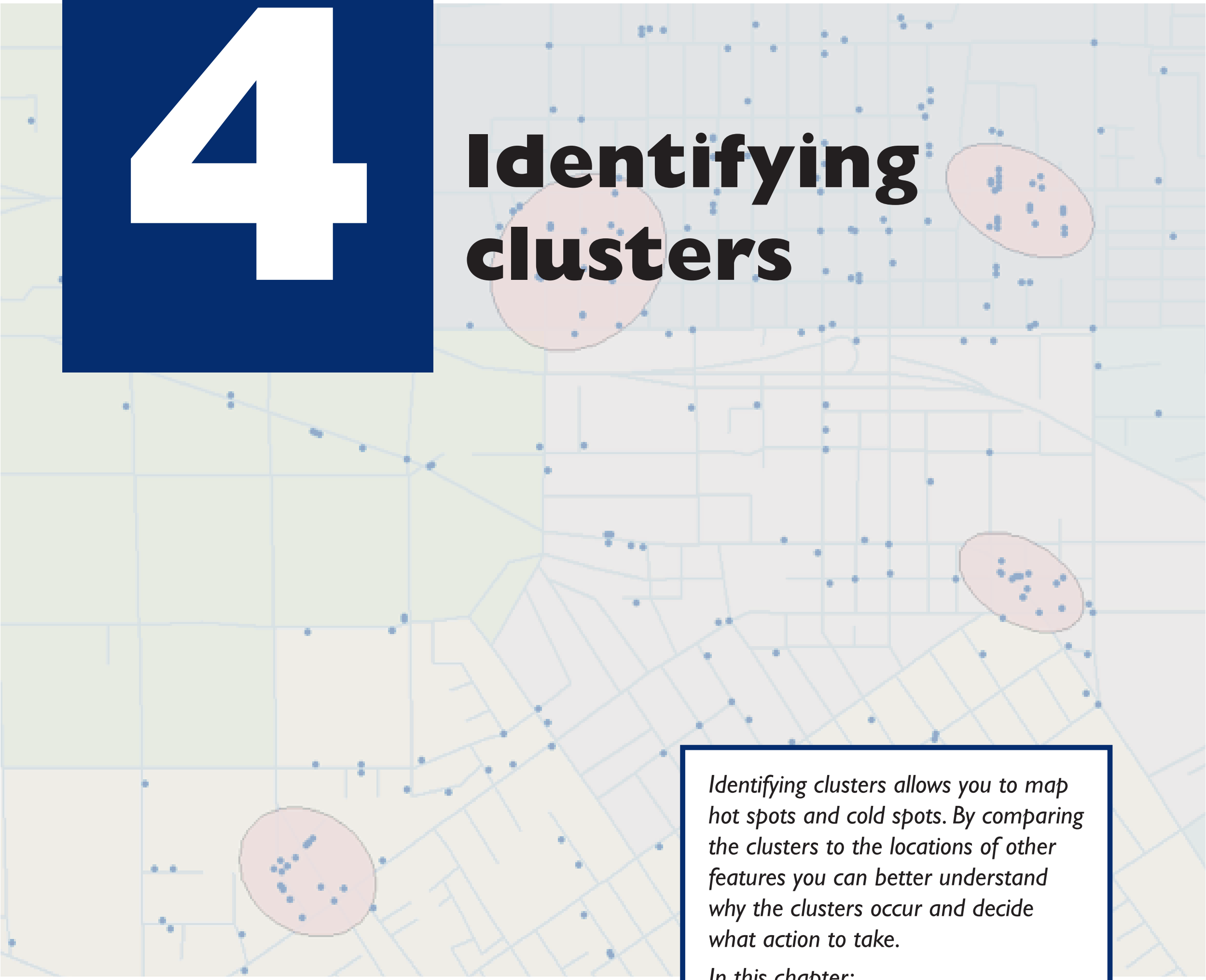
4 Identifying clusters



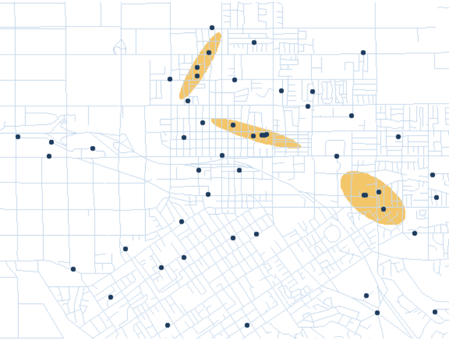
Identifying clusters allows you to map hot spots and cold spots. By comparing the clusters to the locations of other features you can better understand why the clusters occur and decide what action to take.

