Lecture 01: Spatial Data

Theory and Tools (a.k.a. GIS Tools Lab.)

Bruno Conte

18/Sep/2023

Spatial data in economics: this course

- Introduce students to conceptual and practical aspects of **spatial data**
  - What is spatial (geographical) data?
  - How is it used in **research in economics**?
  - Which tools (i.e. computer systems/languages) do we need to work with it?
- Main goal: concepts + tools = practice with real-world data
  - Concepts: types and formats of spatial data
  - Tools: programming in `R` and `RStudio`
- Course's **main philosophy**: a course by an economist working with spatial data
  - Rather than a course by a spatial data's specialist!

Spatial data in economics: this course

This course is about how we, (in principle) **economists**, can use spatial data to empirically answer reserch questions of our interest.

You will learn

- What is spatial data and its applications in economic research
- Basic `R` programming
- Most common spatial data operations
- Introductory (spatial) data visualization

You will not learn

- All state-of-art GIS tools available in `R`
- To write an efficient `R` code*
- To handle big data*
- To solve every possible problem

[*] This is up to you.

Spatial data in economics: this course

Good references

1. Donaldson, D. and Storeygard, A., 2016. The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, *30(4)*, pp.171-98.

2. Lovelace, R., Nowosad, J. and Muenchow, J., 2019. Geocomputation with R. Chapman and Hall/CRC.

3. Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009

4. Wickham, H. and Grolemund, G., 2016. R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.".

Spatial data in economics: schedule

1. Introduction to (spatial) data and programming in R [18.Sep.2023]
    - Introduction to spatial data and examples in economics
    - Basic R programming: set up and practice
2. Spatial data basics: vector data + assignment [21.Sep.2023]
3. Basic operations with vector data + assignment [25.Sep.2023]
4. Geometry operations and miscelanea + follow-up [28.Sep.2023]
5. Raster data and operations + assignment [02.Oct.2023]

6. Take-home exam [03.Nov.2023]

Spatial data in economics: evaluation

1. Class participation (10%)
2. Practical assignments (3 x 10%, in teams)
3. Take-home exam (60%, `pdf` by email):
    - Research idea: spatial data + economics = **research question**
    - Replication of tasks: data + tools = empirical motivation
    - Make sure that you register to it (on almaesami)!

- Any questions?

Getting started: what is Spatial Data?

Europe in 1789 (before the French Revolution)

Europe in 1812 (before the French Invasion of Russia)

Europe in 1815 (after the Congress of Vienna)

Satellite picture of fires (and deforestation) in the Brazilian Amazon

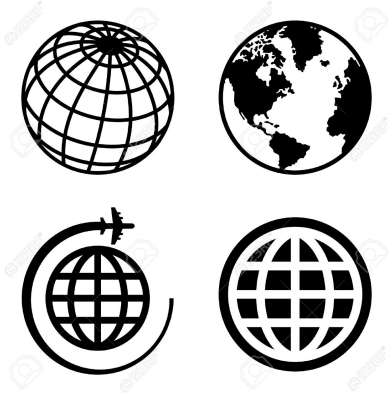

Deforestation in the Brazilian Amazon
N
VENEZUELA
GUYANA
SURINAME
FRENCH GUIANA
COLOMBIA
ECUADOR
PERU
BRAZIL
BOLIVIA
A
B
C
D
E
F
G
H
Amazon
Primary Forest Loss 2020
Very High
High
Medium
Amazon Boundary (biogeographical)
km
0 150 300 600
Amazon Conservation


