GIS data, QGIS and application in Development

Etienne Madinier[1 2]

[1]Paris School of Economics - EHESS, [2]ENS de Lyon

October 2022

email: etienne.madinier@gmail.com

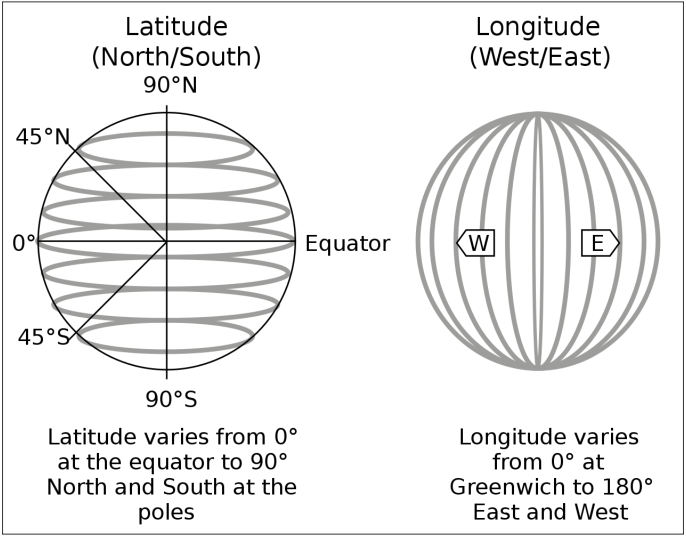
Any questions from last week?

Today

1. Coordinates and Projections
   - Coordinate Reference Systems
   - Projected coordinate systems
2. Measures and measurement errors
   - Vectors and rasters
   - Zonal statistics
   - Measurement errors
   - Distances and areas
3. Guidelines
   - Building a dataset
   - Group project
4. Tutorial

Geographic coordinate systems

- On the globe, each location is coded using degrees



Geographic coordinate systems

- This is sufficient if you want to overlay different maps/ different data sources
- Provided all maps are in the same coordinate systems
- **Exactly the same**

Geographic coordinate systems

- This is sufficient if you want to overlay different maps/ different data sources
- Provided all maps are in the same coordinate systems
- **Exactly the same**
- The Australian plate is moving towards the North East at a speed of about 7cm per year
- This is troublesome for some high precision applications (ex.: precision farming or in-vehicle navigation)
- The reference system has been updated on January 1, 2017 to reflect the 1.7m move of the plate since the implementation of Geocentric Datum of Australia 1994 (GDA94)
- new datum: Geocentric Datum of Australia 2020 (GDA2020)
- From 2020 onwards, a time-dependent frame: Australian Terrestrial Reference Frame (ATRF)

Geographic coordinate systems: WGS1984

- The World Geodetic System (1984) is the most common geographic coordinate system
- Used in cartography, by satellite navigation (including GPS)
- Many remote sensing product and more generally gridded data are distributed in WGS84

- The system corresponds to a model of the earth
- The date corresponds to the year in which the model has been fixed.

Geographic coordinate systems: WGS1984

Coordinates of origin	close to the Earth center of mass (uncertainty around 2cm)
Meridian of zero longitude	102m east of Greenwich meridian
Semi-major axis	6 378 137.0 m
Semi-minor axis	6 356 752.314 245 m
Flattening	1/298.257 223 563
Circumference at the equator	≃40 075.017 km
Length of a meridian	≃40 007.863 km

Geographic coordinate systems

Assuming that you have two sets of data in the same geographic coordinate system

- you can easily match the features of the two layers
- all you need is to "overlay the maps"

Figure 1: Housing centers in France

(a) Housing centers localisation and migratory routes

(b) Refugee center openings by commuting zones between 2004 and 2012

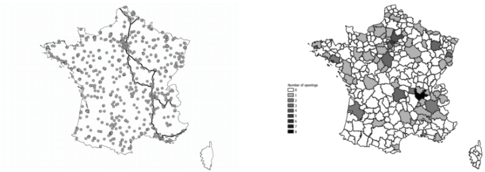


Figure 1a shows the location of all housing center for refugee in France in April 2018 and the migratory routes (from IOM - monitoring flows). Figure 1b is a map of the French commuting zones. White areas are commuting zones that did not have any opening between 2004 and 2012. Colored areas are commuting zones that had at least one opening during the period.

Projected coordinate systems

- For many applications, geographic coordinate systems are not sufficient

Projected coordinate systems

- For many applications, geographic coordinate systems are not sufficient
- 1° in latitude is, at the equator, is 110.6km and 111.7km at poles
- 1° in longitude is, at the equator, is 111.3km, 55.8km at 60° and (almost) 0km at poles

Projection systems: why it matters for economists

- overlay different data sources
- computation of distance, with the disclaimer that some applications do take into account the curvature of the earth

Projection systems: why it matters for economists

- overlay different data sources
- computation of distance, with the disclaimer that some applications do take into account the curvature of the earth
- computation of areas

Projection systems: why it matters for economists

- overlay different data sources
- computation of distance, with the disclaimer that some applications do take into account the curvature of the earth
- computation of areas
- avoid to puzzle the audience with a strange map

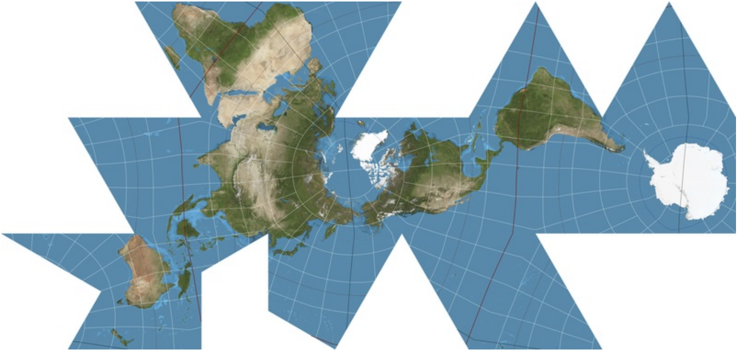


Figure: Dymaxion projection by Buckminster Fuller (1943), based on the projection of earth surface on an icosahedron

