Week 12:
Introduction to Spatial Data & GIS

URBST 200 | Adjunct Lecturer: Erin Lilli | November 14, 2022

- GIS data is spatial (where things are located) and is visualized via mapping at various geographic scales. Geographical space is "define as having positional data relative to the Earth's surface" (Huisman & de Bay, 2009, p. 27)
  - Changes in the earth's geography is natural, human-made, or a mix of both
  - In urban studies, we also understand the earth's geography through human-made political, economic, and social boundaries and points in space.
    - ➢ What are some examples of these?
- Since the late 1970s, GIS capabilities have rapidly developed and become far more accessible to a variety of users for a range of purposes.

➢ Brainstorm some uses for GIS as it pertains to urban studies and urban planning:

Principles of Geographic Information Systems: An introductory textbook, Eds. Otto Huisman and Rolf A. de By (2009). https://webapps.itc.utwente.nl/librarywww/papers_2009/general/principlesgis.pdf

- GIS = geographic information system
- Per Huisman & de Bay (2009), “a GIS is a computer-based system that provides the following four sets of capabilities to handle georeferenced data”:
  1. Data capture and preparation
  2. Data management, including storage and maintenance
  3. Data manipulation and analysis
  4. Data presentation

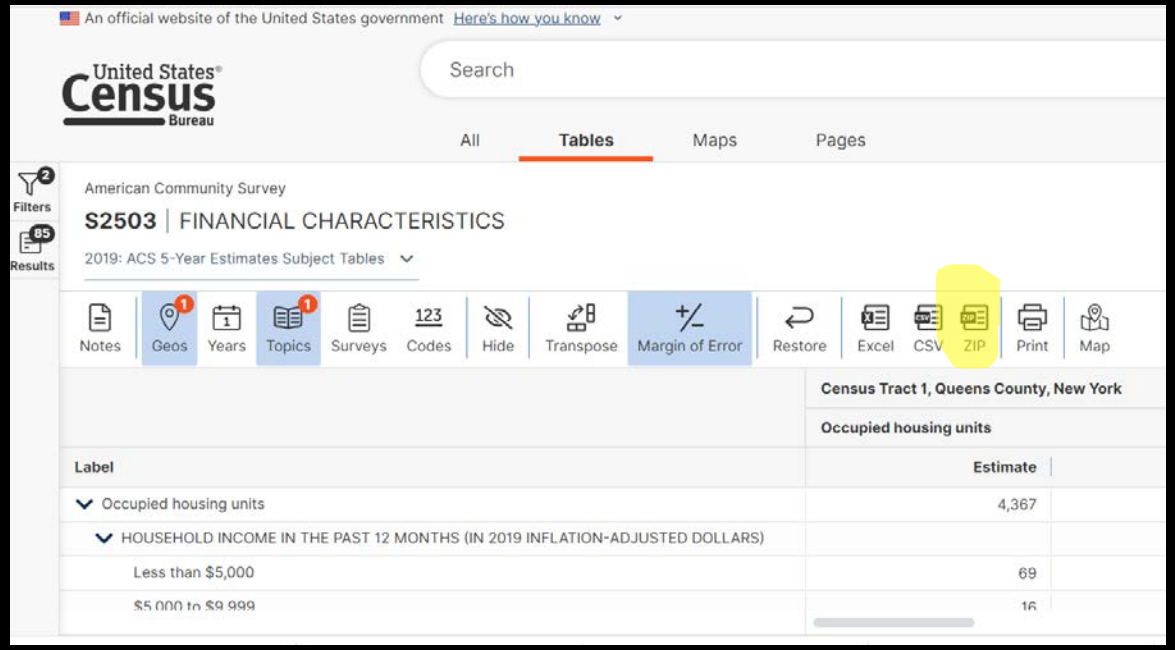
➢In URBST 200, how are we working with these capabilities?