Urban slums in India

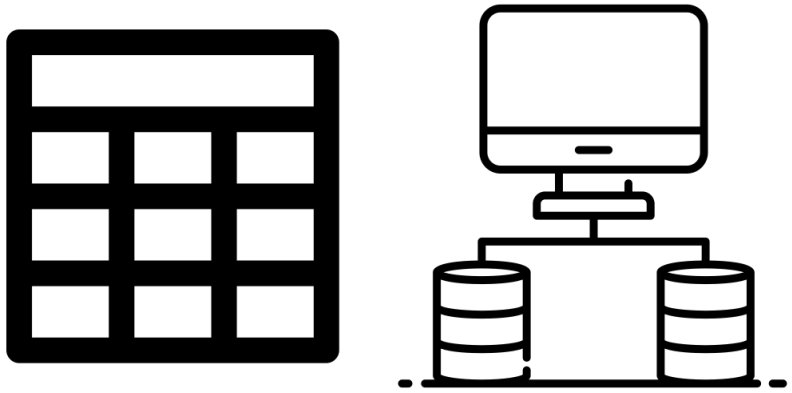
What is Spatial Data?

- Data/information that has a geographical attribute
  - **Much more** than coordinates on a standard dataset
  - Polygons, areas, distances, height, overlaying, intersections, ...

- Common aspect: **unstructured data** (i.e. unconventional data format)



- Our goal: manipulate it into the **structure** required by research



What is GIS?

- GIS = Geographic Information Systems
  - (old) Systems used to manipulate/process spatial data (**1980's**)
  - 1990's: rise of user-friendly, **desktop softwares** (ArcGIS, QGIS)
  - Data Science revolution: full integration of GIS tools into data-processing pipelines; i.e. computer routines that process (potentially spatial) data in modern languages (e.g. `R`)
- **Examples:**
  - Firm processing purchases across branches
    - Is revenue larger in branches *closer to public transportation?*
  - HR firm allocating seasonal workers across plants
    - Choose workers based on residence (reduce commuting time)?

How is Spatial Data used in Economics?

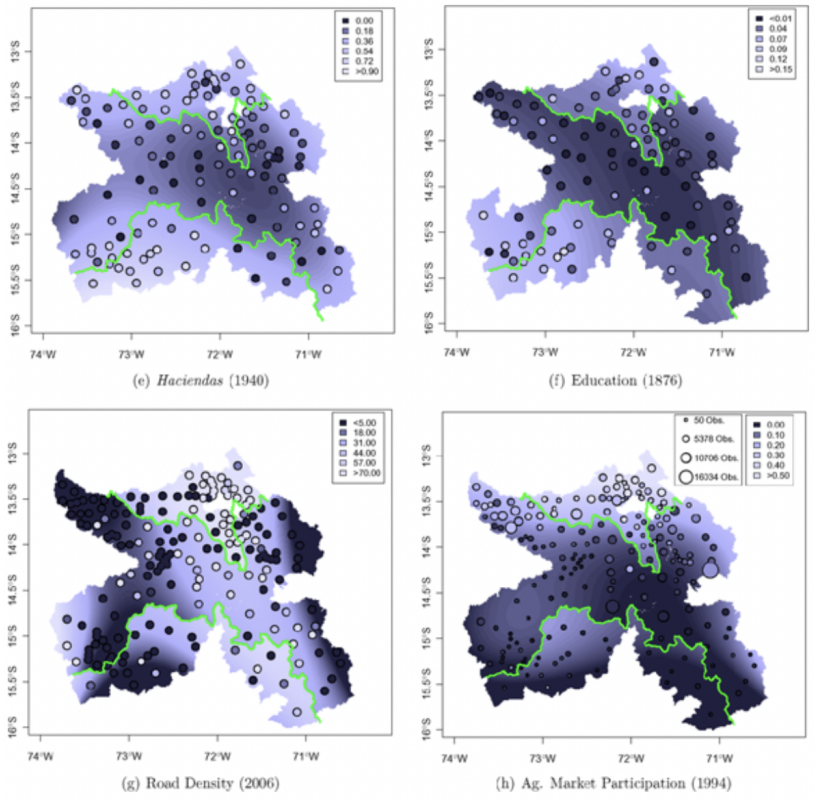
Spatial Data in Economics

- **Motivation:** research questions that requires structuring spatial data.
  - Spatial data = unstructured
  - GIS tools: manipulating spatial data to the required structure
- Applications in economic research:
  - Cholera in London (Snow, 1856)
  - Colonial institutions and development in Peru (Dell, 2010)
  - Railroads and welfare in India (Donaldson, 2018)
  - Climate change and urbanization in Africa (Henderson et al., 2017)