In this chapter:

- *Why identify spatial clusters?*
- *Using statistics to identify clusters*
- *Finding clusters of features*
- *Finding clusters of similar values*

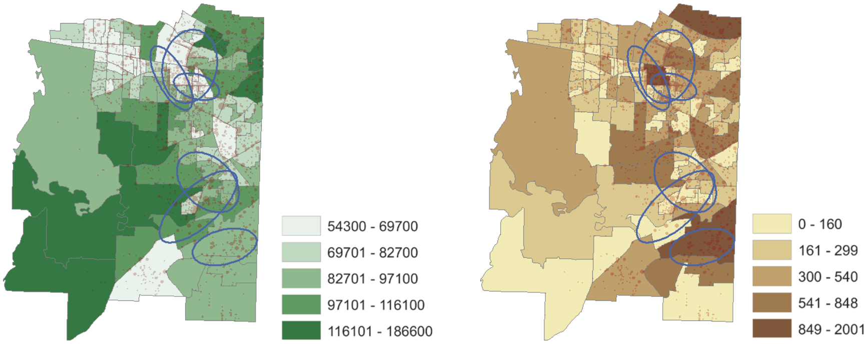
Why identify spatial clusters?

Clusters occur in a geographic distribution when features are found in close proximity or when groups of features with similarly high or low values are found together (hot spots and cold spots). Identifying whether—and where—clusters exist is useful if you need to take action based on the location of one or more clusters—such as assigning a task force to deal with a cluster of burglaries.



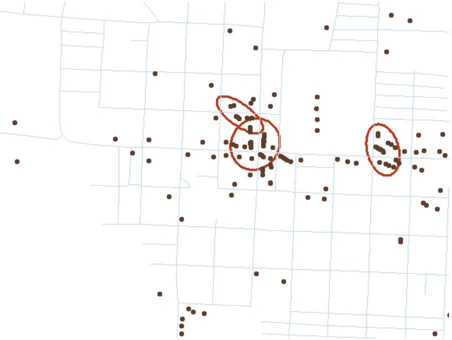
Clusters of burglaries identified using statistical analysis

Pinpointing the locations of clusters can help you examine the causes of the clustering. By comparing the locations of clusters to the other features, you can start to identify possible contributing factors. For example, by comparing the clusters of cases of a particular disease to environmental and economic data, you could see if there are possible spatial relationships between the clusters and any of these factors.



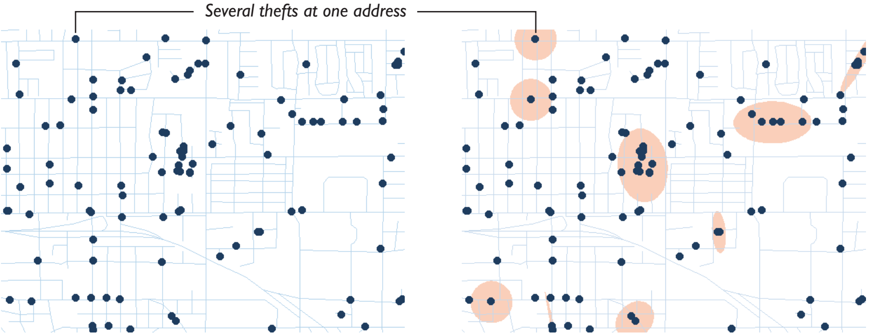
There may be a relationship between significant clusters of emergency calls (ellipses) and median home value (shown by block group, left), but not necessarily with numbers of young adults (right).

By looking at a map, you can draw conclusions about where there are clusters of features. Statistics lets you test those conclusions and validate them by measuring whether features are closer than would occur by chance. You can then map the results of the test, not just the features themselves.



Clusters of businesses

Using statistics takes much of the guesswork out of identifying clusters. If there are multiple events at a single location, such as several auto thefts at the same address, the clusters can be hard to see if you simply map the features. When you use statistics to identify clusters, each event is counted as a unique occurrence.