Principles of Geographic Information Systems: An introductory textbook, Eds. Otto Huisman and Rolf A. de By (2009). https://webapps.itc.utwente.nl/librarywww/papers_2009/general/principlesgis.pdf

1. Data capture and preparation
2. Data management, including storage and maintenance
3. Data manipulation and analysis
4. Data presentation

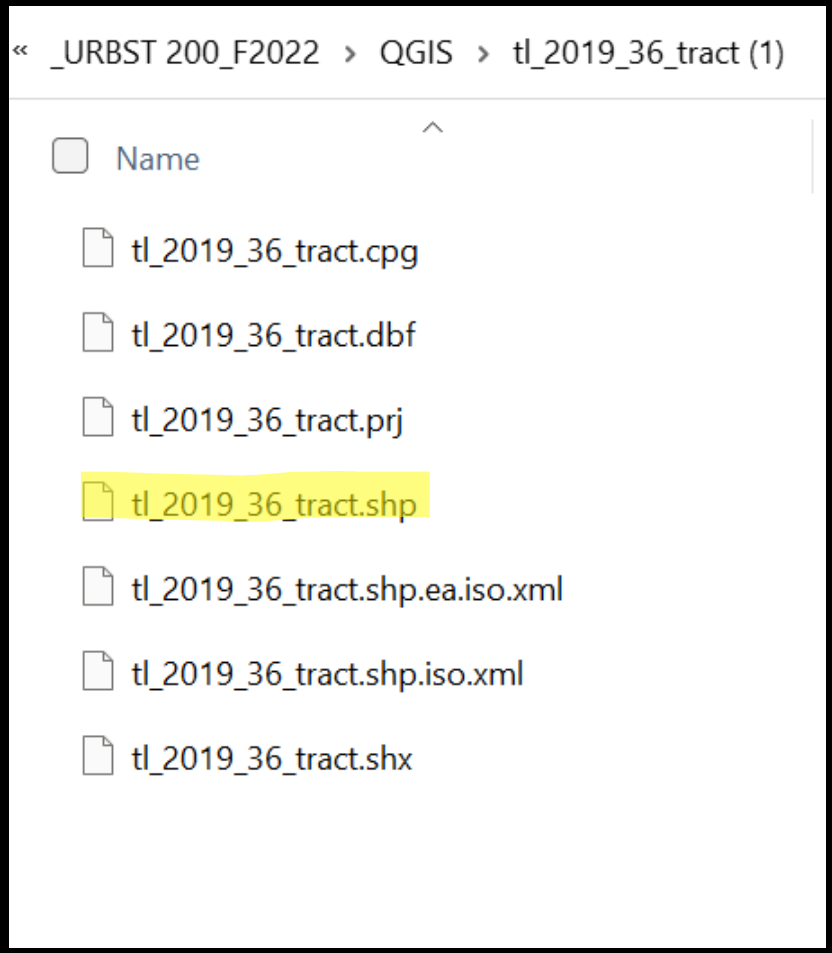
- Capture
  - Census
  - NYC Open Data
- Preparation
  - Cleaning up your data
  - Adding the correct GEOID

	A	B	C
1	GEO_ID	Join_GEO_ID	NAME
2	Geography		Geographic
3	1400000US36081000101	=MID(A3,10,LEN(A3))	Census Trac
4	1400000US36081000102		Census Trac
5	1400000US36081000103		Census Trac



1. Data capture and preparation
2. Data management, including storage and maintenance
3. Data manipulation and analysis
4. Data presentation

- Labeling your data files logically
- Keeping your files organized where they can be easily accessed for a QGIS project
  - This includes any shapefiles



1. Data capture and preparation
2. Data management, including storage and maintenance
3. Data manipulation and analysis
4. Data presentation

- Formatting data so it maps what you want , e.g.,
  - converting to percentages of a total
  - combining categories as needed
  - Identifying which data to map