Application 01: John Snow's Cholera Maps in Soho (London)

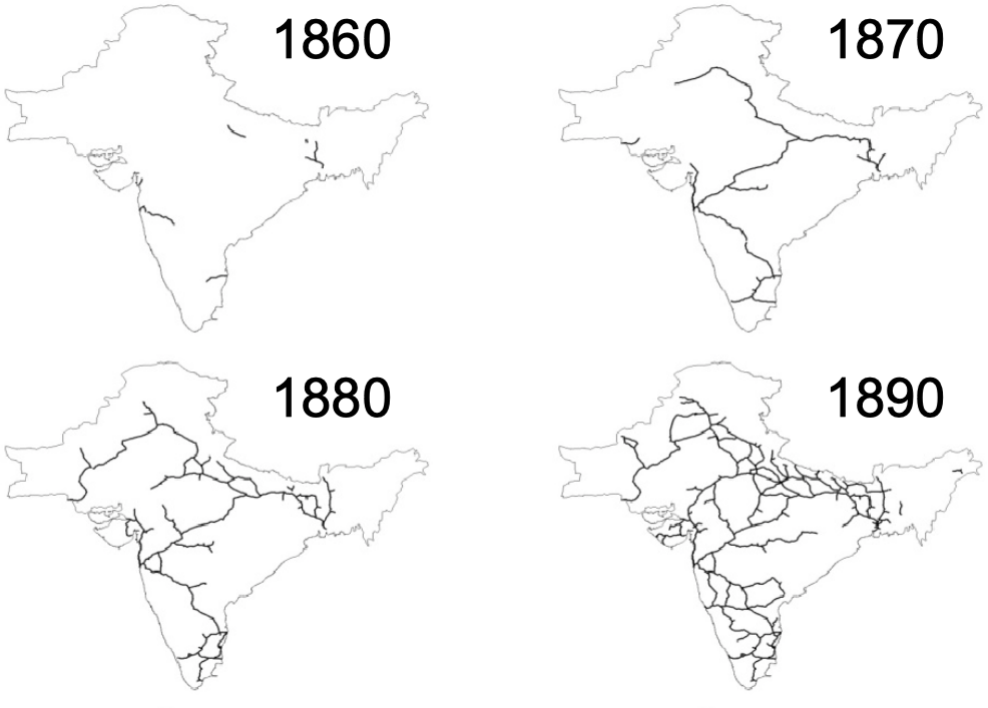
- **Cholera outbreak** in mid 19th century
- Former theory: transmission by air
- John Snow's hypothesis: germ-contaminated water
  - Different rates between locations with different water suppliers
  - Higher rates for those supplied by (polluted) Thames River
- Snow's finding: revolution on public sanitation

Application 02: Long-term consequences of the Mita (colonial) system in Peru



- Spanish empire required **forced labor** to work on silver mines (Potosí)
- Workers from high lands (Mita regions): resistent to the harsh mine conditions
- Mita boundaries: regions that provided more/less conscripts (discontinuously!)
- **Dell's findings:** long-lasting development differences
    - Economic channels: land ownership inequality, less public services, ...

Application 03: Transportation integration and welfare in India



- Vast **expansion of railroad network** in British colonial India
- Standard trade theory: welfare gains from market integration
  - Lack of evidence within countries
- **Donaldson's findings:** improved trade conditions increased welfare
  - Integrated remote areas (reduced price gaps, more trade flows)
  - Welfare gains (real income) from intraregional trade

How to work with spatial data in R?

Working with data (including spatial) in R

- What is `R`?
  - Computer language for statistical computing and graphics
  - Open source, free access
  - Developers' community (CRAN)
    - Development of **libraries** (packages) for specific applications
- `RStudio`: integrated development environment (IDE)
  - User-friendlier environment to work with `R`

R Basics

Basics of programming and data work in R

Open `01_class01.R` on your own computer, where we will cover the following topics. The subsequent slides here are for reference only.

