Lesson 5: Spatial Data Quality

## INTRODUCTION

This lesson defines data quality and spatial data quality. You will learn about measuring spatial data accuracy and quality as well as the five facets used to measure accuracy and quality. You will view images and supporting data to be able to understand how to determine what is considered relevant and necessary while creating measuring accuracy and quality. Positional accuracy is also exhibited and explained in this lesson. The process for completing the National Standard for Spatial Data Accuracy (NSSDA) is outlined and explained as are locational errors. The last portion of the lesson focuses on Data Aggregation and why this is necessary.

## LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Define key terms relating to the accuracy and quality of spatial data, locational errors, and data aggregation.

2. Identify and review data in accordance with the National Standards for Spatial Data Accuracy.

3. Identify and edit locational errors from primary/secondary data sources.

## LEARNING SEQUENCE

|  |  |
| --- | --- |
|  | |
| Required Reading | Read the following:  Spatial Data Quality   * Data Quality * Measuring Spatial Data Accuracy and Quality * National Standard for Spatial Data Accuracy (NSSDA) * Locational Errors * Data Aggregation |
| Assignments | Complete the following assignments:   * Lab 5: Spatial Data Quality * Quiz 5: Spatial Data Quality |

## INSTRUCTION

## Data Quality

### What is Data Quality

Before learning about quality this lesson will take a few moments to define and discuss the concept of quality. The ISO standard 9000 defines quality as: the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs.

What the standard is saying is that quality is measured against some need defined explicitly or implicitly. The standard is also alluding to the fact that you cannot define quality without first knowing what the end result will be as the requirements of the end result will define what sort of quality is needed. While quality data and items are very important in our daily lives, we should also not forget that the quality of spatial data is especially important for a number of reasons.

### Spatial Data Quality

Spatial data quality is important. History shows where knowledge of spatial data played key roles in wars, vacations, exploration, or even real estate. Societies really do rely on accurate and precise spatial data, and that reliance is increasing every day.

Technology is the enabler for a society’s reliance on accurate and precise spatial data as spatial data is now being infused into every facet of our lives and onto every device that we interact with. If poor quality spatial data is utilized in the one of these systems there may be grave consequences. Wanting high quality spatial data is not necessarily a new movement. Quality spatial data was always of interest, but high quality spatial data was not only of great interest it was also of great value.

Historically cartographers have been able to keep up with demand for high-quality maps and spatial information. However, as newer technologies emerge, and the number of map users and producers increase, we are experiencing some new challenges in making sure that we produce high-quality data.

### Disparate Data

What is also interesting is that cartographers are now challenged with having too much disparate data! Traditionally, cartographers were well acquainted with the data on the map as they typically spent quite a bit of time with the data producer pouring over the data and translating it to a visual medium.

Today data collection is so widespread and dispersed that it is likely that the cartographer has never met the data producer. This leaves the interesting problem of having a cartographer use and trust the data that they are receiving from a data producer without spending significant amounts of time with the data producer to get a deep understanding of how they created the data. How does a producer describe the data to a user, and how do we, as cartographers and GIS users, determine which data sets used between competing data sets?

A cartographer needs to know if there are there standards to follow. The good news is that there are some standards by which data producers can document their data and there are standards that producers can adopt when producing their data. Whether you should use one competing data set over another is still the unanswered question as it will depend on your evaluation the documentation and standards used by the data producers, and the quality that you need.

## Measuring Spatial Data Accuracy and Quality

### Why Measure Spatial Data Accuracy and Quality?

When measuring spatial data accuracy and quality, there are five facets of quality and accuracy that we can reference. These five facets are: lineage, positional accuracy, attribute accuracy, completeness, and logical consistency. This section will define and discuss each of these five facets in more detail. Additionally, for each of the five facets of accuracy and quality, an aerial photograph and a vector data set will be reviewed to see real world examples of how accuracy and quality are reported.

### Accuracy and Quality

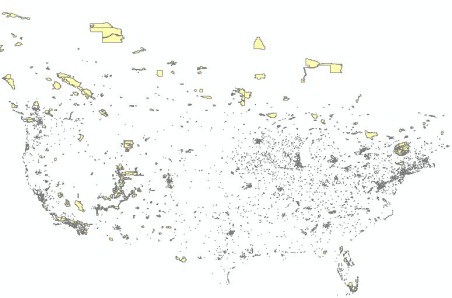
This is the aerial photograph that we will be considering when we are talking about accuracy and quality. There may be a case where you are presented with this aerial photograph to be used in your GIS, but there will not be any documentation, or a useful filename. Without any of that information how would you know where this photograph was taken, when this photograph was taken, whether it has been orthorectified, or if this was originally a true color or infrared photograph?

