Chapter 1

# Why spatial is special

## Introduction

Spatial statistics are special. They are not just traditional statistics that we happen to apply to spatial data. They explicitly use some aspect of geography, some notion of space, in the calculation of the statistics. Sometimes this is just incorporating x,y coordinates. Sometimes it is the shape or the orientation of our spatial data. Sometimes it is the distance between locations. But one of the most powerful uses of geography is defining spatial relationships, or what is nearby. These relationships are then used in spatial statistics algorithms to do things like compare local neighborhood averages to global averages to find hot spots, for example. Incorporating nearby features and neighborhoods often provides valuable information that would otherwise be lost in traditional statistical analysis.

This is not to say that traditional statistics are not useful in answering spatial questions. Traditional, nonspatial statistics are appropriate for spatial data if we understand their limitations and ensure that we are not violating underlying assumptions of those methods (more on that later). Some of the methods we'll cover in this book are not inherently spatial. Still, we'll make the case that interpreting these analyses in a spatial context can be very powerful.

Thinking about things like adding spatial variables (for example, variables that represent the distance to roads or water bodies), visualizing and analyzing results spatially, and using results to make spatial decisions demonstrates how these traditional statistical methods can be part of a larger spatial workflow.

## The first law of geography

So, why does incorporating space matter so much? That leads us to what we often refer to as the first law of geography, also known as Tobler's law. Tobler's law states that **near things are more related than distant things**. In other words, things that are closer together are more related than things that are farther apart. This seems intuitive when we say it, but the reality is that many traditional statistical and machine learning–based approaches ignore this reality. In fact, many traditional statistical approaches have underlying assumptions that data is independent. But if Tobler's law holds true, which we know it often does, spatial data (a.k.a. most data) is rarely independent. And those spatial relationships, that dependence between data, is actually a unique and incredibly valuable characteristic of the data.

Ignoring it not only violates statistical assumptions but leaves a lot of potential information on the table.

Spatial statistics build on the concept of spatial relationships in a way that helps us gain a deeper, more valuable understanding of our data. It can help us find patterns that would go unnoticed; it can help us make predictions more accurately, and it can help us make decisions confidently.

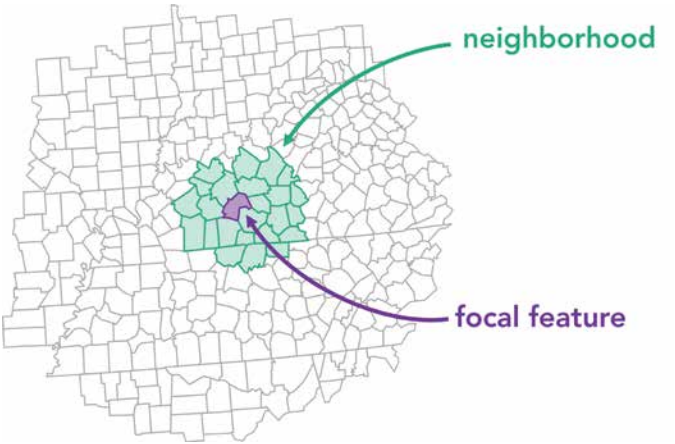
## Defining spatial relationships

Now we know that spatial statistics use geography. But how?

Any tool or method that uses spatial relationships will require us to define or conceptualize what it means to be neighbors. In other words, we need to define what it means to be related in space (and sometimes even in time). This concept of spatial relationships can feel a bit abstract, but actually it's quite simple. Features in our data can be related in any number of ways. Perhaps two parcels are touching, which makes them neighbors. Perhaps all parcels

within a particular drive time of each other are related. Ultimately, every feature will have its own well-defined **neighborhood**.

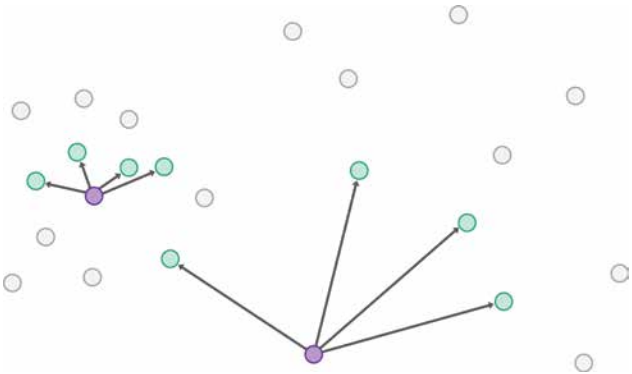
To explore this concept, we'll call the feature whose neighborhood we're defining the **focal feature**.