Projection systems: basic theory

- a (map) projection system is a function linking the earth surface to a plane or to a developable surface

Projection systems: basic theory

- a (map) projection system is a function linking the earth surface to a plane or to a developable surface
- $x = f_1(\phi, \lambda)$ and $y = f_2(\phi, \lambda)$

x, y are coordinates on the plane

ϕ, λ are coordinates on the earth

f_1, f_2 are continuous functions, except on few points and lines

Projection systems: basic theory

- a (map) projection system is a function linking the earth surface to a plane or to a developable surface
- $x = f_1(\phi, \lambda)$ and $y = f_2(\phi, \lambda)$

x, y are coordinates on the plane

ϕ, λ are coordinates on the earth

f_1, f_2 are continuous functions, except on few points and lines

- infinite number of potential functions

Projection systems: basic theory

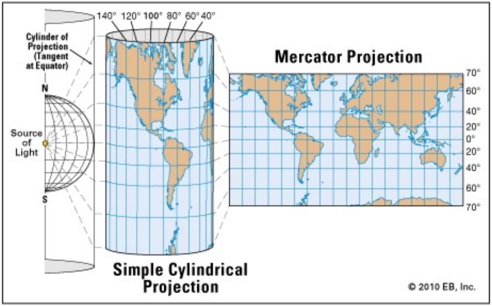
- The main properties:
  - area preserving
  - shape preserving
  - equidistant (in one dimension)

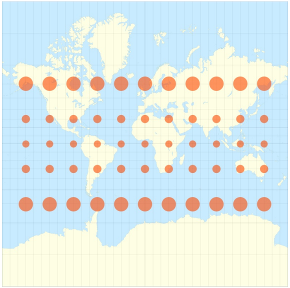
Projection systems: basic theory

- The main properties:
  - area preserving
  - shape preserving
  - equidistant (in one dimension)
- There is no projection system able to preserve the 3 properties (nor the first 2 together)

Projection systems: main types

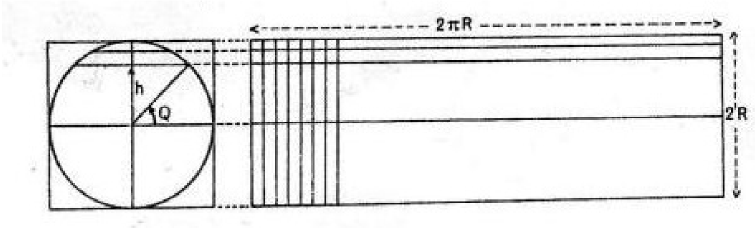
Cylindrical projection: Mercator projection (1569)

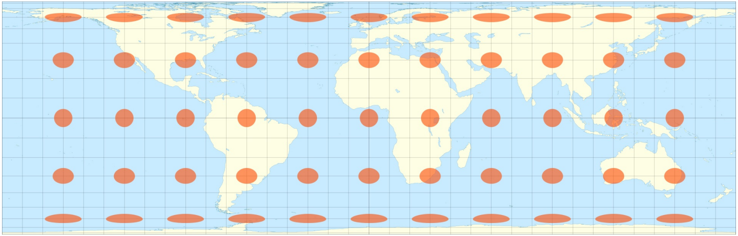




Projection systems: main types

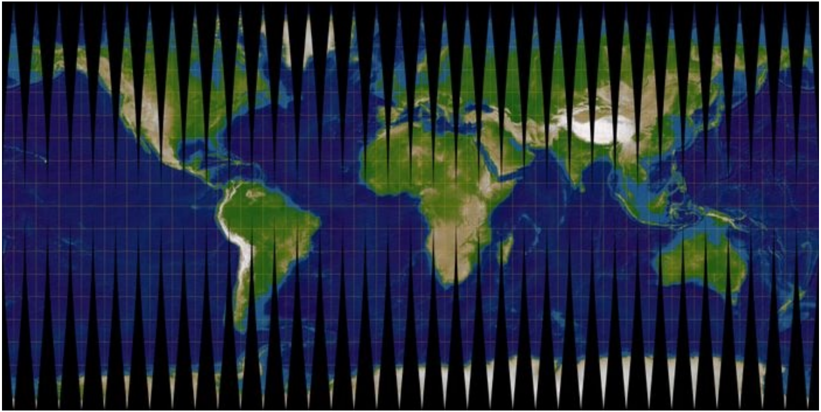
Cylindrical projection: Lambert cylindrical equal-area (1772)





Projection systems: main types

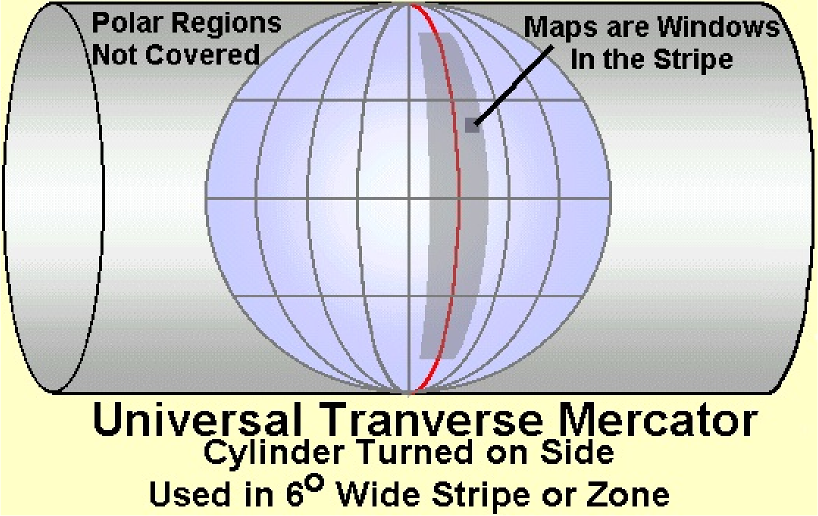
Cylindrical projection: UTM projection
→ The tangency line is chosen at any meridian