The answer is that you would not know and you could not find out who produced this data as they have the documentation. Therefore, if you receive this photograph without the documentation, you should not use it in a GIS until you have all of the accuracy and quality questions answered.



### Vector Data Set Image

This is the vector data set that we will be considering we were talking at accuracy and quality. Like the aerial photograph example, you may be presented with this data set with no documentation. Without the documentation, how are you even to know what this data represents, when it was created, or what the accuracy of it is?



### Lineage

The first facet of spatial data accuracy and quality is lineage. Lineage can be thought of as the history of the data set. Lineage should describe the source material from which the data set was derived and if there were any transformations or modifications made to the data set along the way. You should use lineage to determine whether or not the data set is fit for your usage.

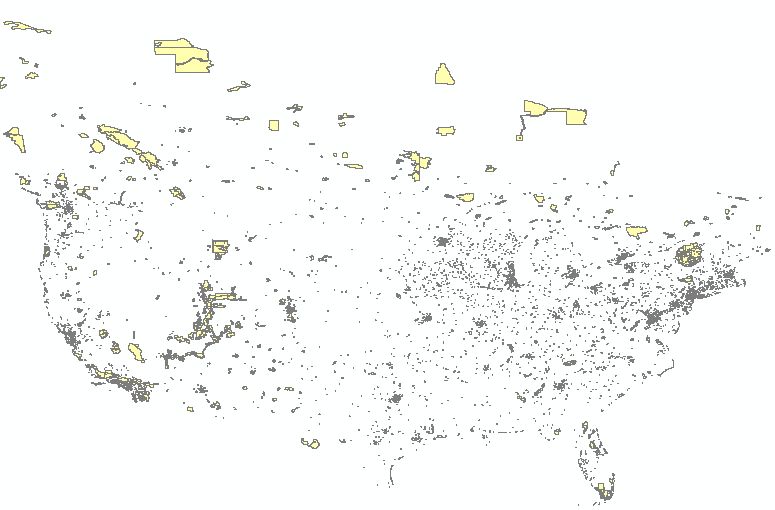
Consider the aerial photograph and the lineage included with the image. If you open the documentation for this photograph, you would see the lineage information included. The lineage information provided tells us who created the data, when the data was published, and how they created the data down to which camera they used and how they transformed it once they got out of the camera. Good lineage information like this allows you to start to determine whether you can use this data set for your purposes or not.



* Originator: USDA-FSA Aerial Photography Field Office
* Publication Date: 20081015
* Process Description: Imagery was flown with Leica ADS40 digital sensors to capture 0.75m raw data. Raw data is then downloaded using Leica GPro software into 12 bit TIFF format. The raw TIFF imagery is then georeferenced and reprojected using GPS/INS 200Hz exterior orientation information (x/y/z/o/p/k). The final DOQQ sheets contain a 300m minimum buffer. These final DOQQ tiles are edge inspected to the existing MDOQQ sheets for accuracy validation.

### Vector Data Set

## Now consider the vector data set. Remember, that this point, we do not even know what this data set represents. The lineage should provide us with enough information for us to know what this data set represents and where it came from. In this lineage documentation, we can see who originated the data, when it was published; a brief abstract explaining what is contained within the data set, the purpose the data set meets, and the process description explaining how the data set was extracted from its parent data set.



* Originator: Tele Atlas North America, Inc.
* Publication Date: 20080401
* Abstract: U.S. and Canada Parks represents parks and forests within the United States and Canada at national, state, and local levels.
* Purpose: U.S. and Canada Parks provides thousands of parks and forests at national, state, and especially local levels.
* Process Description: The following steps were performed by ESRI: Extracted the data set from Tele Atlas StreetMap Premium for ArcGIS v. 9.0. Created ArcGIS® layer file (.lyr), projection file (.prj), and spatial indices.

## 

### Positional Accuracy

The second facet of spatial data accuracy and quality is positional accuracy. An object is considered to have high positional accuracy if the values of its coordinates are close to the “true” position. Positional accuracy can be specified and measured for both horizontal and vertical. Additionally, positional accuracy is often specified within degrees of compliance with spatial registration standards often set by the client, governing, or professional body.

As we consider positional accuracy one of the first questions should be: “what quality of positional information do GIS users need?” Generally, we require information to be near to the “truth” although “near” is fuzzy and may allow some leeway in position accuracy.