Functionally, spatial relationships are actually represented as a series of weights. Large weights mean lots of influence between features (a strong spatial relationship), whereas small weights represent less influence. For most definitions of spatial relationships, weights can either be binary or continuous. With binary weights, a feature is either included in the neighborhood or it is not. With continuous weights, the magnitude of the weight determines how important the relationship is or how influential the neighboring feature is to the focal feature. There are countless ways to define spatial relationships. Let's explore some of the most common approaches.

## Number of neighbors

The number of neighbors method defines a neighborhood using a user-specified number of features (in this case, four) that are closest to the focal feature.

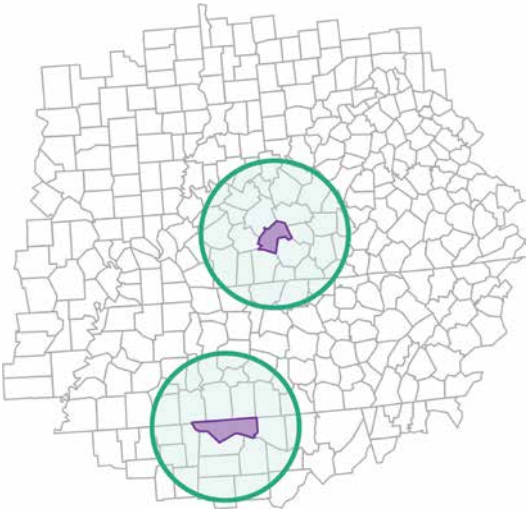


This method is a very powerful, very popular way of defining spatial relationships. Notice how the distances can vary depending on the density of features in the area. Features in dense areas have close neighbors (and therefore smaller neighborhoods), and features in sparse areas have neighbors that are farther away (and therefore much larger neighborhoods). Of course, the number of neighbors is constant throughout each neighborhood, but the size of those neighborhoods can vary quite dramatically. For this reason, this type of neighborhood is often referred to as an adaptive neighborhood.

This approach is also known as k-nearest neighbors, which may seem a little "mathematical" in its name, but *k* simply represents the specified number of neighbors (in our example, k = 4).

## Fixed distance

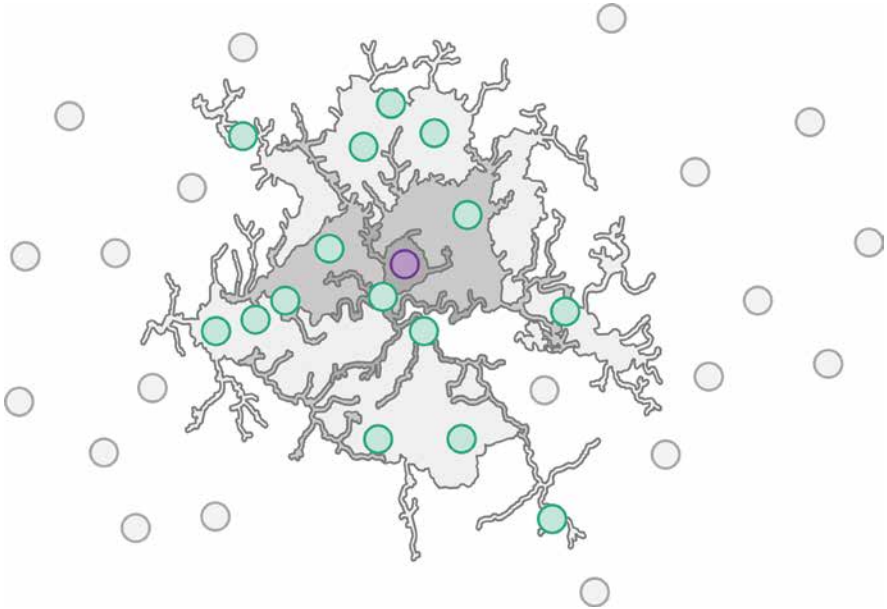
Fixed distance neighborhoods are calculated using a specified Euclidean distance (also known as a straight-line distance or an as-the-crow-flies distance). All features that fall within the specified distance of the focal feature are considered neighbors. Notice how the number of neighbors changes depending on the density of features in the area.