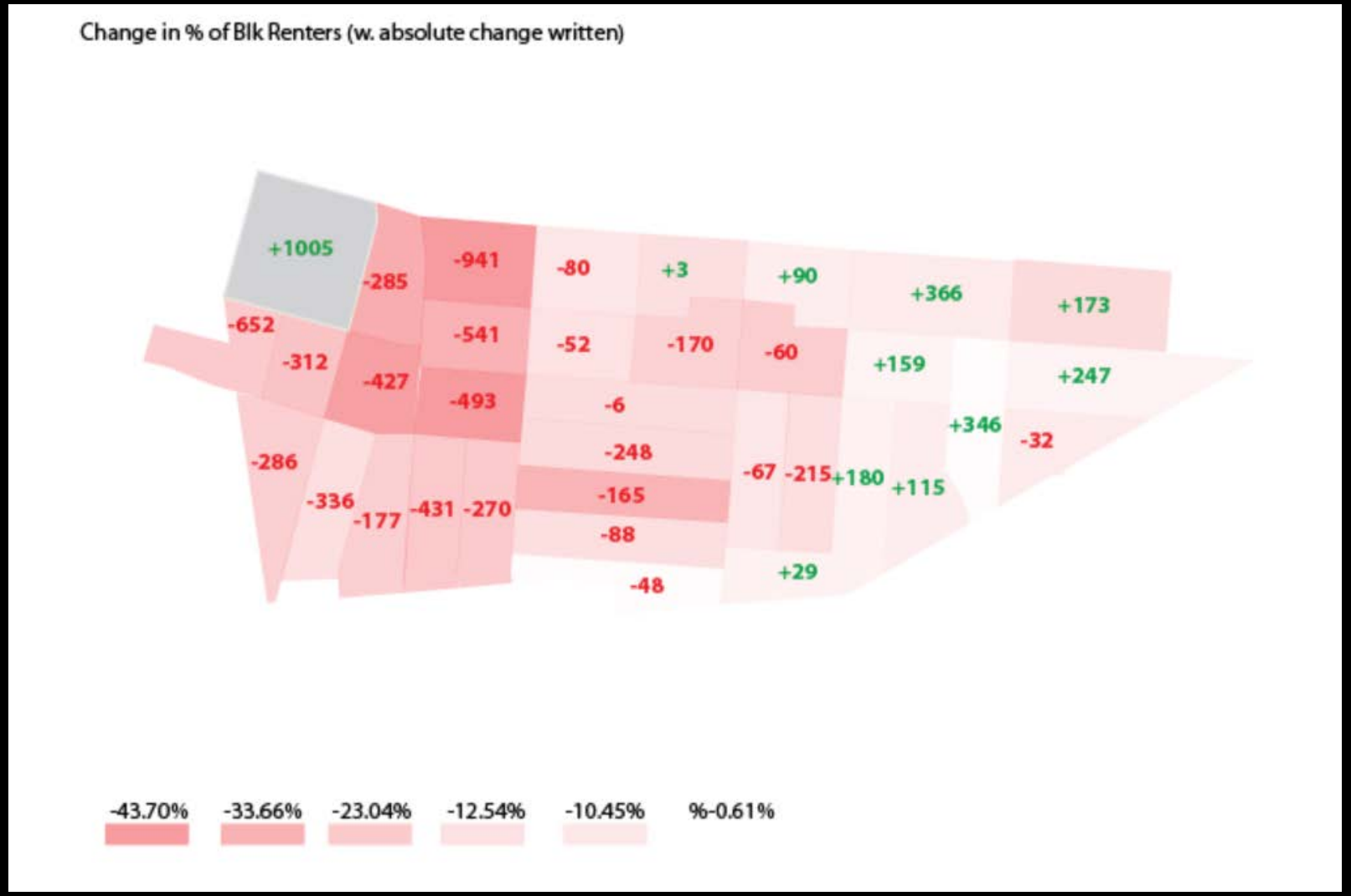
- Merging two time periods of data into one dataset to map them from one CSV file
- Avoid misinterpretation of data

Identifying which part of the dataset you are interested in mapping, if not all of it

