Error, Spatial Bias, Logical Fallacies

Error, Spatial Bias, Logical Fallacies

{: .no_toc }

For our purposes, we will use the work **uncertainty** to describe all problems that arise from our imperfect understanding of the world, our inability to measure it exactly, and our imperfect nature when translating the real world to the GIS. You can think of it using the equation:

$$Uncertainty = Error(precision + accuracy) + ambiguity + vagueness + logicalflaws$$

Data Quality in GIS

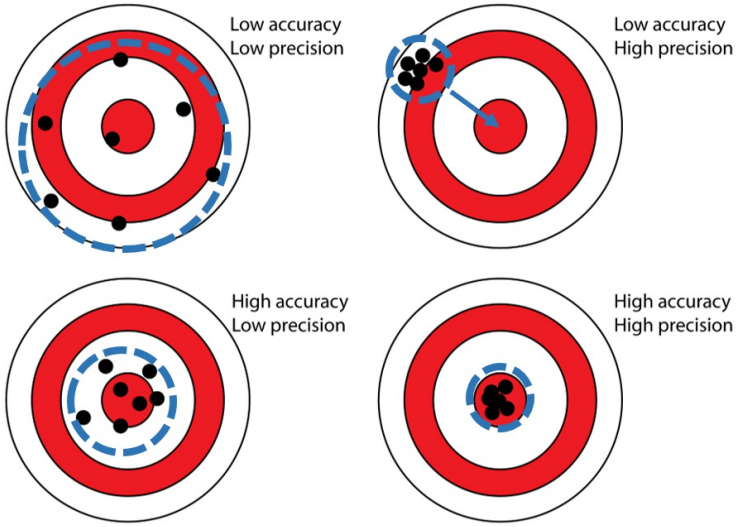
Data quality is mostly a measure of how good a particular dataset is to your needs. It is up to GIS users to assess data to determine if it is appropriate to their needs and use case. There is no standardized measure of data quality for GIS products.
* May pass through many hands before flaws are discovered * You have to trust that the data was collected correctly and that it was processed in a correct and appropriate manner * There is a risk of the user misinterpreting otherwise valid products due to lack of experience, knowledge, or skill.

Unlike say … baking, GIS products don't have obvious evidence of their own inadequacy.

Uncertainty

Errors: Accuracy and Precision

These terms are closely related, but the distinction is **very** important.



Accuracy: The degree to which a set of measurements correctly matches the real world values. How close are we to the real value? * If there is a consistent (systematic) offset from that real world value, our measurements are inaccurate. Inaccurate measurements have a **bias**.

Precision: The degree of agreement between multiple measurements of the same real world phenomena. How repeatable is a measurement? * If you take five measurements of the same feature, how likely are they to be similar? Lack of precision can be attributed to random (**unbiased**) errors.

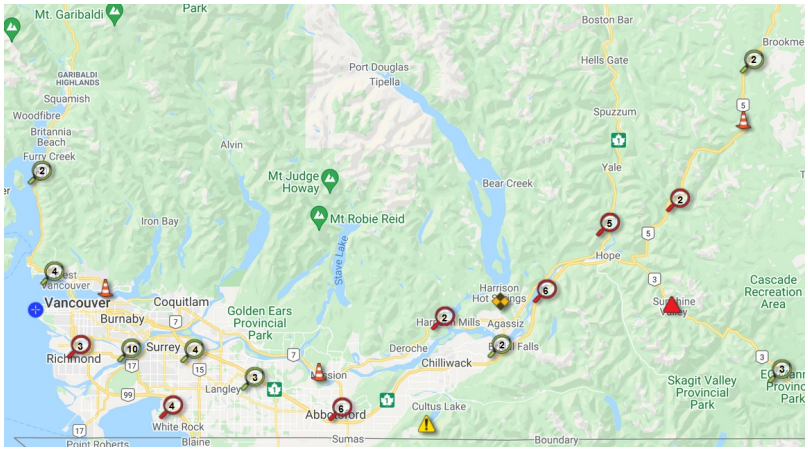
Vagueness and Ambiguity

These two terms are very closely related. The main distinction, its vagueness is about lack of information, ambiguity is about multiple options.