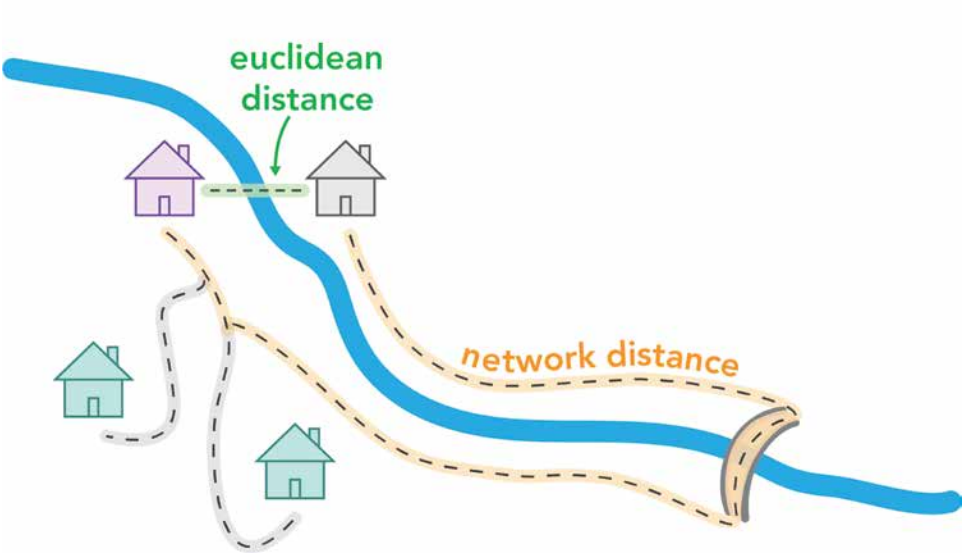
Features in dense areas have many neighbors, and features in sparse areas have far fewer neighbors, using the exact same distance band. For this reason, this type of neighborhood is often referred to as a fixed distance neighborhood, because while the number of neighbors changes, the size of the neighborhood is fixed.

## Network distance

Network distance neighborhoods are like fixed distance neighborhoods, but instead of using a Euclidean distance, they use either a network distance or network travel time.

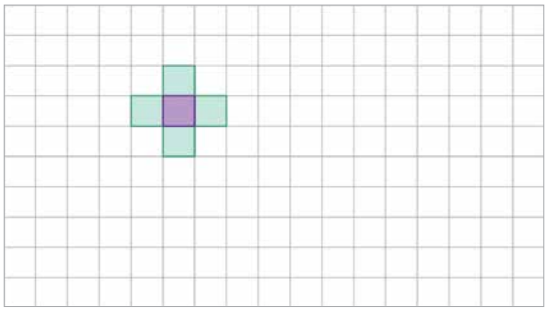


Using a network distance allows our modeling of spatial relationships to be more realistic when looking at human mobility. For instance, two features sitting across a river from each other may have a very short Euclidean distance, but if the only way to get across that river is a bridge 15 miles down the road, then in network distances those features are actually very far apart.

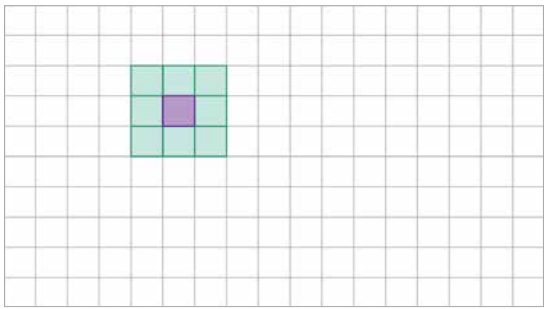


## Contiguity

Contiguity neighborhoods define neighbors based on shared boundaries. There are two flavors of contiguity: edges and edges corners. Contiguity edges defines a neighborhood using all features that share an edge with the focal feature.



Contiguity edges corners defines a neighborhood using all features that share an edge or corner.



The contiguity relationships are applicable only to polygon features.