When defining the quality that you require you should specify a maximum acceptable error. Then you will use that definition as a general guideline for accessing and assessing data sources for your project. Even if you do have a maximum acceptable error, and the positional accuracy complies with the spatial registration standard, should always keep in mind that error is inherent in all measurements and that you should understand and measure error appropriately when using any spatial information.

### Reporting Accuracy

When producing and utilizing spatial data, is important that you represent and report accuracy of the data set correctly. Besides complying with spatial registration standards, one of the simplest ways to make sure that your data’s accuracy is not overrepresented is to pay attention to significant digits.

The amount of detail and reported measurement should reflect its accuracy. For instance, a measurement of 14.4 m implies an accuracy of .1 m. A measurement of 14 m implies an accuracy of 1 m. By including a decimal place and the tens digit, you are implying accuracy of a 10th of a meter. If you apply accuracy to a 10th of a meter, then your data should be accurate to a 10th of a meter. If you have more significant digits that are appropriate for the accuracy of your data, the excess precision should be removed by rounding the number.

## Let’s once again visit our aerial photo and see what the positional accuracy is. In this photograph the positional accuracy complies with the farm service agency digital ortho photo specs. To determine what the specs are, you would need to contact the Farm Service Agency to get a copy. For the vertical positional accuracy, the state that this is attempting accuracy measurement is not applicable as this is only two-dimensional data.

## Air Photo

### Positional Accuracy Example

The vector data set represents parks in the United States and Canada. Review the image and then see how the positional accuracy is reported.



**Horizontal Positional Accuracy:**

* The horizontal positional accuracy is 167 feet for the least generalized level. The least generalized level has the highest horizontal positional accuracy.

The positional accuracy statement for this data set stated that the least generalized level is accurate to 167 feet and that 167 feet is the highest stated accuracy of all positions stored in the data set. At first glance, this may seem to indicate that the data set is accurate to 167 feet. However, what this statement actually means is that the most accurate the data could be is plus or minus 167 feet and the accuracy of some positions in the date set could be worse.

**Vertical Positional Accuracy:**

* *<NULL>*

The vertical positional accuracy is Null, which means they did not enter anything in this field. Assuming this is a two-dimensional only data set, there would be no vertical positional accuracy statement here.

### Attribute Accuracy

The third facet of spatial data quality and accuracy is attribute accuracy. Remember that an attribute is a fact about some location or feature allowing us to distinguish one feature from another in the world and in the data set. In the attribute it can be uncertain. For example, a nominal attribute can have historical uncertainty where one group may claim one set of facts, and the competing group may claim a different.

Ordinal attributes may be uncertain in the case where there is a classification blunder or difference of opinion and people who are classifying the data. For interval and ratio attributes, there can be mathematical errors, errors in measurement, varying resolutions of measurement, or geographic grouping that may insert some uncertainty into the data.

### Measures of Error

When measuring and reporting error for attribute accuracy, these are the recommended methods.

* For interval or ratio attributes the error should be reported as a mean or standard deviation.
* For ordinal attributes, compare the attributes to some source of higher accuracy or get another expert opinion.
* For nominal attributes, compare the information some source of higher accuracy.

Returning to the aerial photograph we are able to see the accuracy of the attributes. The attribute information recorded in the documentation for this aerial photograph discusses the number of pixels per band and the brightness values. As this is a photograph the only attribute is color; therefore there is not much to report.



|  |
| --- |
| Entity\_and\_Attribute\_Information:  Overview\_Description:  Entity\_and\_Attribute\_Overview: 8-bit pixels per band, 4 band RGB and NIR represent brightness values 0 – 255 |

Now let’s consider the park data set. The attribute information for the park data set, notice that it does no. That should not be the case as the answer you accuracy information should be vetted by the data producer and documented here. So if this is not the desired behavior, then why is this example being shown to you? The reason it is being shown is to illustrate how you will often receive the data set with incomplete documentation. In this case, your next step would be to contact the data producer to determine how they verified the accuracy of their attributes.



|  |
| --- |
| Entity\_and\_Attribute\_Information:  Overview\_Description:  Entity\_and\_Attribute\_Overview: *<NULL>* |

### Completeness

The fourth facet of spatial data accuracy and quality is completeness. Completeness refers to how exhaustive a set of features are within a data set and which features were selected to be included within the data set. Additionally, if the definitions are used to describe features of the data set and if the definitions vary from any standard definitions these definitions should be noted.