AQ	AR	AS	AT	AU	AV	AW	AX	AY	BD	BE	BF	BG	BH
Owner's Business Type	Non-Profit	Owner's Busines	Owner's First Na	Owner's L	Owner's H	Owner's H	Owner’s H	Owner’s H	LATITUDE	LONGITUD	COUNCIL_	CENSUS_T	NTA_NAM
INDIVIDUAL	N	N/A	CHARLES	GANSA	1370	PACIFIC ST	BROOKLYN	NY	40.6774	-73.9452	36	313	Crown Hei
CORPORATION	Y	BROOKLYN PACT	DANIEL	MORITZ	1044	NORTHERI	ROSLYN	NY	40.67611	-73.9357	36	309	Crown Hei
CORPORATION	Y	BROOKLYN PACT	DANIEL	MORITZ	1044	NORTHERI	ROSLYN	NY	40.67685	-73.9352	36	309	Crown Hei
CONDO/CO-OP	N	N/A	SHIRA	BEERY	10	PLAZA STR	BROOKLYN	NY	40.67547	-73.9706	39	159	Park Slope
INDIVIDUAL	N	N/A	JOEL	FRIEDMAN	116	SANFORD	BROOKLYN	NY	40.6697	-73.9302	41	349	Crown Hei
INDIVIDUAL	N	N/A	JOEL	FRIEDMAN	116	SANFORD	BROOKLYN	NY	40.6697	-73.9302	41	349	Crown Hei
INDIVIDUAL	N	N/A	JOEL	FRIEDMAN	116	SANFORD	BROOKLYN	NY	40.6697	-73.9302	41	349	Crown Hei
INDIVIDUAL	N	N/A	JOEL	FRIEDMAN	116	SANFORD	BROOKLYN	NY	40.6697	-73.9302	41	349	Crown Hei
CONDO/CO-OP	N	N/A	JAKE	ABRAHAM	20	PLAZA STR	BROOKLYN	NY	40.67553	-73.9703	39	159	Park Slope
INDIVIDUAL	N	N/A	MORRIS	LIEBERMA	823	FRANKLIN	BROOKLYN	NY	40.67022	-73.9581	35	325	Crown Hei
CORPORATION	N	1260 ATLANTIC A	JAMES	COAKLEY	134	WEST 29 S	NEW YORI	NY	40.67852	-73.951	36	315	Crown Hei
INDIVIDUAL	N	ALGORITHM REA	BIANCA	ROBINSON	1110	ST. JOHNS	BROOKLYN	NY	40.67115	-73.9415	35	339	Crown Hei
CORPORATION	N	GUILD OF ATHEN	FRANCES	CAPERCHI	366	PARK PLAC	BROOKLYN	NY	40.676	-73.9655	35	207	Prospect H
OTHER	N	BEC NEW COMM	DAN	MATHEW	67	HANSON F	BROOKLYN	NY	40.67121	-73.9337	36	351	Crown Hei
PARTNERSHIP	N	USD 1370 DEAN	SETH	WEISSMAI	27	WEST 20 S	NEW YORI	NY	40.67651	-73.9433	36	313	Crown Hei
CORPORATION	N	1407 DEAN MAN	YEHUDA	COHEN	702	SAINT MA	BROOKLYN	NY	40.67648	-73.9424	36	313	Crown Hei
PARTNERSHIP	N	737 DEVELOPME	EYAL	OVADIA	298	TOMPKINS	BROOKLYN	NY	40.65948	-73.9379	41	87401	Prospect L
CORPORATION	N	1750 DEAN ST, LI	SHLOMI	COHEN	138-78	QUEENS B	BRIARWO	NY	40.67577	-73.9297	36	307	Crown Hei
CORPORATION	N	899 PACIFIC STRI	VINCENT	DING	899	PACIFIC ST	BROOKLYN	NY	40.68052	-73.9655	35	203	Prospect H
NYC AGENCY	N	DEPT OF HEALTH	SHEILA	BENJAMIN	42-09	28TH STRE	QUEENS	NY	40.67353	-73.9358	36	345	Crown Hei

nit_Issuance_Feb 2022

Avoid misinterpretation of data



1. Data capture and preparation
2. Data management, including storage and maintenance
3. Data manipulation and analysis
4. Data presentation

- Per Huisman & de Bay (2009), the following need to be addressed when presenting your data:
  - The message to communicate
    - Patterns, trends, deviations from a norm, etc.
  - The audience reading your data
  - The medium through which your data will be presented
    - Print (size, color?)
    - Digitally
    - Interactive
  - The rules of aesthetics
    - maps should be printed north-up
    - use clear georeferencing; with intuitive use of symbols etc.
    - Size and arrange title, scale bar, key etc. in clear and legible way

Challenges to Spatial Data, according to safegraph.com

1. Data standardization

- Data scientists and GIS analysts often spend up to 90% of their time just cleaning data before using it due to a lack of standardization in how the data collected.
- This can include differences across datasets in timestamps and time zones and/or units of measurements.
- A standard is also sometimes only as good as its adoption rate, and there can be barriers to this as well. For example, the standard's creator(s) may charge money, require re-sharing of data, or impose some other obligation that makes people and organizations hesitant to adopt that standard. And remember: a standard doesn't have to perfectly fit all cases; it just has to fit enough so that a critical mass of people or organizations agree to it and derive value from it.