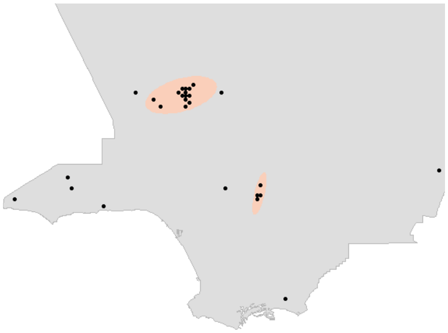


Locations of auto thefts (left) and clusters of auto thefts (right)

When using statistics to identify clusters, you can calculate the probability that the clusters are not due to chance, so you can be more confident in any decisions you make based on the results of the analysis.

Clusters can be formed using the locations of features alone, or formed using the location influenced by an attribute value (clusters of features with similar values).

Clusters based on location alone are composed of discrete features. Often the features are points, such as auto accidents, earthquake epicenters, or cases of measles.

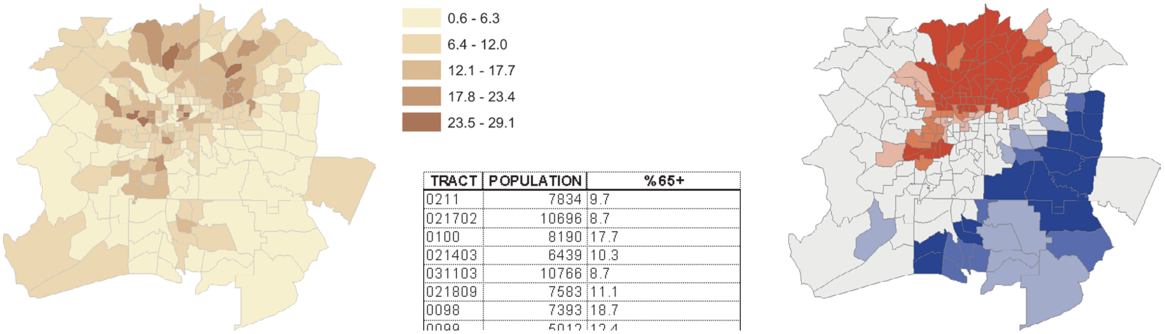


Earthquakes in Los Angeles County in March 1974, forming two clusters

The features can also be discrete areas, such as cut areas in a forest. However, since the centroids of the areas are used to identify the clusters, the results may be misleading, especially if the areas are elongated or convoluted. (See "Using statistics with geographic data.")

Clusters of features having similar attribute values can be composed of discrete features (points or areas), spatially continuous data, or data summarized by contiguous areas (such as census tracts or counties). The attributes are interval or ratio values.

A market analyst might locate clusters of coffee shops with low sales, to find out where additional marketing is needed. A social service agency planning senior services could identify clusters of census tracts where a high percentage of the population is seniors.

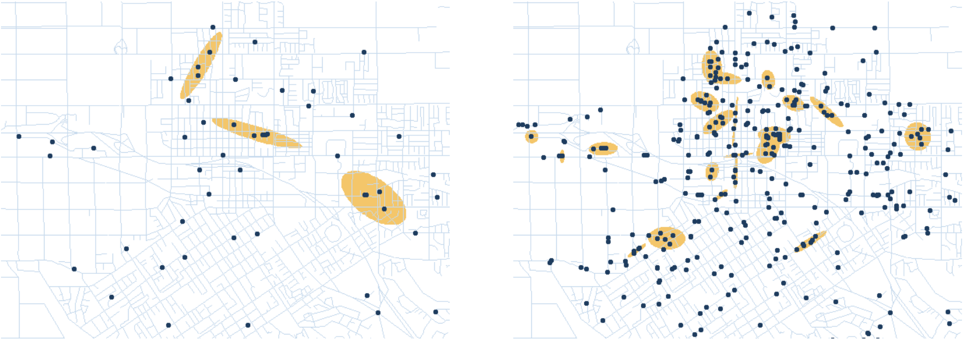


TRACT	POPULATION	%65+
0211	7834	9.7
021702	10696	8.7
0100	8190	17.7
021403	6439	10.3
031103	10766	8.7
021809	7583	11.1
0098	7393	18.7
0099	5012	12.4

Percent age 65 and over by census tract (left) and statistically significant clusters of tracts with a high percentage of seniors (orange) or a low percentage (blue).