### Sources of Incompleteness

There are many ways in which a data set can be complete. For instance, there may be a discrepancy between a data description and the data collected. For instance, if a data set reports to have all of the fire hydrants in Houston, Texas, however the data set only contains fire hydrants in Houston, Texas that have the color red then the data set does not match the description, and therefore it is incomplete.

Another source of the incompleteness is inappropriate data modeling. For instance, if we are collecting fire hydrants for our data set, but we do not record whether the fire hydrant is functional or not, and do not include some other important attribute the data set will be considered incomplete because it does not have the normal appropriate attributes. A data set can be rendered incomplete if there is a loss of data during data exchange or through an inappropriate user action. One more, and perhaps the most common source of incompleteness, is simply to make a temporal change in the data set.

#### Example 1

**Aerial Photograph Incompleteness**

Once again we will revisit the aerial photograph and consider whether it is a complete data set or not.



Completeness:

* *<NULL>*
* Is Null acceptable for aerial photos?

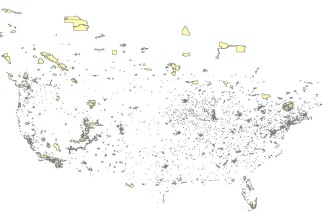
The documentation for the aerial photograph is null in the completeness section. Normally, a null field in the description of any of the data quality fields is not acceptable, however, thinking about what completeness means, is null an acceptable response for an aerial photograph?

We could argue that yes, null is an acceptable response in this case because a photograph records everything within its viewpoint leaving nothing out. So long as the date taken is included in the lineage, then a user would be able to determine that too much time has lapsed to be used for their purpose.

#### Example 2

**Park Data Set Incompleteness**

Returning to the park data set let’s consider whether this is a complete data set or not.



Completeness:

* After processing, the data set is checked for drawing display and number of records and file sizes are compared with source materials. This data set provides six levels of generalization.

The documentation for the park data set states that after they process the data they compare the number of records and file sizes with the source materials. It then goes on to list even though it is not shown here that there are six levels of generalization displayed in the data set.

This is a reasonable completeness statement showing that the authors have at least verified against the source material that they were able to extract all the targeted features and that the data was generalized to six different levels.

### Logical Consistency

The fifth facet of spatial data quality is logical consistency. Logical consistency deals with logical rules of structure and how we model the real world in our databases. Logical consistency also includes the idea of semantic and ontological modeling. Logical consistency can refer to the structure of the data set, what format it stored in, and how the features are represented as spatial phenomenon.

#### Aerial Photograph Logical Consistency

Let’s see what the metadata says in the logical consistency section for the aerial photograph.



Logical Consistency:

* *<NULL>*
* Is Null acceptable for aerial photos?

For the aerial photograph, the logical consistency does not have anything recorded, and is therefore a null value. Null can be considered an acceptable value for logical consistency since a raster simply stores a value per pixel and there is no underlying complex structure.

#### Parks Data Set Logical Consistency

Looking at the park data set we are able to see what the metadata has stored for logical consistency.



Logical Consistency:

* The shapefile is converted to SDC (Smart Data Compression) format using either ESRI SDC Data Development Kit Professional 2 (DDKP2) or tools in ArcGIS. The SDC data set is then loaded into ArcSDE® to verify and validate the geometry.

For logical consistency, it states: the shape file is converted to a smart data compression format using either as ESRI SDC development kit professional or tools at our GIS. The SDC data set is then loaded into ArcSDE® to verify and validate geometry. The statement discusses the formats and the way in which it was converted. This could also be considered part of lineage and may be better suited for that section.

### Concluding Thoughts on Spatial Data Quality

To wrap up this section, here are some concluding thoughts on spatial data quality. When considering spatial data and its quality, we need to keep in mind a few important concepts. First, spatial data is important and is valuable and widely used; therefore, high-quality spatial data should be consumed and produced.

It is important to know who created the data and how it was created so that way you can properly evaluate worth of the data to ensure it fits your quality needs. You should measure error and determine maximum acceptable error when determining which data sets to use. Remember that not all generous projects need pinpoint accuracy, however, that doesn’t mean that you should include data at any accuracy. Don’t forget to consider what is not in the data set.

Another aspect that needs to be considered is that know-how reality is modeled in the data set. For example, if you have downloaded a road data set, is important to know which data structure it is stored in, how the roads are represented, and whether topology is included. You should use these five metrics of data quality every time you consider using and creating a data set. If you’re an easy or unsure about the documentation, contact the organization or person who created the data, or don’t use the data set.