- A good standard should allow your datasets to communicate easily with other datasets. To do this, it should be able to identify data points under a series of guidelines, often summarized as the "S.I.M.P.L.E." formula:
  - **Storable** – Data point IDs should be able to be stored in places that don't require Internet access.
  - **Immutable** – Data point IDs shouldn't change over time, except in extreme circumstances.
  - **Meticulous** – Data points should be uniquely identifiable across all systems they're in.
  - **Portable** – Standardized IDs should allow data points to smoothly transition from one storage system or dataset to another.
  - **Low-cost** – The standard should be inexpensive, or even free, to use for data transactions.
  - **Established** – The standard needs to cover almost all data points it could be applied to.

https://www.safegraph.com/guides/geospatial-data-integration-challenges

Challenges to Spatial Data, according to safegraph.com

2. Address standardization

- There are many different elements to addresses: street name, building unit number, city, region, country, mailing code, and so on.
- Some databases may not have these pieces of information in a standard order, or may not even have all of them. This can make it difficult for a computer program or algorithm to tell if two or more addresses point to the same location.
- In your dataset, do “US”, “USA”, “U.S.A.”, “the (United) States”, and even “America” all refer to the same country? Can it tell if the abbreviation “St.” stands for “street” or “saint”, and in which cases either one applies?

https://www.safegraph.com/guides/geospatial-data-integration-challenges

Challenges to Spatial Data, according to safegraph.com

3. Lack of institutional knowledge

- Traditionally, geospatial data and geographic information systems (GIS) have been in a class of their own, separate from data science or other engineering fields. So only a small group of people in these latter fields (about 5%) actually know how to work with geospatial data. It doesn't behave the same way as, say, tabular data, so many organizations struggle to ingest it into their workflows because there is a skills gap.
- To address this, organizations can look within and host webinars, hackathons, or meetups; attend conferences; or hire a specialized recruiting agency to attract contacts with specialized geospatial data know-how.
    - Ideally, you're going to want someone with strong programming skills and a background in statistics. They should also know how to make data products, visualizations, workflows, and pipelining routines. Finally, you'll want someone who's familiar with machine learning, distributed computing, and (obviously) GIS software.

https://www.safegraph.com/guides/geospatial-data-integration-challenges

Challenges to Spatial Data, according to safegraph.com

4. File size/processing times

- Like any type of data science analysis, geospatial analytics require the right systems and infrastructure.
- Depending on the size of datasets you're working with, basic tools like Excel [and QGIS] might be sufficient, or you need to invest in more robust data-management systems.

5. Data quality

- A lot of bad data exists, often caused by a lack of expertise in how to collect and process it, or just simple human error. [we're pretty safe with the Census and NYC OpenData]
- Four steps to check data before using it.
    1. Make sure it comes from reliable sources.
    2. Evaluate what it's capable of, including any gaps it may leave and any assumptions you might make about it.
    3. Determine how much work it will take to get the data ready for use.
    4. Based on what you know the data can (and can't) do, draw up a plan for what specific function(s) it will serve in your operations.

➢Other challenges to spatial data can be hound here

https://www.safegraph.com/guides/geospatial-data-integration-challenges

THE BAD NEWS: THE PITFALLS OF SPATIAL DATA (from O'Sullivan and Unwin, 2010, Ch. 2)

1. **Spatial autocorrelation** is a complicated name for the obvious fact that data from locations near one another in space are more likely to be similar than data from locations remote from one another.
   - This is one of the main things that separates statistical inferences made with spatial data from non-spatial data—once can't achieve a "random sample" of spatial data.
   - If spatial autocorrelation were not commonplace, then geographic analysis would be of little interest and geography would be irrelevant.
   - If geography is worth studying at all, it must be because phenomena do not vary randomly through space. The existence of spatial autocorrelation is therefore a given in geography. Unfortunately, it is also an impediment to the application of conventional statistics.