What's the time period of the data?
By identifying a cluster, in many cases you're assuming the features are related in time as well as in space. For static features, such as vacant parcels, you use a snapshot of the current condition. If you're analyzing events that individually take up little time—such as crimes or earthquakes—you need to define the period to use. For example, a crime analyst looking for clusters of gang-related assaults might include incidents for the past five or six months; assaults from longer ago are unlikely to be related to any current incidents. On the other hand, a geologist looking for clusters of earthquakes over time, to possibly predict future seismic activity, would likely include events for the entire available period of record—at least a hundred years.

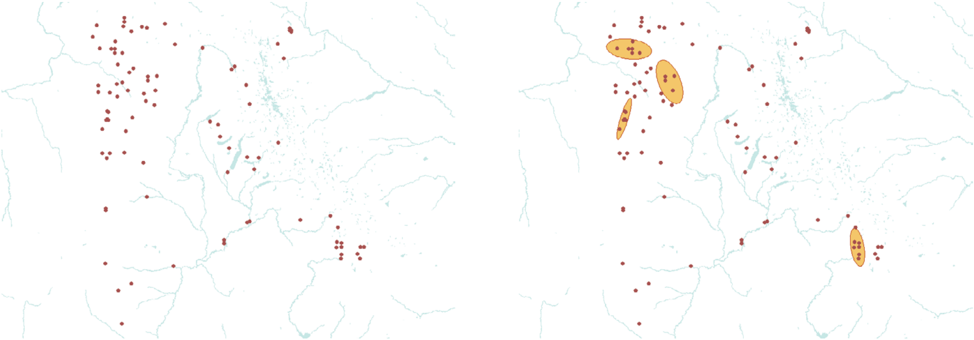
Even for the same type of feature, the time period will change depending on the purpose of your analysis. A crime analyst trying to find burglaries possibly committed by the same group of people would include only recent burglaries. If the analyst were trying to identify clusters representing ongoing high-crime areas, she'd include burglaries over several years in the analysis.



Cluster of burglaries occurring over one month (left) and over one year (right).

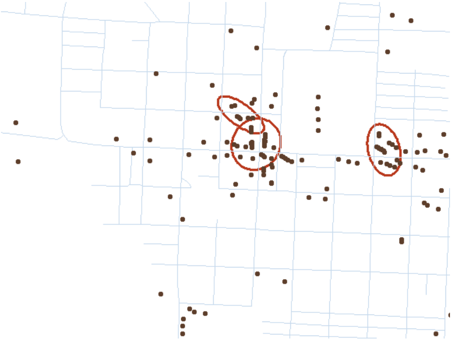
What measure of distance are you using?
Clusters are usually defined using straight-line (Euclidean) distance between features. However, you can also use other measures of distance, such as travel time or cost. For example, a crime analyst would want to identify clusters of burglaries using driving time between the crimes rather than straight-line distance, especially if the incidents are separated by a barrier. Two burglaries on opposite sides of the river may be near each other using straight-line distance but in fact may not be very close in terms of travel time.

One method for finding clusters of discrete features is to specify the distance features can be from each other in order to be part of a cluster, and the minimum number of features that make up a cluster. The method, described by CrimeStat author Ned Levine, is known as nearest neighbor hierarchical clustering.



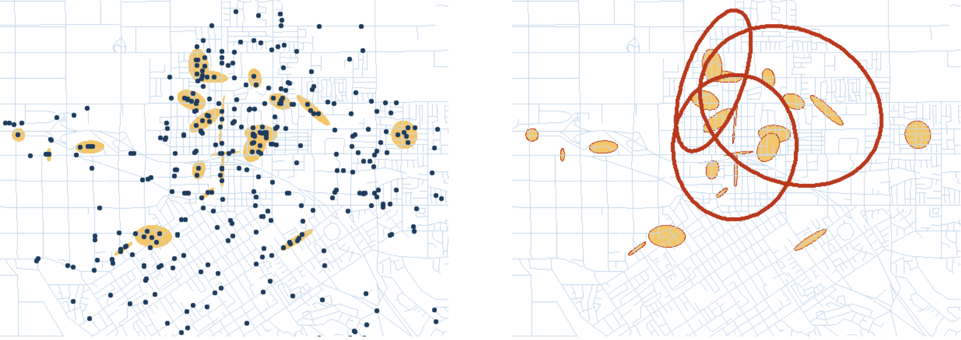
Moose sightings during March, over several years (left) and clusters of sightings (right)

Nearest neighbor hierarchical clustering uses the distance between features (similar to nearest neighbor analysis for analyzing the pattern of a distribution of features).