Vagueness: When something is not clearly stated or expressed. Vagueness arises when boundaries or definitions are poorly defined. eg. The word "bank" can refer to a financial institution or a riverside. * Boundaries may not be clearly defined * Where does a forest end? On a coastline we might have a pretty distinct boundary. But what about * Data may be an estimate * Stats Canada census long form - only samples 20% of the population. Questions pertaining to: racial identity, income, hosing cost, etc. are all drawn from estimates. It is highly unlikely the number presented by Stats Canada are **exactly** representative of the real value. * The position of objects are unclear or changeable * Fraser river boundary file - Is it the low water line? mean water level? high water line?

Ambiguity: Uncertain in meaning. Ambiguity arises when something can be interpreted in more than one way. eg. stating that an action "may be appropriate" reduces the clarity about whether or not the action should be performed. * Labels can often apply to multiple

features: does “London” refer to London, UK or London, Ontario, Canada? * Multiple points of interest may be located near a label: which one does it refer to?

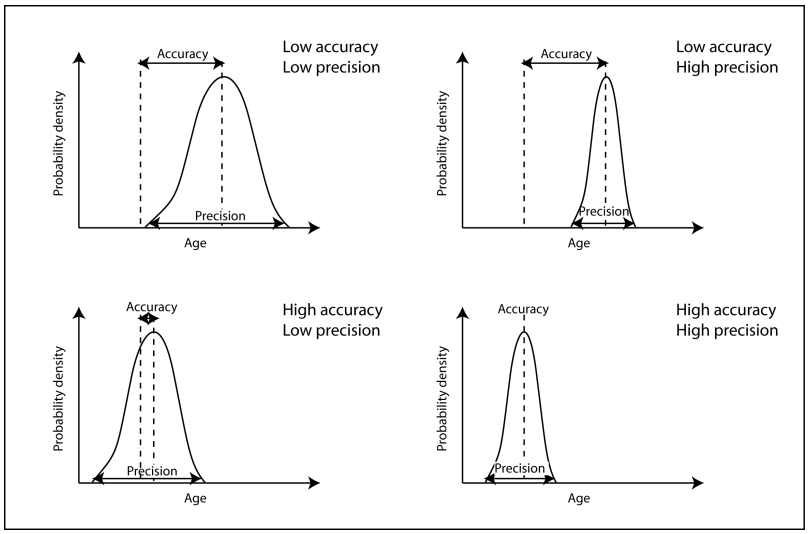


Misunderstanding

There are a near infinite number of ways we can misunderstand or misinterpret a system. The misunderstanding of a problem must be evaluated on a case by case bases.

Quantifying Errors

Even if there is no standard measure of uncertainty in GIS, it can be helpful to use some statistical methods to quantify error. These measures won’t tell us for sure that we are correct, but they can give us some insight.



Quantifying Accuracy

We can gauge the accuracy of an observation/estimate by comparing them to know true values. This is not always feasible, but when we have enough information, we can do it. If

$$x$$

is an observed/estimated value,

$$t$$

is a the true value, and we have N total:

Mean Absolute Error (MAE): The absolute value of the error for each estimate, averaged over all values. Gives us an idea of how close the set of observations/estimates tend to be to the truth.

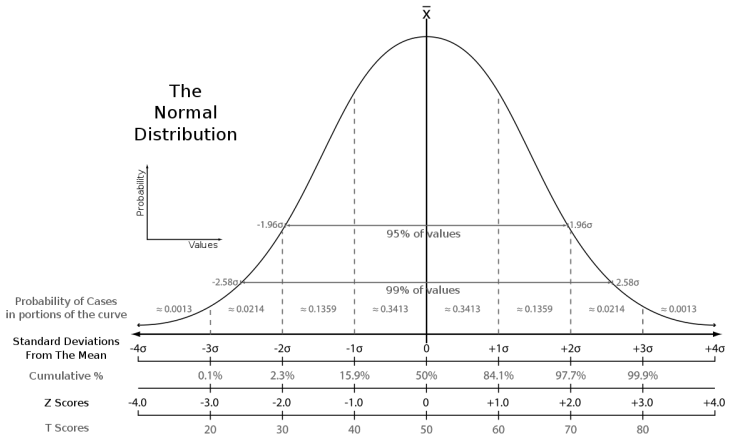
$$MAE = \frac{\sum_{i=1}^{N} |x_i - t_i|}{N}$$

Mean Squared Error: The squared value of the error for each estimate, averaged over all values. Similar to MAE, but more harshly penalized **large** deviations. But, squaring means the value is **not** in the same units as the original value.