Projection systems: main types

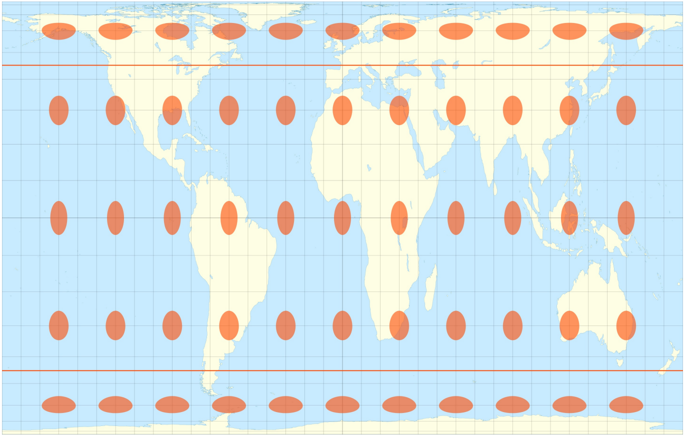
Cylindrical projection: UTM projection

→ The tangency line is chosen at any meridian



Projection systems: main types

Cylindrical projection: Gall-Peters cylindrical equal-area (1855)
→ The projection relies on two standard parallels at $45°N$ and $45°S$



Projection systems: main types

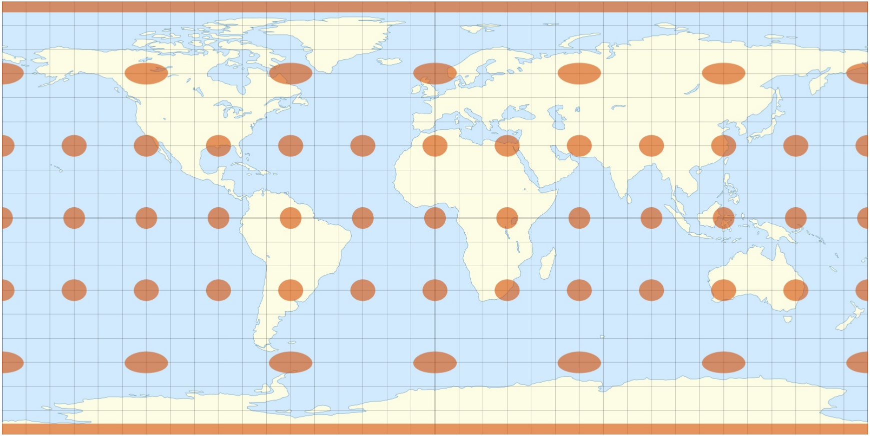
Cylindrical projection: equidistant cylindrical projection, attributed to Marinus of Tyr (AD 100) , with the special case of the “platte carrée”

- The projection maps meridian and circles of latitude to, respectively, parallel vertical and parallel horizontal lines.
- The projection is neither equal area nor conformal
- equidistant along meridians

→ Little use in navigation or cadastral mapping but very often used to visualize the world

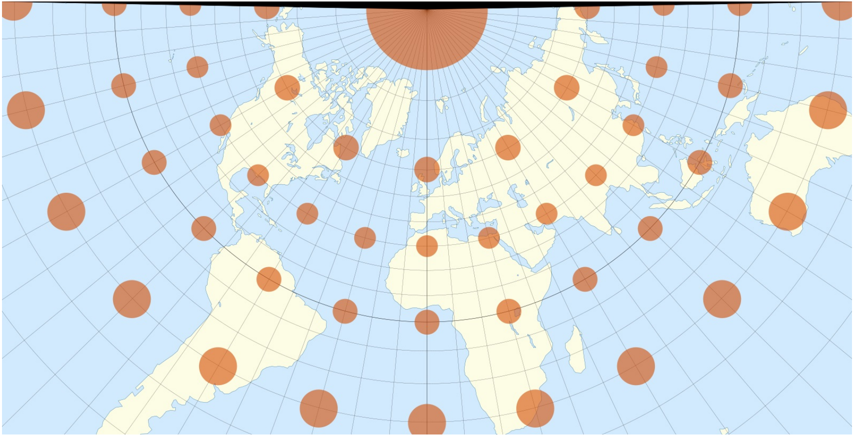
Projection systems: main types

equidistant cylindrical projection



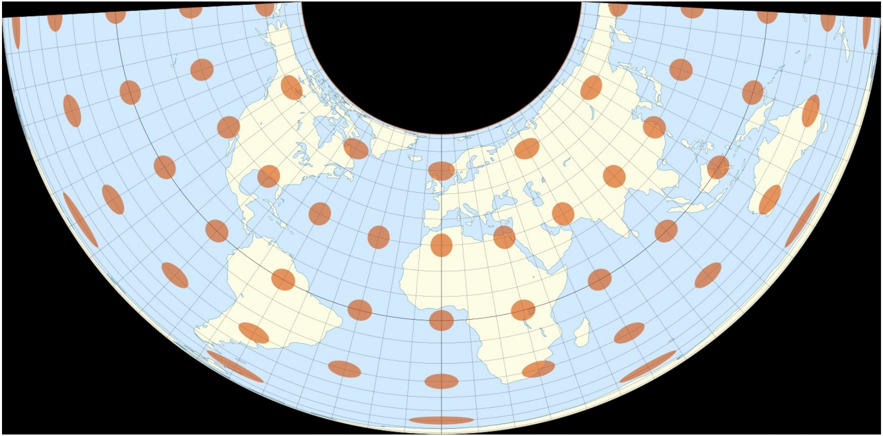
Projection systems: main types

Conic projection: Lambert conformal conic projection (1772)
Very much used for regional or national mapping systems. One tangency line or two standard parallels Tissot indicator:



Projection systems: main types

Conic projection: Albers equal-area conic projection (1805)



Projection systems: main types

Many other projection systems...