## National Standard for Spatial Data Accuracy (NSSDA)

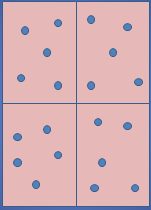
### Process for Completing the NSSDA

The United States national standard for spatial data accuracy defines test statistics, methods, and reporting of positional accuracy. There is a five step process to complete the national standard. The five-step process consists of selecting test points, defining independent control data sets, collecting measurements from both sources, calculating positional accuracy statistics, and reporting the accuracy of the metadata for the data set. Let’s go over each one of these five steps in more detail.

#### Step 1

**Select Test Points**

The first step outlined by the national standard for spatial data accuracy is to select test points in your data set. These test points must be well-defined and easily identifiable and should select anywhere between 20 and 30 different test points evenly distributed throughout the data layer. When selecting these test points you must make sure they are easy to find and measure in both the data set being tested and the independent data set which is being evaluated against.

This image includes four examples of test points that are reasonably well spread across the extent of the data set. The test points capture information from the extremities and in the central area. Notice the test points are not clustered closely together nor are they only representing a small portion of the data set.

#### Step 2

**Define Independent Control Data Set**

The second step outlined by the national standard for spatial data accuracy is to define an independent control data set. The independent control data set identifies the true values to be determined from the source data set. The independent control data set must be acquired separately from the data set being tested and should be three times more accurate than the expected accuracy of the test data set. The accuracy of the independent control set should always be reported in the metadata.

#### Step 3

**Collect Measurements**

The third step outlined by the national standard for spatial data accuracy is to collect test point coordinate values from both the test data set and the independent data set. You should record the numbers in an appropriate and similar numeric format paying mind to significant digits. You should also use similar commonsense recording computer accuracy statistics.

#### Step 4

**Calculate Positional Accuracy**

The fourth step outlined by the national standard for spatial data accuracy is to calculate positional accuracy using appropriate accuracy statistics worksheets. You should calculate the sum of the set of squared differences between the test data set coordinate values in the independent data set coordinate values. Calculate the average of the sum and then calculate the root mean square error which is the square root of the average. To get the final statistic for the standard, you should multiply your root mean square error by 1.73084 horizontal and 1.9600 for vertical. This represents the standard error of the mean at the 95% confidence level.

#### Step 5

**Report Accuracy in Metadata**

The fifth and last step outlined by the national standard for spatial data accuracy is to report the accuracy of the metadata. You should report the accuracy in a complete description in the data set explaining how you did the accuracy assessment and the outcome of the results. You should include this in the spatial accuracy section of the metadata.

## Locational Errors

### What is locational error?

Locational error means that a feature in the data set is not representative of its true or real position on the map. The feature can be displaced relative to the actual ground it represents, which is notice positional absolute accuracy, or it can be displaced relative to some other object, which is notice positional relative accuracy.

Locational errors occur from many different sources. For instance, locational error can result from distortion and aerial photographs, or distortion and satellite images. It can also be the result of a data conversion gone wrong or in proper conversion of one projection to another or from mixing projections in a project.

### Combining Different Projections

One of the most common ways in which locational errors occur is by combining different projections in the same map project. For instance, some data sets may be in the State Plane Coordinates system and some may be in the latitude longitude coordinate systems which are both based on different projections and properties of the earth. Although, most software completes conversions on-the-fly, it is best to convert the data into a common projection which is more accurate and more appropriate for when you are doing analysis.

### Scaling Data Improperly

Another common way locational errors occur is when data is used to the scale in which it was not intended to be used. As a reminder, a data scale is defined as a scale in which the data was collected and where it is considered to be accurate. Every time data is manipulated or combined with other data new errors can arise if the two data sets have different data scales.

If you are going to aggregate data which have different data scales, it is valid to aggregate data from a detailed level to a more generalized level. However you must then state that the entire data set exists at the most generalized level. It is not valid to go from a generalized level to a very detailed level.

## Data Aggregation

What is Data Aggregation?

Data aggregation is the grouping of spatial data at a scale that is smaller than the level at which the data was collected. In other words that means combining data so that it displays well at a smaller scale. There are some problems that we need to keep in mind when we aggregate data. As we create a coarser spatial data set we lose spatial resolution and attribute detail as we combine multiple features into fewer features. For some tasks, this may be the desired outcome, however this is not always the case and it is up to you to determine where the loss of spatial is and if the detail is appropriate. While we would often like to work with very detailed data, many times the only data available is aggregated data. Therefore you should be aware of the fact that much data available is only at the aggregate level.