$$MSE = \frac{\sum_{i=1}^{N} \left(x_i - t_i\right)^2}{N}$$

Root Mean Squared Error (RMSE): Same as MSE, but taking the square root, puts RMSE back in the units of the estimate.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - t_i\right)^2}{N}}$$



Quantifying Precision

We can quantify the precision of an estimate by looking at the spread of a dataset. If

$$x$$

is an observation/estimate, and

$$\bar{X}$$

is the s

Standard Deviation (

$$\sigma$$

): It is similar to RMSE, but instead of characterizing error, it the dispersion of a dataset.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2}{N}}$$

Confidence Intervals (CI): The standard deviation can be used to gauge our confidence in the average value of an estimate. If all the x values are close together, we have higher confidence in $\bar{X}$. If they are more dispersed, we have lower confidence in $\bar{X}$. It is the ratio of σ to the $\sqrt{N}$, multiplied by a [z score](#)

$$CI = \frac{\sigma}{\sqrt{N}} z$$

Qualifying Ambiguity and Vagueness

Ambiguity and Vagueness are more difficult to quantify numerically. The key to addressing these issues is to present things clearly and thoroughly. It is important to be as thorough, explicit, and transparent as possible when conducting and describing your work.

Sources of Error & Uncertainty

Data Measurement and Entry

In many cases, sources of error are out of our control. The tools we rely on to collect information are only so precise, there is a limit to how close we can measure things. **Sometimes** they are thing we can control, such as typos furring tabular data entry or digitizing errors when manually creating shapes. There are three types of error than can occur when features are digitized (especially when done by hand):

- **Gaps**: Places where there should be a feature, but there is not
- **Overlaps**: Where one polygon sits over another polygon
- **Slivers**: Where a new feature is created between two features when it should not be

Data Resolution

Data resolution impacts both accuracy & precision. * Lower resolution data is by definition less precise, but not necessarily less accurate. * High resolution data can be very precise, but still be biased. * What do you do if your data are collected at different resolutions? * Resolution applies to **Time** as well as **Space** * Maybe your data is out dated? * Maybe your analysis spans a long time period - what census years should you use? How do you work with data that spans multiple years? - I'll give an example of that next week.

Data Processing, Conversion, and Projection

Even with 'perfect' data, a series of complex GIS operations on good-quality, highly precise, and error-free data can still add uncertainty. * Each task performed in a GIS increases that level of uncertainty! * Tasks may perform generalizations, may re-project data into different coordinate systems, may perform mathematical transformations that introduce errors (raster to vector conversion and vice versa). * Tasks can be performed incorrectly or with the different settings than intended

Logical Fallacies: Flaws in our Thinking

A logical fallacy is a flaw in our reasoning that undermine the logic of our argument. They can be made both by accident and on purpose. They can often be identified because there is a lack of evidence to support the claims/decisions made, or evidence being presented in a misleading or untrue.

- "Hasty generalizations" are an example of logical fallacies: 'I saw a violent protester on TV … Protesters are inciting violence.'

Labels, Boundaries

Since geographic phenomena often don't have clear, natural units, we are often forced to assign zones and labels in our work (eg. census tracts). * This is a convenient way of simplifying complex processes. * However, these boundaries/labels are often **vague** and/or **ambiguous**. * They may be difficult to defend because they are arbitrary. * Where to draw a boundary, and what to call a zone are likely to vary significantly between different people/groups.

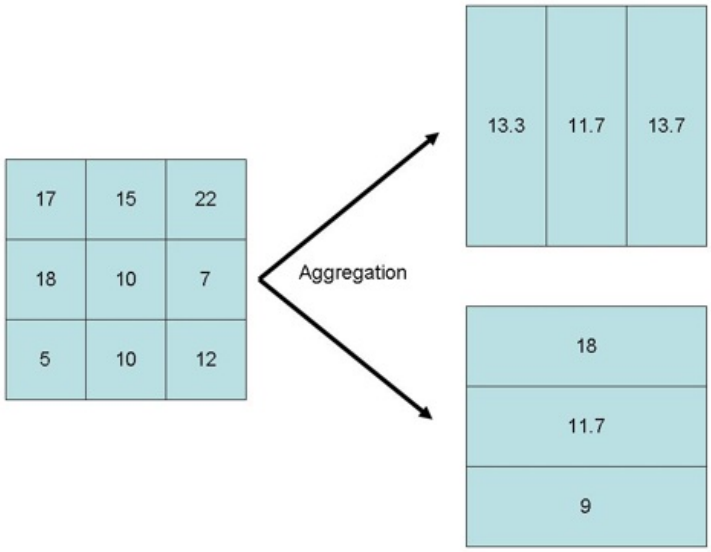
Data Aggregation

Much of the data we use to learn about society is collected in aggregate. We take average values for many individuals within a group or area (eg. **Census Data**) * This lets us explore the average case for each group/area * It allows us to explore the make up or attributes of different groups or regions * This can be useful for comparing different areas, to determine where resources should be and where they can be taken away