Projection systems: crude wrap-up for economists

Pay special attention when a software reprojects “on-the-fly”

- QGIS displays WGS84 using “platte carrée” projection

Projection systems: crude wrap-up for economists

Pay special attention when a software reprojects “on-the-fly”

- QGIS displays WGS84 using “platte carrée” projection
- The first map you upload in some software (QGIS, ARCGIS) will determine how the next ones will be displayed
- Some operations do “virtually” reproject maps stored in different coordinate systems before executing further commands
- Risky business for you because most operations do not

Projection systems: crude wrap-up for economists

Pay special attention when a software reprojects “on-the-fly”

- QGIS displays WGS84 using “platte carrée” projection
- The first map you upload in some software (QGIS, ARCGIS) will determine how the next ones will be displayed
- Some operations do “virtually” reproject maps stored in different coordinate systems before executing further commands
- Risky business for you because most operations do not
- Work with a consistent projection system for each set of operation

Projection systems: crude wrap-up for economists

How to identify coordinate systems?

- check the properties of your layer
- name and geodetic datum
- EPSG code: European Petroleum Survey Group (EPSG, 1986-2005) distributes a database of geodetic parameter sets ◂ EPSG or ◂ EPSG
- The most common coordinate systems are available in your favourite GIS software

Projection systems: crude wrap-up for economists

- Which projection should you choose?

Projection systems: crude wrap-up for economists

- Which projection should you choose?
- Be consistent. Use the same projected coordinate system all along

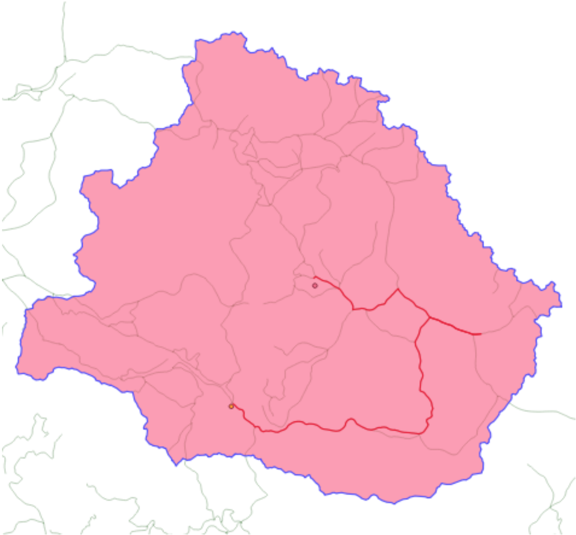
Projection systems: crude wrap-up for economists

- Which projection should you choose?
- Be consistent. Use the same projected coordinate system all along
- When working on a specific area, try to have a projection system with a tangency or a secant line close by. Distortion will be minimal for most of our applications, as economists

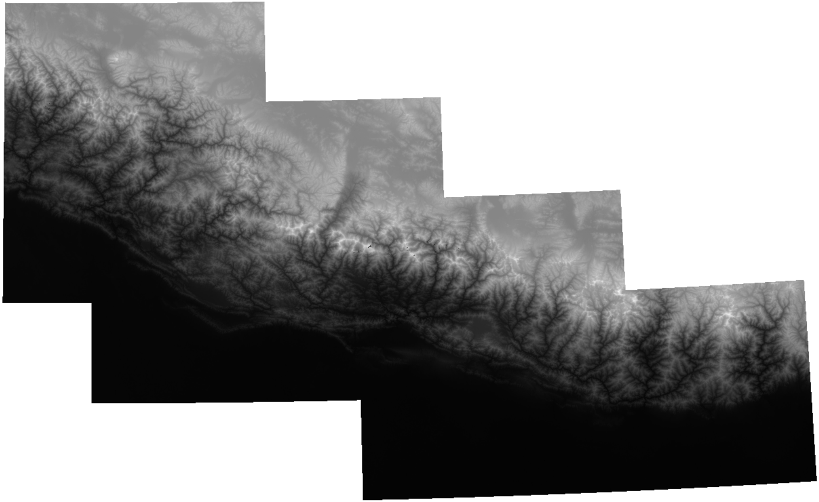
Projection systems: crude wrap-up for economists

- Which projection should you choose?
- Be consistent. Use the same projected coordinate system all along
- When working on a specific area, try to have a projection system with a tangency or a secant line close by. Distortion will be minimal for most of our applications, as economists
- Especially if you work at the world level, pay special attention to what matters for you
  - areas may be extremely important if you work on food production and population density
  - distances are very important if you work on trade routes. This is however challenging because it is not possible to preserve distances in all directions.
  - shapes may matter more for some applications.