➢ In URBST 200 we are just doing descriptive statistics, although it is possible to perform inferential statistics and to map the likelihood of an outcome spatially, but that is beyond the scope of this course.

O'Sullivan, D. and Unwin, D. (2010) 'The pitfalls and potential of spatial data', in *Geographic Information Analysis*, Oxford University Press, pp. 33-41. – available here

Spatial autocorrelation – cont.

- Spatial autocorrelation introduces redundancy into data, so that each additional item of data provides less new information than is indicated by a simple assessment based on *n*, the sample size. This affects the calculation of confidence intervals and so forth.
  - Your spatial data may exhibit positive autocorrelation, negative autocorrelation (rare), and noncorrelation or zero autocorrelation.
- Describing and modeling patterns of variation across a study region, effectively describing the autocorrelation structure, is of primary importance in spatial analysis.
- There are two kinds of spatial variation:
  1. First-order spatial variation: occurs when observations across a study region vary from place to place due to changes in the underlying properties of the local environment.
     - For example, the rates of incidence of crime might vary spatially simply because of variations in the population density, such that they increase near the center of a large city.
  2. Second-order variation: is due to interaction effects between observations, such as the occurrence of crime in an area making it more likely that there will be crimes surrounding that area, perhaps in the shape of local "hotspots" in the vicinity of bars and clubs or near local street drug markets.

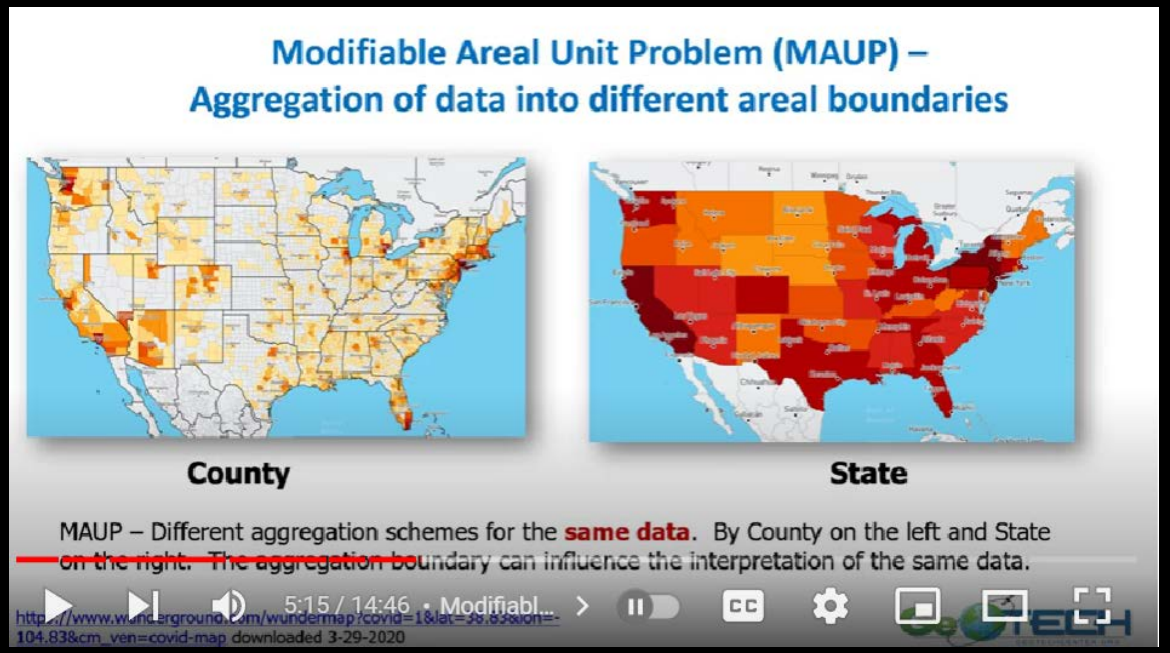
O'Sullivan, D. and Unwin, D. (2010) 'The pitfalls and potential of spatial data', in *Geographic Information Analysis*, Oxford University Press, pp. 33-41. – available here

2. Modifiable Area Unit Problem

- A major difficulty with spatial data is that they are often aggregates of data originally compiled at a more detailed level.
- The best example is a national census, which is collected at the household level but reported for practical and privacy reasons at various levels of aggregation such as city districts, counties, and states.
- The problem is that the aggregation units used are arbitrary with respect to the phenomena under investigation, yet the units used will affect statistics determined on the basis of data reported in this way.
  - This difficulty is referred to as the modifiable areal unit problem (MAUP).
- If the spatial units [e.g. census tract, NTA, school district, senate district etc.] in a particular study were specified differently, we might observe very different patterns and relationships.

