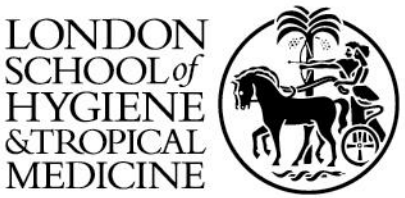
Introduction to Spatial Data in R

Emily Nightingale & Ruoran Li



In the course so far

- Basics of R
- Data manipulation in R
  - *tidyverse* set of packages
  - *dplyr* syntax (select, mutate, filter, group by, summarise)

Yesterday's task: Working with data from The Humanitarian Data Exchange (HDX) to investigate population coverage of each health facility.

The next two sessions

- Types of spatial data
- Coordinate reference systems
- Working with spatial data in R
  - Reading/writing
  - Simple manipulations
- Plotting spatial data
- Spatial data operations

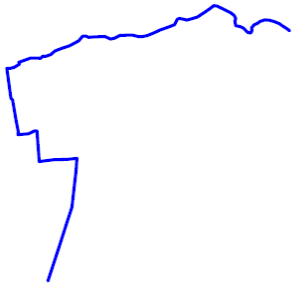
A great (free) resource: https://r.geocompx.org/

Today

13:00-14:00	Solutions to session 2 exercises
14:00-14:15	Break
14:15-15:15	Intro to spatial data
15:15-15:30	Break
15:30-16:00	Start on exercises and questions