Vector data

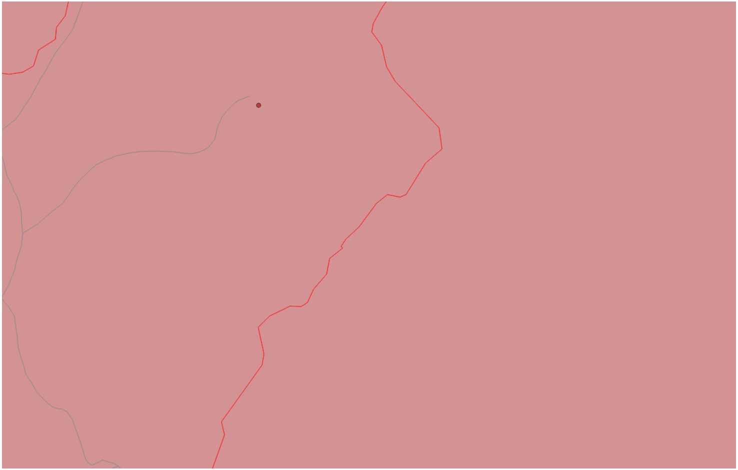


Raster data



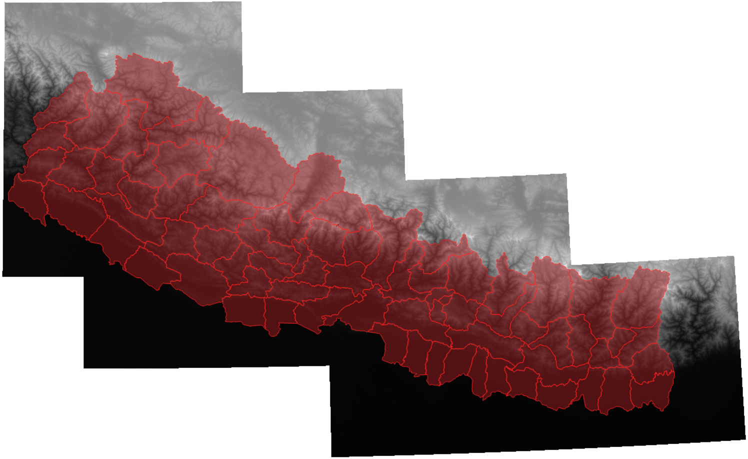
Resolution

Be careful: Vecotr data are “infinitely precise”, but this is artificial



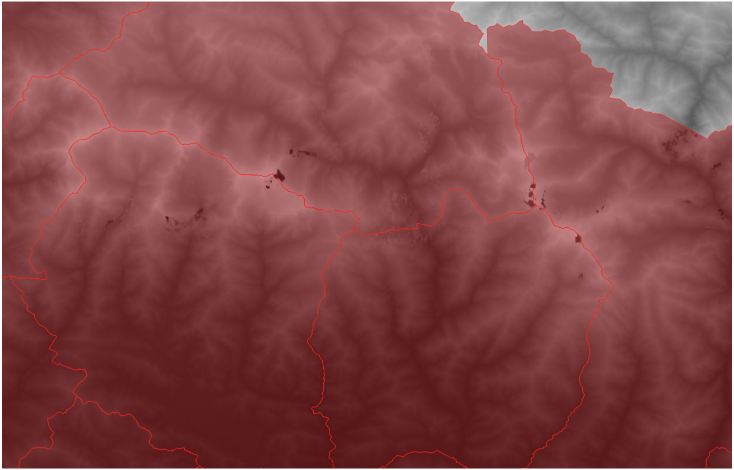
Resolution

Be careful: Resolution has slightly different meaning between vector and raster data



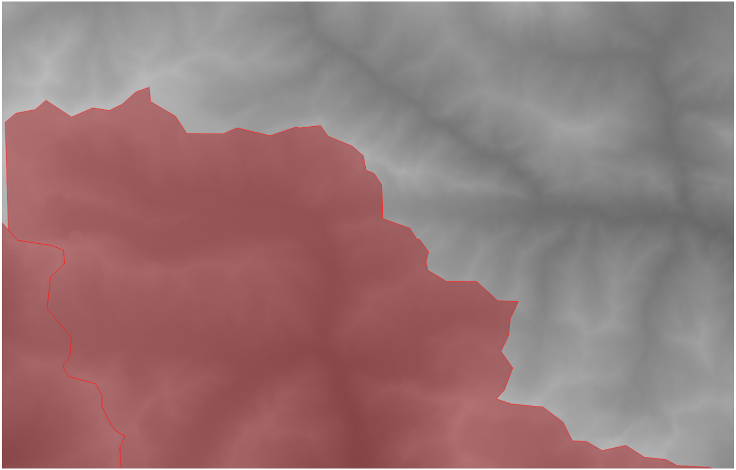
Resolution

Be careful: Resolution has slightly different meaning between vector and raster data



Resolution

Be careful: Resolution has slightly different meaning between vector and raster data



Zonal statistics

- Zonal statistics are extremely important and of widespread use in many applications in economics
  - average rainfall in a district
  - land quality around a household location
  - deforestation in or around a forest concessions
  - land use in a country
  - nighttime light over a grid cell

Zonal statistics

- Zonal statistics are extremely important and of widespread use in many applications in economics
  - average rainfall in a district
  - land quality around a household location
  - deforestation in or around a forest concessions
  - land use in a country
  - nighttime light over a grid cell
- Zonal statistics are summary statistics of raster data value within each polygon
  - mean, standard deviation
  - median
  - min, max, range
  - total
  - count

Zonal statistics

- Zonal statistics are extremely important and of widespread use in many applications in economics
  - average rainfall in a district
  - land quality around a household location
  - deforestation in or around a forest concessions
  - land use in a country
  - nighttime light over a grid cell
- Zonal statistics are summary statistics of raster data value within each polygon
  - mean, standard deviation
  - median
  - min, max, range
  - total
  - count
  - for integer rasters: median, variety, majority and minority

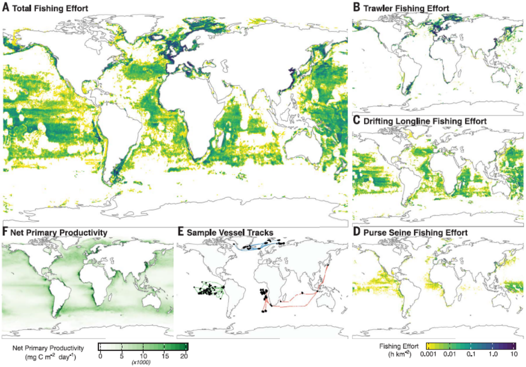
Remember ?

Some measurement errors result from much more technical aspect, and should not be neglected !

Levels of aggregation

- Kroodsma et al. (2018, Science) find that industrial fishing occurs in > 55% of ocean area and has a spatial extent more than four times that of agriculture.

Fig. 1. The spatial footprint of fishing. (**A** to **D**) Total fishing effort [hours fished per square kilometer ($h\ km^{-2}$)] in 2016 by all vessels with AIS systems (A), trawlers (B), drifting longliners (C), and purse seiners (D). (**E**) Examples of individual tracks of a trawler (blue), a longliner (red), and a purse seiner (green). Black symbols show fishing locations for these vessels, as detected by the neural network, and colored lines are AIS tracks. (**F**) Global patterns of average annual NPP [expressed as milligrams of carbon uptake per square meter per day ($mg\ C\ m^{-2}\ day^{-1}$)] are shown for reference.



Levels of aggregation

- Amoroso et al. (2018, Science) re-analyse the same dataset and reply that, depending on the level of analysis, higher resolution data reduced footprint estimates by factors of > 10 and > 5, for all vessels and (regional) trawlers respectively.

Levels of aggregation

- Amoroso et al. (2018, Science) re-analyse the same dataset and reply that, depending on the level of analysis, higher resolution data reduced footprint estimates by factors of > 10 and > 5, for all vessels and (regional) trawlers respectively.