Clusters of businesses. Features must be within a specified distance of each other to be considered part of a cluster.

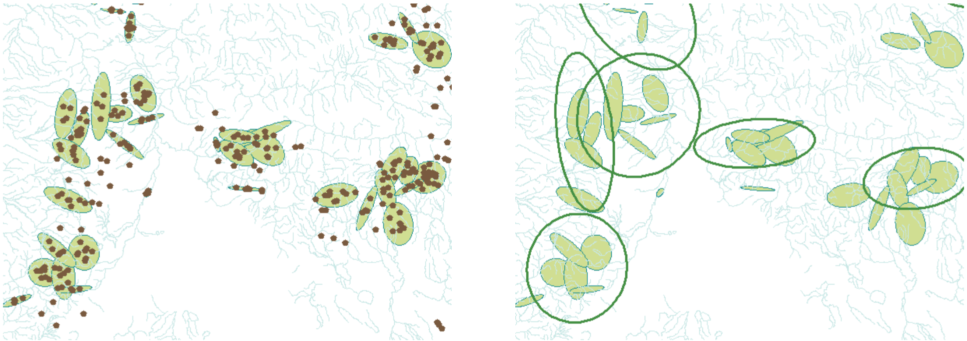
It is hierarchical because, after delineating a first set of clusters, the routine continues on to group the clusters into larger clusters.



Clusters of burglaries (left) and the clusters grouped into three larger clusters (right)

Hierarchical clustering shows you how the features are clustered at several geographic scales, such as how burglaries cluster at both the neighborhood and citywide scale. In crime analysis, for example, as Ned Levine describes in his CrimeStat documentation, you could identify first-order clusters at the neighborhood level, where officers would intervene for specific incidents, and higher-order clusters, which might correspond to areas requiring community policing.

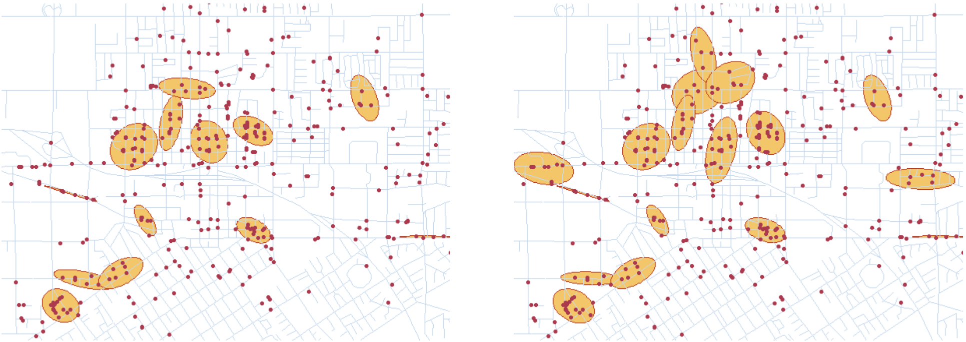
In wildlife management, clusters of individuals of a species might identify habitat areas that need to be protected, while clusters of these habitat areas might define larger management areas requiring development of corridors linking the clusters, and other strategies.



Clusters of wolverine sightings (left) and the clusters grouped into larger clusters, potentially defining habitat areas (right)

HOW NEAREST NEIGHBOR HIERARCHICAL CLUSTERING WORKS

For nearest neighbor hierarchical clustering, you specify a probability level, which the GIS uses to calculate the distance within which features will be considered a cluster. With a lower probability, more features will be included in a cluster, but you'll be less certain that the features actually represent a cluster.



Clusters of assaults calculated using a high probability (left) and a low probability (right)

There is a range within which the distance between two features may occur owing simply to chance. This range—termed the confidence interval—is calculated based on the probability level you specify.

If the distance is greater than the high end of the range, the features are farther apart than you would expect by chance. Since you're trying to find features that are closer than you would expect by chance (clusters), you're interested in the lower end of the range. This value is the minimum, or "threshold," distance.

The confidence interval is calculated using the mean distance that would occur between points in a random distribution—that is, the mean distance between points if there were the same number of points as in the dataset you're analyzing spread over the same study area, and you knew the points were randomly distributed. This is termed the "mean random distance."