- **Concepts covered:**
  1. `R` basics: environment, main elements (vectors, lists, `data.frame`), libraries
  2. Basic data wrangling with `dplyr`
     - Filtering, mutating, merging
  3. Data visualization with `ggplot2`

- **Setting up `R` (or in `RStudio`)**:

```
# # Install packages (only first time)
# install.packages('data.table')
# install.packages('tidyverse')

# Load them:
library(data.table)
library(dplyr)
```

Note: warning messages are OK!

Basics of programming and data work in R (1/3)

- `R` is versatile working environment
  - Can handle **different** elements (e.g. datasets, images, texts) contempotaneously
- Setting the **local** environment: working directory

```
getwd() # tells you the current wd
```

```
## [1] "/Users/brunoconteleite/Dropbox/Teaching/02-gis-unibo"
```

- Types of elements in `R` environment:
  - Vectors, `data.frame()`, `list()`, among (many) others
  - To check (or clean) current environment: `ls()` (or `rm()`)

Basics of programming and data work in R (2/3)

- **Data wrangling:** manipulating raw data with `dplyr`
  - Creating new variables, filtering datasets, arranging, merging, reshaping
- Pipe syntax: uses `%>%` operator. Example if merging datasets:

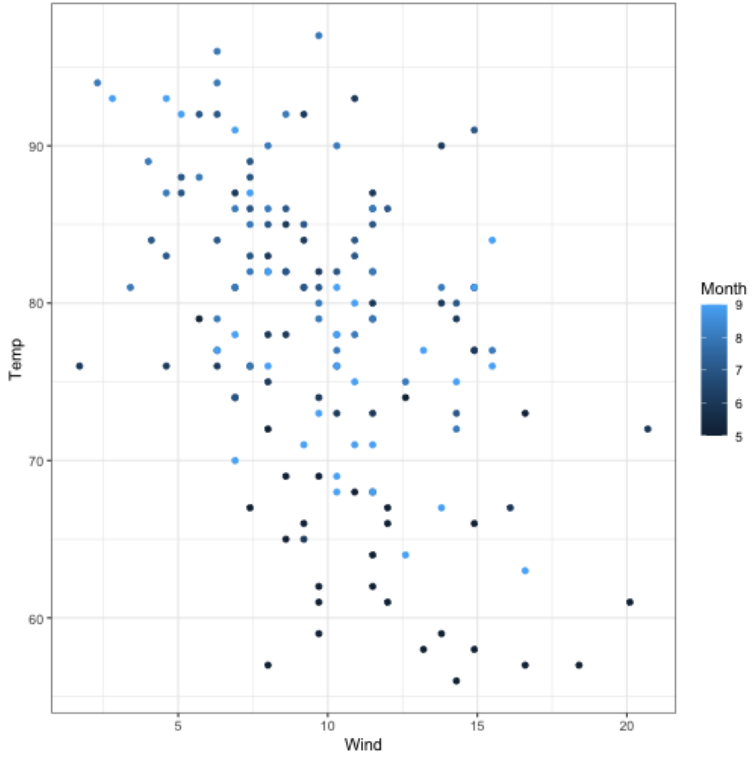
```
df <- merge.data.table(a,b,by = 'Month') # is equivalent to:
df <- a %>%
  left_join(b,by = 'Month')
```

- Same reasoning with many other `dplyr` data-wrangling functions; e.g. `mutate()`, `filter()`, `select()`, `summarise()`, `arrange()`
- Check wiki here

Basics of programming and data work in R (3/3)

- **Data visualization** in `R` with `ggplot()`. Syntax that maps data → geometry → visuals