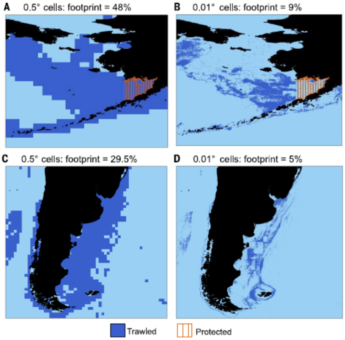


Fig. 1. Effect of grid resolution on the perception of fishing footprint. The areas in dark blue show the trawling footprints estimated for 2016 with (**A** and **C**) an equal-area grid with 0.5° resolution at the equator; (**B** and **D**) an equal-area grid with 0.01° resolution at the equator. The hatched area shows an example region of the North Pacific where all trawling was prohibited.

Levels of aggregation

- zonal statistics are perfect if gridded data are at a much finer resolution than the polygon layer
- what if coarser gridded data?
  - if you have points and all that matter is the value of the grid below
    → Extract (multi)values to points

Levels of aggregation

- zonal statistics are perfect if gridded data are at a much finer resolution than the polygon layer
- what if coarser gridded data?
  - if you have points and all that matter is the value of the grid below
    → Extract (multi)values to points
  - if you have polygons... **only pixels which have at least 50% of their area in the polygon are taken into account.**
    → Potential solution: transform pixels into polygons and intersect the two layers to have a wieghted average

Intersect or union

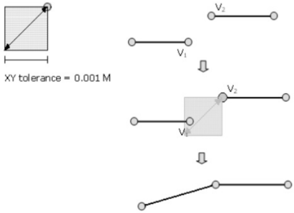
- intersect keeps only features existing on both layers
- if you want to keep unmatched features, use **union**

Intersect or union

- intersect keeps only features existing on both layers
- if you want to keep unmatched features, use **union**
- Some polygons layer may contain polygons with some overlap (lack of precision)
- Intersect or union will create very tiny polygons along borders

Intersect or union

- intersect keeps only features existing on both layers
- if you want to keep unmatched features, use **union**
- Some polygons layer may contain polygons with some overlap (lack of precision)
- Intersect or union will create very tiny polygons along borders
- Set XY tolerance at a higher level
- The tolerance defines the size of a square treated as a point in geoprocessing



Buffers

Buffer can be use to create weighted neighbourhood matrices

Projection systems

Through projections some measure errors could correlate, specially if you focus on distances or areas

Distance: on the map

On the globe, knowing the geographic coordinates, use the Great Circle Distance formula

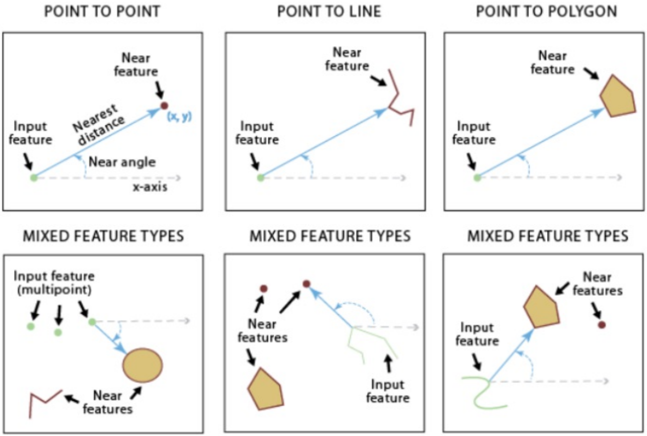
$$d_{ij} = 111.12 \times \cos^{-1}\left[\sin(La_i)\sin(La_j) + \cos(La_i)\cos(La_j)\cos(Lo_i - Lo_j)\right] \quad (1)$$

with d_{ij} the distance in kilometres between i and j

using a projected coordinate system, use the distance matrix in QGIS

Distance: on the map

- distances can be computed between points and polylines
- examples: distance to roads, railway lines, shoreline, borders,...
- use QGIS nearest neighbour analysis

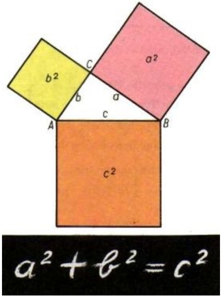


Distance and area: the world is not smooth

- In many applications in economics, we do need to compute areas
- The easy way: add field in the attribute table of your polygon layer
- Parcel area computed in this may differ from what you get from cadastral maps.

Distance and area: the world is not smooth

- In many applications in economics, we do need to compute areas
- The easy way: add field in the attribute table of your polygon layer
- Parcel area computed in this may differ from what you get from cadastral maps.
- The world is not smooth...



Distance and area: the world is not smooth

- When needed, you should adjust your area computation using a global digital elevation model (GEDM)
- For distance, it may also matter but if elevation matters, you may want to compute travel time or optimal route using the digital elevation model

Distance and area: the world is not smooth

- When needed, you should adjust your area computation using a global digital elevation model (GEDM)
- For distance, it may also matter but if elevation matters, you may want to compute travel time or optimal route using the digital elevation model

- Computing accurate distances may require to take into account traveling costs such as
  - frontiers
  - rivers
  - vehicles
  - asphalte

Distance to freedom

The promise of freedom: fertility decisions and the escape from slavery by Treb Allen published in the *Review of Economics and Statistics*, 2015

- The Fugitive State Law (1850) increased the legal protection of slaveholders attempting to capture fugitive slaves who escaped to the North, levied large fines on anyone aiding fugitives, and removed virtually all legal rights of captured fugitives

Distance to freedom

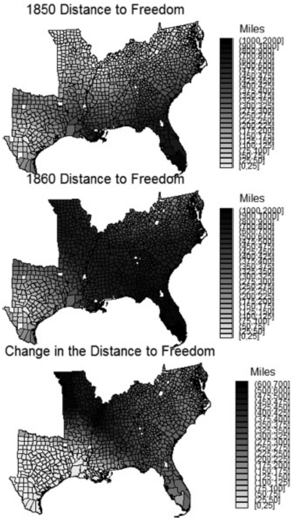
The promise of freedom: fertility decisions and the escape from slavery by Treb Allen published in the *Review of Economics and Statistics*, 2015

- The Fugitive State Law (1850) increased the legal protection of slaveholders attempting to capture fugitive slaves who escaped to the North, levied large fines on anyone aiding fugitives, and removed virtually all legal rights of captured fugitives
- Why it matters?