Types of Spatial Data

Types of Spatial Data - Vector



Points

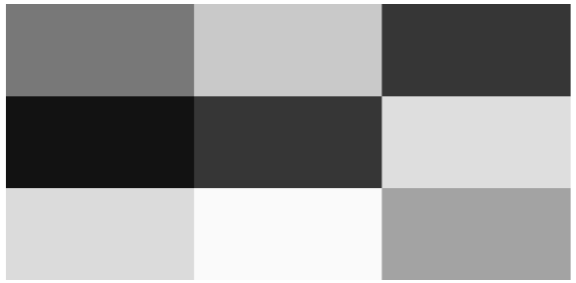
Lines

Polygons

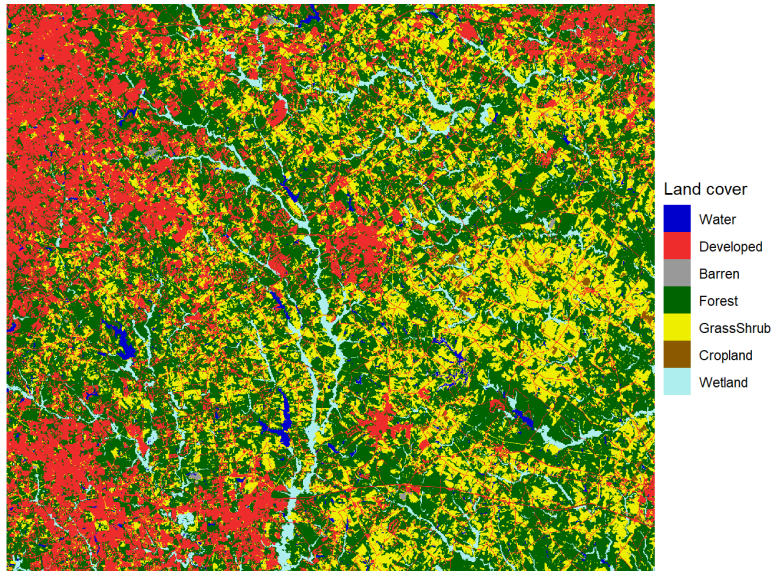
Types of Spatial Data – Raster

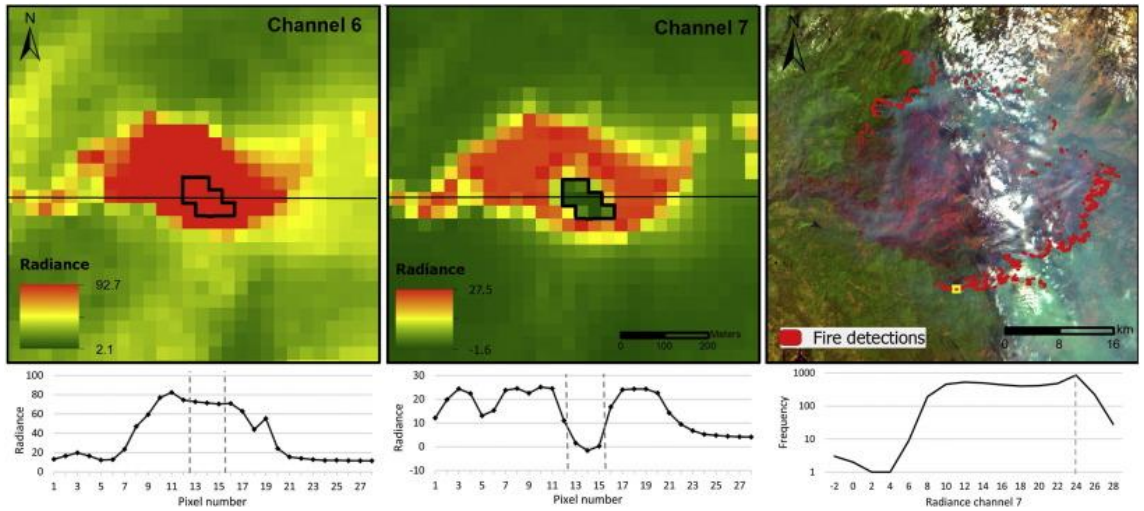
53	21	79
93	79	13
14	2	36

Raster values



Raster image





Types of Spatial Data

A range of file types may be used to store different types of spatial data:

- **Vector data:** shapefile, geoJSON, csv
- **Raster data:** geoTIFF, image file types e.g. jpeg

In this course we will be using vector data stored in the form of **shapefiles** and **csv**.

You may also have your own spatial data – can you identify what type they are?

Aside: Sourcing spatial data

Type	Name	Description	Link
Basemaps	Natural Earth	Free to download cultural, physical, and raster basemap data.	http://www.naturalearthdata.com/downloads
	USGS Earth Explorer	Free satellite and aerial imagery.	https://earthexplorer.usgs.gov/
Boundaries	Global Administrative Areas Database (GADM)	Allows you to download shapefiles (or GeoJSON and KMZ) of levels of administrative area per country or for the world. Freely available for academic or non-commercial use.	https://gadm.org/index.html
	GeoBoundaries	Individual country, global composite and simplified (for visual) admin boundaries	https://www.geoboundaries.org/
Features	OpenStreetMap	Crowd-sourced High spatial resolution cultural vector data such as buildings, roads, and waterways.	https://gisgeography.com/openstreetmap-download-osm-data/
	Humanitarian Data Exchange (UN OCHA HDX)	Humanitarian data available for specific countries or globally, including indicators, location of health facilities. Some data is spatial (shapefiles), some is in csv format which could be appended to different administrative regions.	https://data.humdata.org/dataset
	WorldPop	Geospatial data on population distribution, demographics and dynamics focusing on lower and middle income countries.	https://hub.worldpop.org/
Gazetteers	GeoNames	When you only have a place name or postcode, you can use this gazetteer to search and identify the coordinates.	https://www.geonames.org/