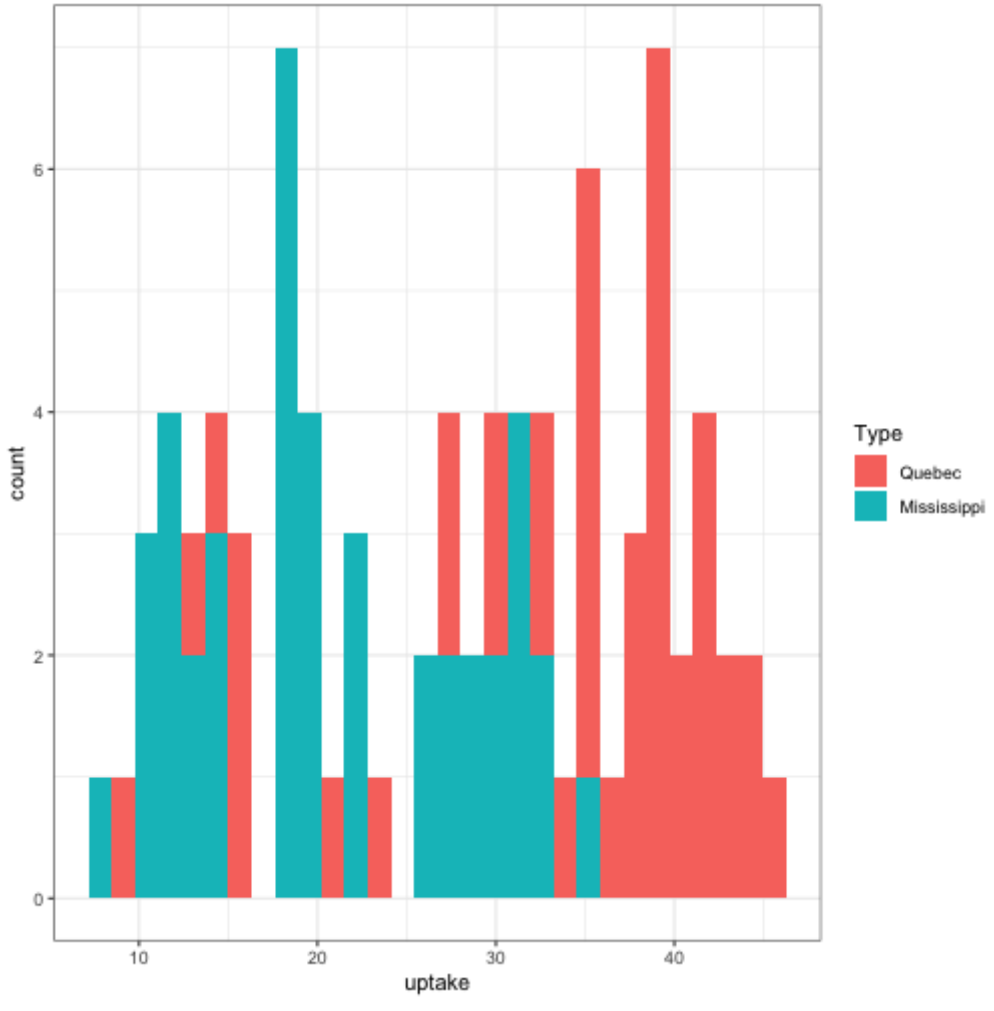
```
library(ggplot2)
try( # ignore this
p <- ggplot(data = data) +
  geom_GEOM(mapping = aes(MAPPINGS)) +
  THEME()
)
# Example
p <- ggplot(data = airquality) +
  geom_point(mapping = aes(Wind,Temp, c
  theme_bw()
```