O'Sullivan, D. and Unwin, D. (2010) 'The pitfalls and potential of spatial data', in *Geographic Information Analysis*, Oxford University Press, pp. 33-41. – available here

https://www.youtube.com/watch?v=CISjONu-5Qg&ab_channel=MichaelBuzzelli



https://www.youtube.com/watch?v=CwVxvOmx2AM&ab_channel=GeoTechCenterConceptModules

3. Ecological Fallacy

- Arises when a statistical relationship observed at one level of aggregation is assumed to hold because the same relationship holds at a more detailed level.
  - For example, we might observe a strong relationship between income and crime at the county level, with lower-income counties being associated with higher crime rates. If from this we conclude that lower-income individuals are more likely to commit a crime, then we are falling for the ecological fallacy.
  - In fact, it is only valid to say exactly what the data say: that lower-income counties tend to experience higher crime rates. What causes the observed effect may be something entirely different—perhaps lower-income families have less effective home security systems and are more prone to burglary (a relatively direct link); or lower-income areas are home to more chronic drug users who commit crimes irrespective of income (an indirect link); or the cause may have nothing at all to do with income.

O'Sullivan, D. and Unwin, D. (2010) 'The pitfalls and potential of spatial data', in *Geographic Information Analysis*, Oxford University Press, pp. 33-41. – available here

4. Scale

- The geographic scale at which we examine a phenomenon can affect the observations we make, and this must always be considered prior to spatial analysis.
  - For example, at the continental scale, a city is conveniently represented by a point. At the regional scale, it becomes an area object. At the local scale, the city becomes a complex collection of point, line, area, and network objects.
  - The scale we work at affects the representations we use, and this in turn is likely to have effects on spatial analysis; yet, in general, the correct or appropriate geographic scale for a study is impossible to determine beforehand, and due attention should be paid to this issue.

O'Sullivan, D. and Unwin, D. (2010) 'The pitfalls and potential of spatial data', in *Geographic Information Analysis*, Oxford University Press, pp. 33-41. – available here

5. Nonuniformity of Space and Edge Effects

- A significant issue distinguishing spatial analysis from conventional statistics is that space is not uniform.
  - For example, we might have data on crime locations gathered for a single police precinct. It is very easy to see patterns in such data...and the patterns may appear particularly strong if crime locations are mapped simply as points without reference to the underlying geography.
  - There will almost certainly be clusters simply as a result of where people live and work, and apparent gaps in (for example) parks or at major road intersections. These gaps and clusters are not unexpected but arise as a result of the nonuniformity of the urban space with respect to the phenomenon being mapped.
- A particular type of non uniformity problem, which is almost invariably encountered, is due to edge effects.
  - These arise where an artificial boundary is imposed on a study, often just to keep it manageable. The problem is that sites in the center of the study area can have nearby observations in all directions, whereas sites at the edges of the study area only have neighbors toward the center of the study area.
  - Unless the study area has been very carefully defined, it is unlikely that this reflects reality, and the artificially produced asymmetry in the data must be accounted for.

O'Sullivan, D. and Unwin, D. (2010) 'The pitfalls and potential of spatial data', in *Geographic Information Analysis*, Oxford University Press, pp. 33-41. – available here

Example of edge effects:

Definition of “patient area” when including and excluding offer and demand outside. Focus on the IRIS named “Fournes-en-Weppes”- (IRIS no. 592 500 000), the Nord department are circled in blue, whereas neighboring IRIS from the three departments of Somme, Aisne and Pas-de-Calais are yellow.
a) without consideration of offer and demand beyond the boundary;
b) with consideration of offer and demand beyond the boundary
(source: https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-017-0119-3)