Calculations

A more mathematically sophisticated way to calculate surface density is to use the kernel method. When applying the kernel method, Spatial Analyst draws a circular neighborhood around each sample point (not each cell) and then applies a math function that goes from 1 at the location of the point to 0 at the neighborhood boundary. Think of a kernel as a smoothly curved surface that is fitted over each point.

When a kernel function is applied to each data point, the effect is like that of an elevation surface, except that the density value for each cell is calculated by adding the values of all the kernel surfaces where they overlay the cell center.

The kernel function generally creates a smoother-looking surface than one created with the simple method.
Using Attributes to calculate density

Sometimes points represent populations or the number of occurrences at a particular location.
For example, the number of people in a village or the number of thefts at convenience stores are attributes that are assigned to points.
Instead of calculating the density of the points, you can also calculate the density of the populations or events that those points represent.

\[
\frac{686 \text{ people}}{3.14 \text{ square kilometers}} = 218.47 \text{ people per square kilometer}
\]

The cell values for a kernel density surface are determined by finding the sum of attribute values in overlapping kernels. In this case, a math function is applied that goes from the attribute value at the location of the point to 0 at the neighborhood boundary.

In both situations, the population, not the number of occurrences, is distributed throughout the surface.
Using Cell, Neighborhood, and Zonal Statistics

You need statistics to describe your data, to add validity to your research, and to make sound decisions. Traditionally, statistics are used on a random but representative subset and the results are extrapolated to the larger group. In other words, you can ask a question of a subset of the population and make inferences about the entire population from the subset's answers. This subset of the population is called a sample. Inferential statistics, however, don't always work as well with geographic data. When this is the case, descriptive statistics are applied.

The methods of inferential statistics don't transfer easily to geographic data for two main reasons.
- First, inferential statistics assume that you want to estimate the characteristics of a population from a sample. With geographic data, however, you often have the entire population to work with, so you use descriptive statistics rather than inferential statistics.
- Second, inferential statistics does not include tools for representing geographic data.

Spatial Analyst provides a set of statistical functions, which makes descriptive statistics part of your geographic analysis. For example, you can compare the difference between values over time, cell-by-cell, or you can construct a statistical filter to weed out unwanted values. You can also assess past trends or the current status of features, or reveal the underlying structure of the data. Frequently, the results of statistical analyses, end up as inputs for further analyses. They become pieces of a more elaborate puzzle and necessary steps in the modeling process. Statistical methods are also used to clean up data, smooth out the rough edges, and make the data more digestible.

Comparing Raster Datasets using Cell Statistics

Statistics are useful for describing certain tendencies in your data. You may want to know the average value, the highest value, or how many different types of values exist in the dataset.

For a single raster dataset, statistics are automatically generated. The minimum, maximum, and mean values, as well as the standard deviation of values are presented in the layer’s properties.

You can also use statistics to create new raster datasets. While the statistical functions are divided into three basic groups (cell, neighborhood, and zonal statistics), each group utilizes the same statistical methods.

Cell statistics allow you to compare two or more raster datasets on a cell-by-cell basis. In other words, cells occupying the same location but belonging to different rasters can be evaluated together using basic descriptive statistics. This is especially useful when comparing time-series data, such as annual changes in land use.
Statistical Methods

You can use the following statistical methods for cell, neighborhood, and zonal functions:

- **Majority**: determines the value that occurs most often.
- **Minority**: determines the value that occurs least often.
- **Maximum**: determines the highest value.
- **Minimum**: determines the lowest value.
- **Range**: computes the difference between the highest and lowest values.
- **Mean**: computes the total of the values divided by the number of observations.
- **Median**: computes the central value, where there are an equal number of values above and below it.
- **Standard Deviation**: computes the spread or dispersion of values from the mean.
- **Sum**: computes a total value by adding all of the values.
- **Variety**: determines the number of different values.

To be evaluated, the cell values must be numbers, either a measurement or a code representing a feature or the ranking of a feature. If the values are measurements, like temperature, all of the statistical methods described above are valid. For example, you can find the mean temperature from a range of different temperatures. However, if the values are codes, such as unique numbers representing features, some of the statistical methods may not be useful. If you had, for example, a raster dataset with values ranging from 1 to 20 and each value was code indicating a specific vegetation type, then finding the sum of those values would not make sense. On the other hand, you could use variety to find out how many types of vegetation occur.

The Cell Statistics Function

You can use the Cell Statistics function to find trends or detect change between corresponding cells in multiple raster surfaces. The assessment works exactly like the mathematical operators in the Raster Calculator, such as multiplying or adding raster datasets together. In this case, however, the function uses one of the statistical methods mentioned previously to compare the corresponding cells. For example, you can create a raster with cell values indicating change from one land use type to another.
Monitoring change in the landscape is a way of discovering trends in urban development, habitat loss, and crop management, for example. To monitor the change in the landscape, you need snapshots of the data for an area that show the area's conditions at different points in time. Once you have the snapshots you can compare them using the Cell Statistics function to see where changes have occurred.

Describing raster datasets using neighborhood and zonal statistics

While you can use statistics to compare corresponding cells from different raster datasets, you can also evaluate a single raster dataset based on neighborhoods or zones.

The Neighborhoods Statistics function considers the values of cells within a specified neighborhood around the processing cell. Neighborhoods are sections of the raster that can be defined in almost any way you want. Neighborhood statistics are output as new raster layers.

The Zonal Statistics function considers the values of cells based on groups of like cells, or zones, in another dataset. Zonal statistics are output as tables.

The Neighborhood Statistics Function

The Neighborhood Statistics function helps you discover trends within a single raster dataset based on the values within a specified group of cells, which surround and usually include the evaluation cell. This technique is like passing a filter over your data and can be used to indicate diversity of feature types or to impose neighboring values over existing values.
Statistics are determined based on the cells in the neighborhood. The corresponding cell in the output raster is given the resulting statistical value. The neighborhood passes over every cell of the input raster, left to right, top to bottom, until all of the cells have been processed. You determine the size and shape of the neighborhood. For example, you can use a rectangular neighborhood of any dimension.

Alternatively, you can use other neighborhood shapes, such as a circle, an annulus (a doughnut shape), or a wedge.
A circular neighborhood is specified by giving the radius of the circle in the number of cells measured perpendicular from the x- or y-axis. The annulus or donut neighborhood requires two radii. The first specifies the number of cells to the inside of the donut and the second, the number of cells to the outer edge. Only cells that fall between the two radii will be considered. The wedge or partial circle is specified by identifying the radius of the circle and two-degree bearings that define the slice of the circle for the neighborhood. The smaller degree reading must be specified first, and only the cells between the two degree readings will be considered when processing each cell. Zero degrees is to the east and 90 degrees is to the north; the bearing increases counterclockwise.

If you don’t specify a neighborhood shape or size, Spatial Analyst uses the default, a 3-cell by 3-cell rectangular neighborhood.

The Zonal Statistics Function

In a raster dataset, any two or more cells with the same value belong to the same zone. A zone can consist of cells that are connected or disconnected, or both. For example, all of the cells in a raster dataset with a value of 1 belong to a zone, all of the cells with a value of 2 belong to another zone, and so on. In fact, every cell in a raster belongs to a zone and a raster dataset can contain one or many zones.

You use the Zonal Statistics function to find trends in data within zones defined by another raster or vector dataset. In this way, the areas of analysis are fixed or constrained by the shape and location of the zones. Instead of a new raster, the Zonal Statistics function produces a table of statistics and a graph.
Zones are composed of regions. A region is a group of connected cells in a zone. A zone that consists of a single group of connected cells has only one region.