- Check wiki here

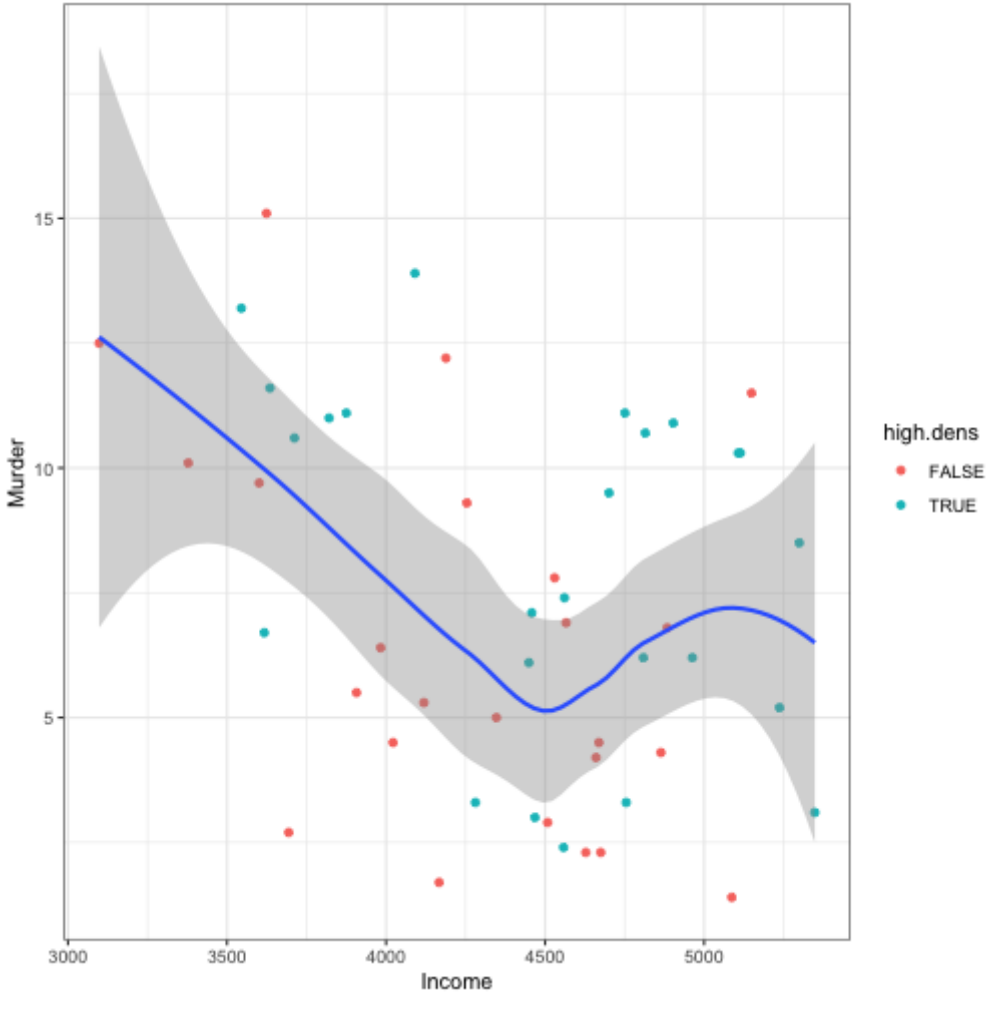
Hands-in: your turn! (1/2)

- Distribution (histogram) of CO2 uptake across plants in US/Canada
- Distinguish plants by state (Quebec/Mississipi)
- Extra: play with different `theme()` parameters of `ggplot()`
- Use the `datasets::CO2` data!



Hands-in: your turn! (2/2)

- Icome vs. Murder rates across US states (scatter plot). Use `state.x77` dataset
- Distinguish between high/low density states
    - High density = (Population/Area) > median: use `mutate()`
- Extra: additional geom layer with non-linear relationship? Use `geom_smooth()`
- Can you remove outliers (i.e. states with Income higher than 6,000)? Use `filter()`



References

- Dell, M., 2010. The persistent effects of Peru's mining mita. *Econometrica*, *78(6)*, pp.1863-1903.
- Donaldson, D., 2018. Railroads of the Raj: Estimating the impact of transportation infrastructure. *American Economic Review*, *108(4-5)*, pp.899-934.
- Henderson, J.V., Storeygard, A. and Deichmann, U., 2017. Has climate change driven urbanization in Africa?. *Journal of development economics*, *124*, pp.60-82.
- Snow, J., 1856. On the mode of communication of cholera. *Edinburgh medical journal*, *1(7)*, p.668.