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

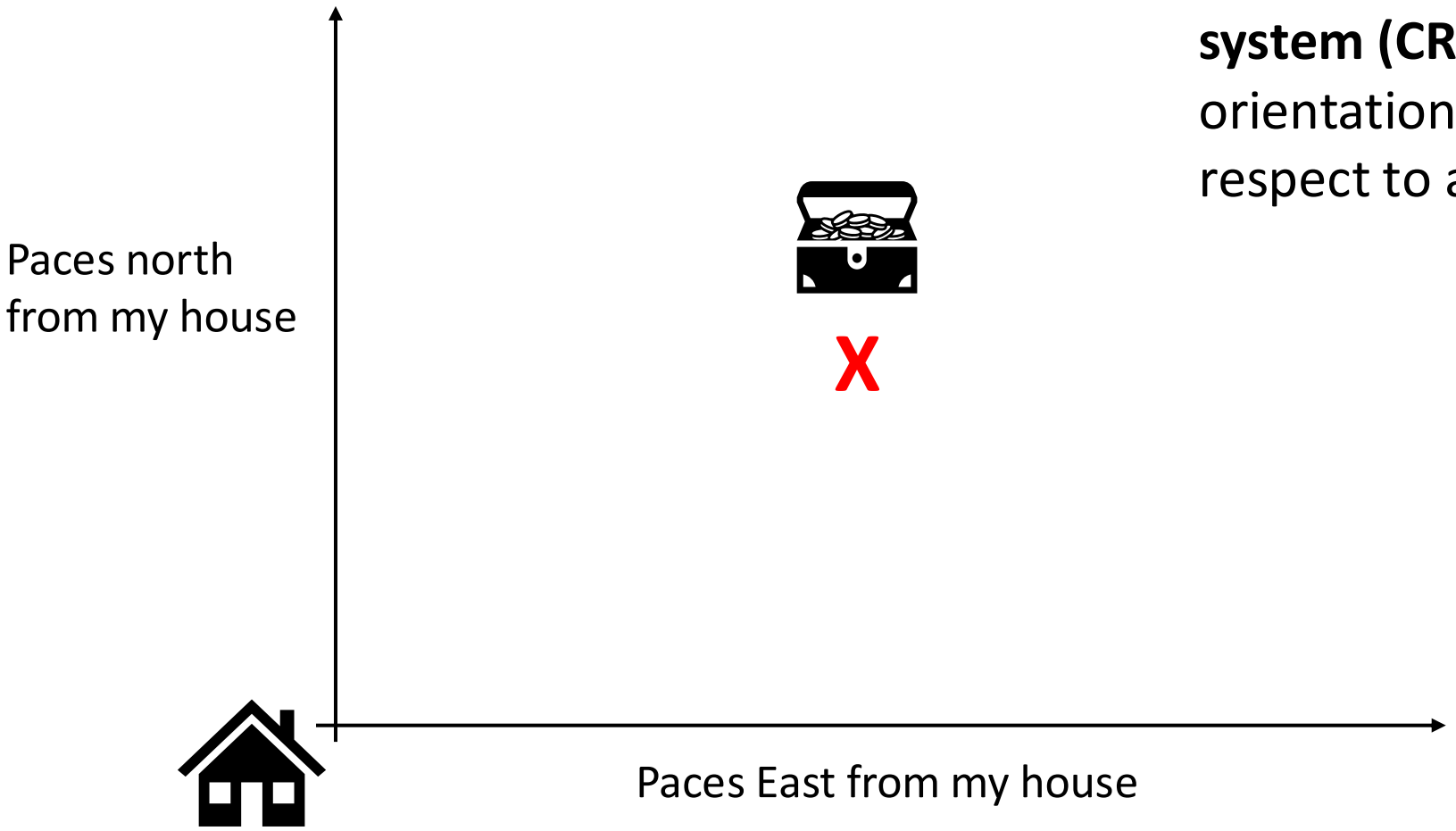
Orienting ourselves:
Coordinate Reference Systems

Coordinate Reference Systems

The treasure is buried at (24,18)...

 ?