  My wife was torn from her mother's embrace in childhood, and taken to a distant part of the country. She had seen so many other children separated from their parents in this cruel manner, that the mere thought of her ever becoming the mother of a child, to linger out a miserable existence under the wretched system of American slavery, appeared to fill her very soul with horror; and as she had taken what I felt to be an important view of her condition, I did not, at first, press the marriage, but agreed to

Distance to freedom

Figure 1.—The Distance to Freedom



The figure indicates the change in the great circle distance to freedom before and after the passage of Fugitive Slave Law of 1850. Prior to the passage of the law, the distance to freedom is defined as the smaller of the distance to the nearest county in the U.S. North and the distance to Mexico. After the passage of the law, the distance to freedom is defined as the smaller of the distance to the nearest entry point into Canada and the distance to Mexico.

Distance to freedom

TABLE 3.—EFFECT OF DISTANCE TO FREEDOM ON SLAVE FERTILITY

	(1)	(2)	(3)	(4)	(5)
Distance to freedom (100 miles)	−0.010	−0.036***	−0.048**	−0.068***	−0.062*
	(0.009)	(0.014)	(0.020)	(0.026)	(0.036)
Distance to freedom squared	0.002	0.004*	0.004*	0.004*	0.004
	(0.001)	(0.002)	(0.002)	(0.003)	(0.004)
Plantation and county controls	Yes	Yes	Yes	Yes	Yes
County fixed effects	No	Yes	Yes	Yes	Yes
Year fixed effects	No	No	Yes	Yes	Yes
Controls × Year	No	No	No	Yes	Yes
State-year fixed effects	No	No	No	No	Yes
Observations	26,481	26,481	26,481	26,481	26,481
R^2	0.594	0.613	0.613	0.614	0.614
Mean fertility	0.814	0.814	0.814	0.814	0.814

The dependent variable is the ratio of the number of children younger than age 5 to the number of women aged 16 to 44 on a plantation. Each observation is a plantation with at least one women aged 16 to 44 in a particular year; plantations are weighted by the number of slaves using a GLS procedure that allows an arbitrary correlation in fertility within a plantation. Plantation controls include the total number of slaves; indicator variables for whether there were one, two, three, four, or five to ten slaves in the household; the fraction of potential mothers under the age of 30, the fraction of adults who were male; an indicator for whether the plantation was rural; and the sex of the slaveholder. County controls include whether the plantation was rural; the value of cotton, rice, tobacco, sugar, and total agricultural production; the change in total agricultural production over the past ten years; the adult white and slave populations; the change in the white and slave populations over the past ten years, and indicator variables for whether there was railroad or water transportation access in the county. Since county-level data are missing for some counties, missing variables are set equal to 0 and indicator variables for missing values are included to control for level differences. Controls × Year indicates that interactions with all demographic and county control variables and the year 1860 are included to allow the control variables to have different effects in 1850 and 1860. Standard errors allowing for spatial correlation between plantations (using a Barlett kernel with a 100-mile bandwidth) are reported in parentheses. Statistically significant at *10%, **5%, ***1%.

Distance to freedom

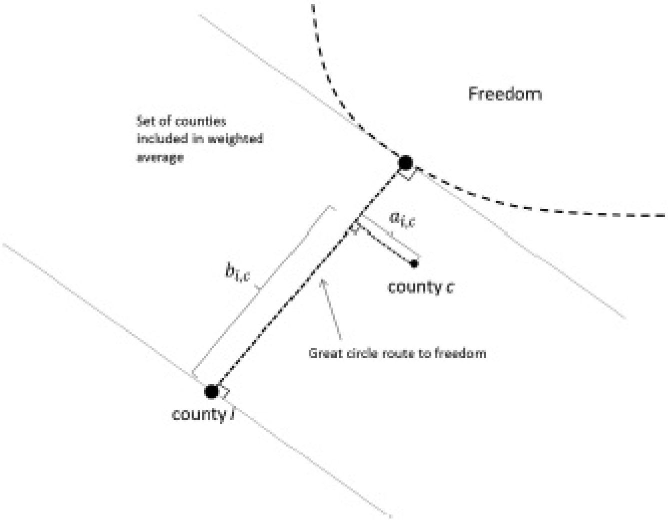
- the great circle distance is a proxy of the travel distance for fugitive slaves
- rivers, railroads, roads matter
- but also the ability to get support all the way through

Distance to freedom

- the great circle distance is a proxy of the travel distance for fugitive slaves
- rivers, railroads, roads matter
- but also the ability to get support all the way through
- it matters if potential fugitives have a reasonable knowledge of these sets of observations

Distance to freedom

FIGURE 2.—CALCULATING THE DIFFICULTY OF THE ROUTE



This figure presents the methodology of determining the difficulty of the route to freedom. The difficulty

Distance to freedom

Table 5.—Difficulty of Route

	(1)	(2)	(3)	(4)	(5)	(6)
Distance to freedom (100 miles)	−0.092**	−0.066**	−0.043*	−0.067***	−0.062**	−0.075*
	(0.039)	(0.029)	(0.025)	(0.026)	(0.026)	(0.042)
Distance to freedom squared	0.008*	0.004	0.004	0.004	0.003	0.006
	(0.004)	(0.003)	(0.002)	(0.003)	(0.003)	(0.004)
Distance × Railroads	−0.015					−0.021
	(0.014)					(0.014)
Distance × Waterways		0.001				−0.001
		(0.006)				(0.008)
Distance × Support of Slavery			−0.029***			−0.038***
			(0.008)			(0.009)
Distance × Free Black Fraction				−0.004		−0.009
				(0.006)		(0.007)
Distance × Slave Fraction					−0.009	−0.016
					(0.011)	(0.012)
Plantation and county controls	Yes	Yes	Yes	Yes	Yes	Yes
County fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Controls × Year	Yes	Yes	Yes	Yes	Yes	Yes
Observations	26,467	26,467	26,431	26,467	26,467	26,431
R^2	0.614	0.614	0.614	0.614	0.614	0.614
Mean Fertility	0.814	0.814	0.814	0.814	0.814	0.814