### Why Aggregate Data?

Why do we aggregate spatial information? There are many reasons, let’s cover the most common four: protect confidentiality, generate data, generalize or summarize data, and simplified maps.

#### Protect Confidentiality

You may wish to aggregate data when we want to protect confidential information. For instance, if we have answers to surveys that were solicited from individual entities, then it is important that we protect this potentially confidential, sensitive, or personal information.

We can protect the individuals from being uniquely identified by aggregating the surveys together into a summary statistic covering a larger geographic area where there are a significant number of individuals participating.

#### Generate Data

You may wish to aggregate data when you generate the data set. A common case when you would wish to aggregate new data is if the volume of data is extremely large or very complex and you wish to simplify the data for public consumption or some other use.

#### Generalize/Summarize Data

Another reason to aggregate data is that you are working with a solution, model, or algorithm that requires aggregated spatial data as inputs. Many models and algorithms are based off of aggregated data as they might work at a particular spatial scale.

#### Simplify Map

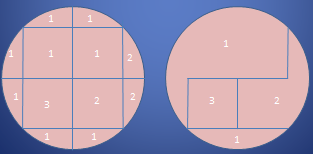
The fourth reason to aggregate data is if you wish to simplify the data for display on a map. In semantic mapping in particular, more data is not always better as it may clutter up the map and make it harder to understand. The cartographer must decide an appropriate scale of detail and this may require aggregation of data.

### Vector Techniques

Now that you know why you would want to perform data aggregation, let’s cover three aggregation techniques that work on vector data. The three techniques will discuss are dissolve, merge, and join.

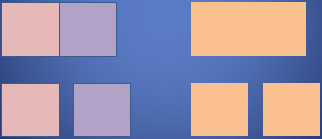
#### Dissolve

The dissolve function combines individual features from a vector data set based on where they share the same attribute value for a particular attribute field. If the features do share the same attribute value, the geometries combined into a single feature, and the attributes are combined into a single record in the attribute table.



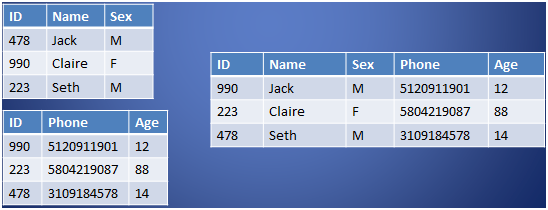
#### Merge

The merge function combines multiple input data sets into a single output file based on geometry. Selected input features can be merged together to create one single feature composed of the geometries of the input features. The input features do not need to touch to be merged. This means that the output merged feature can be a multipart feature.



#### Join

The joint function join attributes from one feature to another feature based on common attribute values. If two tables each have a column that represent the same attributes as the other tables, then the two tables can be joined together to create one larger table. This table would then contain only one column of the same attribute the join which was based off of the remainder columns from both input tables.



### Issues

Anytime we aggregate data there were a few issues that we need to be aware of. The three issues to take note of include: the modifiable areal unit problem (MAUP), the ecological fallacy problem, and the cross area aggregation problem.

#### Modifiable Areal Unit Problem

The Modifiable Areal Unit Problem (MAUP) is when point-based measures of spatial phenomenon are aggregated into districts, resulting summary values that are influenced by the choice of enumeration units. Additionally, if the geometry of the enumeration units change between different time epochs, then it becomes very difficult, or impossible, to perform an apples to apples comparison between two different times.

#### Ecological Fallacy

Ecological fallacy arises when a statistical relationship using aggregate spatial data is attributed to individuals. The user of the data cannot make any inference concerning the cause of the relationship without further analysis.

#### Cross-area Aggregation

Cross-area aggregation is the transfer of aggregate spatial data from one enumeration unit to another. The most common approaches use area waiting, which assumes the data are distributed uniformly within zones, which is rarely the case.

## SUMMARY

In this lesson data quality and spatial data quality were explained. You learned about measuring spatial data accuracy and quality as well as the five facets used to measure accuracy and quality. You viewed images and supporting data to help you understand how to determine what is considered relevant and necessary while measuring accuracy and quality. Positional accuracy is also exhibited and explained in this lesson. The process for completing the National Standard for Spatial Data Accuracy (NSSDA) is outlined and explained as are locational errors. The last portion of the lesson focuses on Data Aggregation and why this is necessary.

## ASSIGNMENTS

1. Lab 5: Spatial Data Quality
2. Quiz 5: Spatial data quality