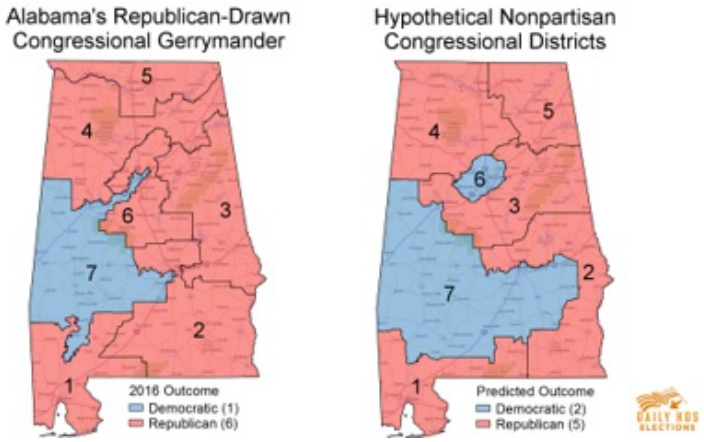
The ***ecological fallacy*** is an issue related to how we interpret statistical data when taking data collected from a group and applying that to individuals within that group. Census data is averaged for an area, the information about individual values is lost. You cannot learn about individuals within the group or area based on the aggregate data for that group/area. Basically - don't make assumptions about individuals.

- The median income for my census tract is $2491/month
    - I don't make that much …

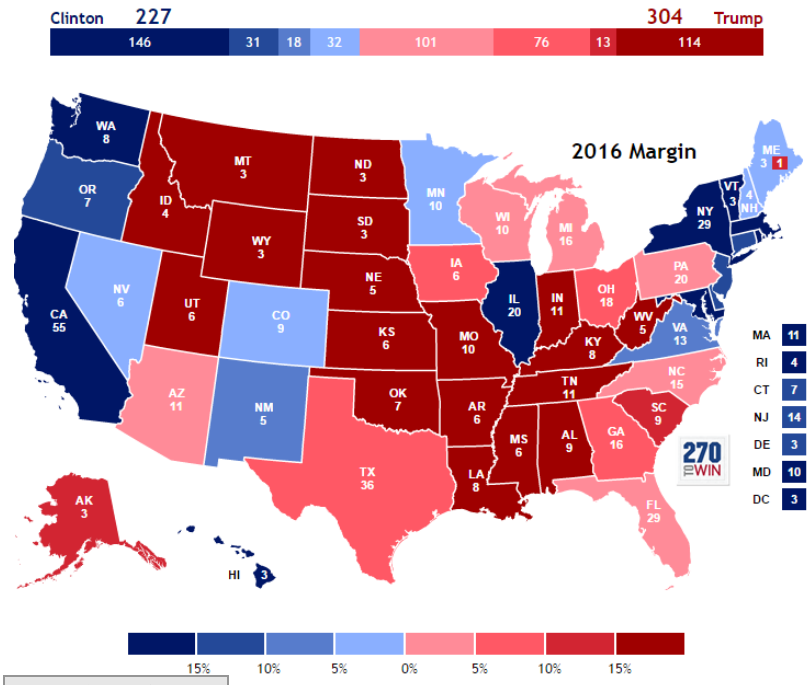
The **Modifiable Aerial Unit Problem** (MAUP) relates to how we choose to draw boundaries. Modifiable, arbitrary boundaries that have little justification can have a significant impact on values given to aggregated (combined) areas. Related to the Ecological Fallacy. When areas are grouped together, the way you choose to group them can change the values of the groups



The big issues is that data collected at a finer level of detail is being combined into larger areas of lower detail. Not only are we losing information, but what information remains can be easily manipulated. We can use this property to imply things about the data that aren't necessarily true



The **Atomistic** fallacy occurs when we assume we can combine already aggregated data and aggregate it again at an even higher level For example: If you take the average income for CTs and average over Metro Vancouver to calculate average income



Processing math: 100%