The dependent variable is the ratio of the number of children younger than age 5 to the number of women aged 16 to 44 on a plantation. Each observation is a plantation with at least one women aged 16 to 44 in a particular year; plantations are weighted by the number of slaves using a GLS procedure that allows an arbitrary correlation in fertility within a plantation. For comparability across regressions, the difficulty variables are Normalized to have a mean of 0 and standard deviation of 1. The difficulty variables are included in levels as well to control for any direct correlation between fertility and the difficulty of the route. Plantation controls include the total number of slaves; indicator variables for whether there were one, two, three, four, or five to ten slaves in the household; the fraction of potential mothers under the age of thirty, the fraction of adults who were male; an indicator for whether the plantation was rural; and the sex of the slaveholder. County controls include whether the plantation was rural; the value of cotton, rice, tobacco, sugar, and total agricultural production; the change in total agricultural production over the past ten years; the adult white and slave populations; the change in the white and slave populations over the past ten years; and indicator variables for whether there was railroad or water transportation access in the county. Since county-level data are missing for some counties, missing variables are set equal to zero and indicator variables for missing values are included to control for level differences. Controls × Year indicates that interactions with all demographic and county control variables and the year 1860 are included to allow the control variables to have different effects in 1850 and 1860. Standard errors allowing for spatial correlation between plantations (using a Barlett kernel with a 100-mile bandwidth) are reported in parentheses. Statistically significant at *10%, **5%, ***1%.

Distance: on the map

- distances can be computed between points and polygons
  - to the edge of the polygon ⇒ nearest neighbour tool
  - to the centroid of the polygon ⇒ get the centroids of the polygon and compute the distance between points

Some general guidelines...

- Database construction is an art, not a science

Some general guidelines...

- Database construction is an art, not a science
- There is a wide range of taste on how to proceed (and not every one likes oil on canvas)

Some general guidelines...

- Database construction is an art, not a science
- There is a wide range of taste on how to proceed (and not every one likes oil on canvas)
- Database construction is the most important step in the life of an empirical paper. However no one, but you, will ever look and care about this work

Some general guidelines...

- Database construction is an art, not a science
- There is a wide range of taste on how to proceed (and not every one likes oil on canvas)
- Database construction is the most important step in the life of an empirical paper. However no one, but you, will ever look and care about this work
- No one, but you and your co-authors
- It becomes more important as the profession (slowly) pays attention to the ability to replicate results

Some general guidelines...

- Database construction is an art, not a science
- There is a wide range of taste on how to proceed (and not every one likes oil on canvas)
- Database construction is the most important step in the life of an empirical paper. However no one, but you, will ever look and care about this work
- No one, but you and your co-authors
- It becomes more important as the profession (slowly) pays attention to the ability to replicate results
- **reproducibility**
- **robustness**
- **intelligibility / memory**
- **modularity**

Some more precise guidelines

- Automation
  - Automate everything that can be automated
    - At the beginning it may look as time waste
    - If it is not automated, keep track of all decisions that were made
  - Write a single script that executes all code from beginning to end
    - The beginning and the end still need to be defined...

Some more precise guidelines

- Automation
  - Automate everything that can be automated
    - At the beginning it may look as time waste
    - If it is not automated, keep track of all decisions that were made
  - Write a single script that executes all code from beginning to end
    - The beginning and the end still need to be defined...
- Directories
  - create separate directories for rawdata, temporary files, clean data and analysis
  - input and output files should not be at the same place
  - make directories portable = use globals and relative paths
  - use / instead of \

Some more precise guidelines

- Layers/Keys
  - create some files with identifiers at different levels.
    - Suppose you have data at the individual, household, village and district level
    - Then create at least 4 files with a comprehensive list of observations for each level of observation

Some more precise guidelines

- Layers/Keys
  - create some files with identifiers at different levels.
    - Suppose you have data at the individual, household, village and district level
    - Then create at least 4 files with a comprehensive list of observations for each level of observation
- Variable names
  - unique
  - meaningful
  - short, as far as possible
  - prefixes and suffixes are very useful and should separated by "_" or "-" in a consistent way

Some more precise guidelines

- Memory
  - Documentation is useful if well maintained and updated
  - Documentation should therefore cover only the very important issues (and things which are not automated)
  - Do not forget that comments within the code can explain what you have done “by hand”
  - The code should be self-containing, for you and for others...
  - Label variables as soon as the name of the variable is not sufficient (**hhsize** might self containing but **hhsize_adj** will be a nightmare 3 years from now)

Some more precise guidelines

- Memory
  - Documentation is useful if well maintained and updated
  - Documentation should therefore cover only the very important issues (and things which are not automated)
  - Do not forget that comments within the code can explain what you have done "by hand"
  - The code should be self-containing, for you and for others...
  - Label variables as soon as the name of the variable is not sufficient (**hhsize** might self containing but **hhsize_adj** will be a nightmare 3 years from now)
- Random comments
  - slow code should not run each time you test a few command lines
  - use spacing and indentation. Your code will be much nicer
  - use files.log when necessary, or at least use print
  - If you produce maps and graphs never forget legends and be consistent in the choice of units (and colours).

Group project

- 4-5 people $\rightarrow$ one single grade
- Output: a poster
- Should include:
  - at least 2 data sources, among which at least 1 GIS
  - 2 maps
  - brief GIS data source description
  - descriptive statistics section, including correlations (or basic regressions)
  - research proposal
  - $\simeq$ 600 words
- Attach the code if any, or a written paragraph describing the GIS treatment (separate document)

Tutorial

- QGis - click and button: Notes and Data
- QGis - python programming: Notes and Data
- Your own project