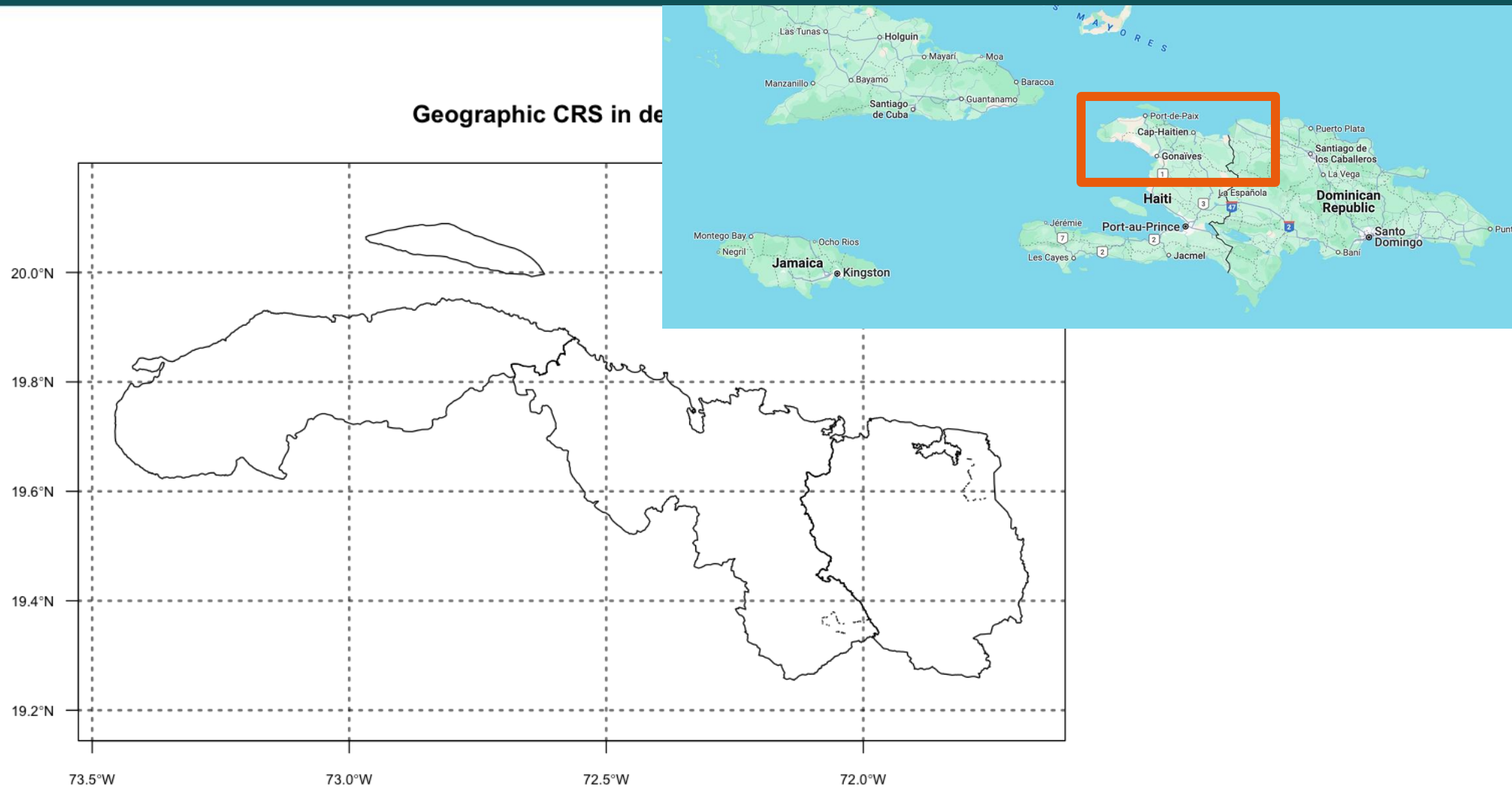
Coordinate Reference Systems



The **coordinate reference system (CRS)** tells us the orientation of the data with respect to a reference point.

Coordinate Reference Systems

Coordinate Reference Systems



A quick reminder that we are not flat-Earthers…

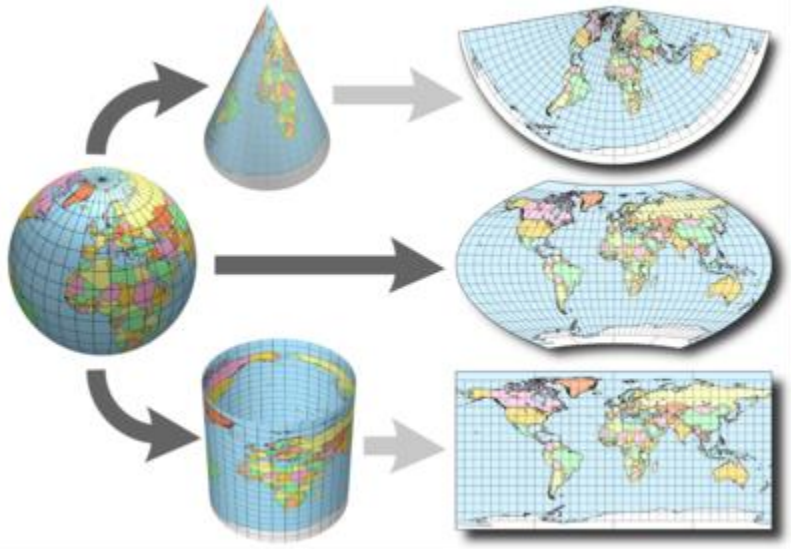
Types of Coordinate Reference Systems

Two types of CRS:

Geographic (Unprojected)

Defined in degrees around Earth's surface

1 degree longitude reflect inconsistent distances



Projected

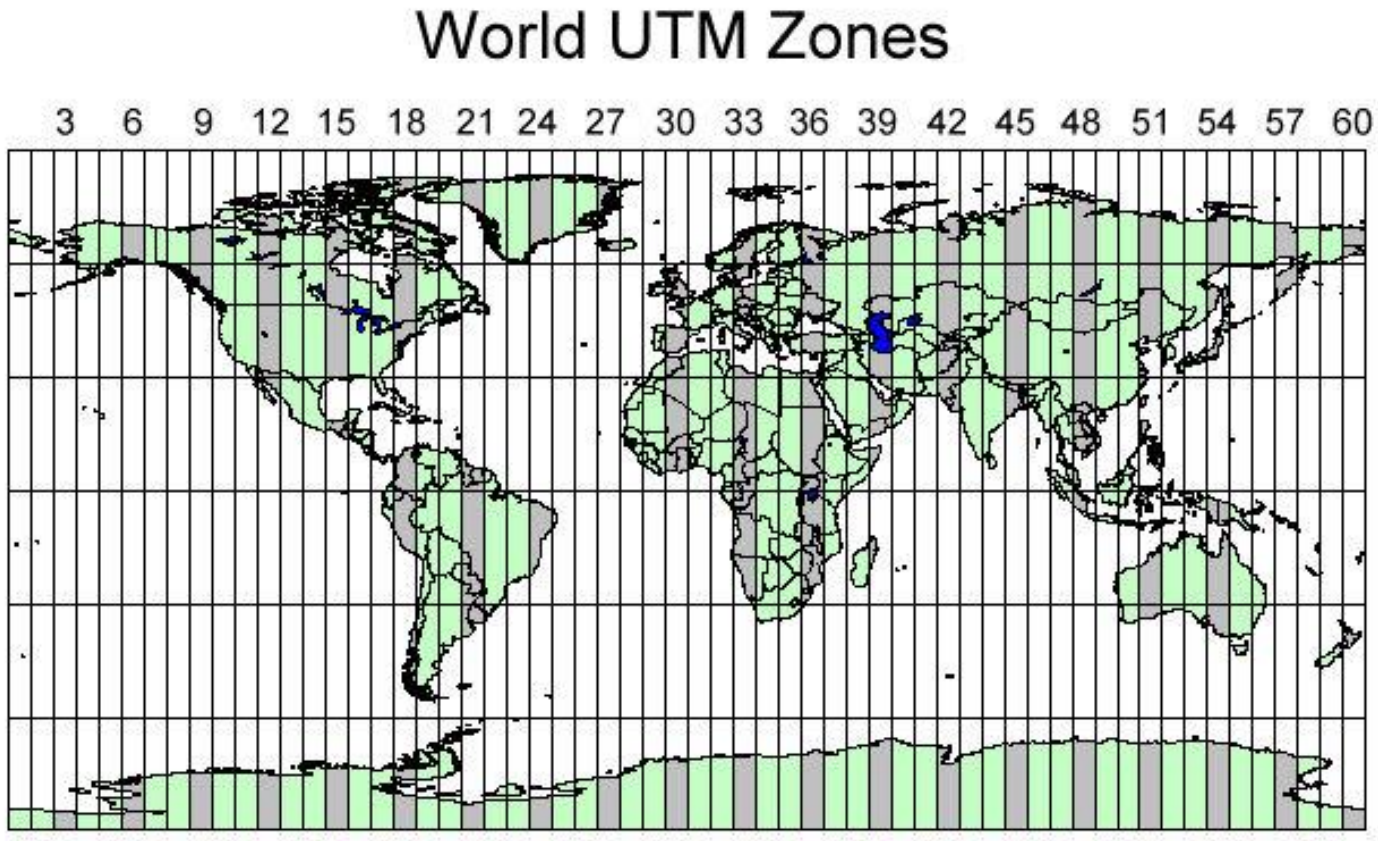
2D transformation of 3D surface

Defined in linear units e.g. km

Accurate measurement of distance/area/angle

Types of Coordinate Reference Systems

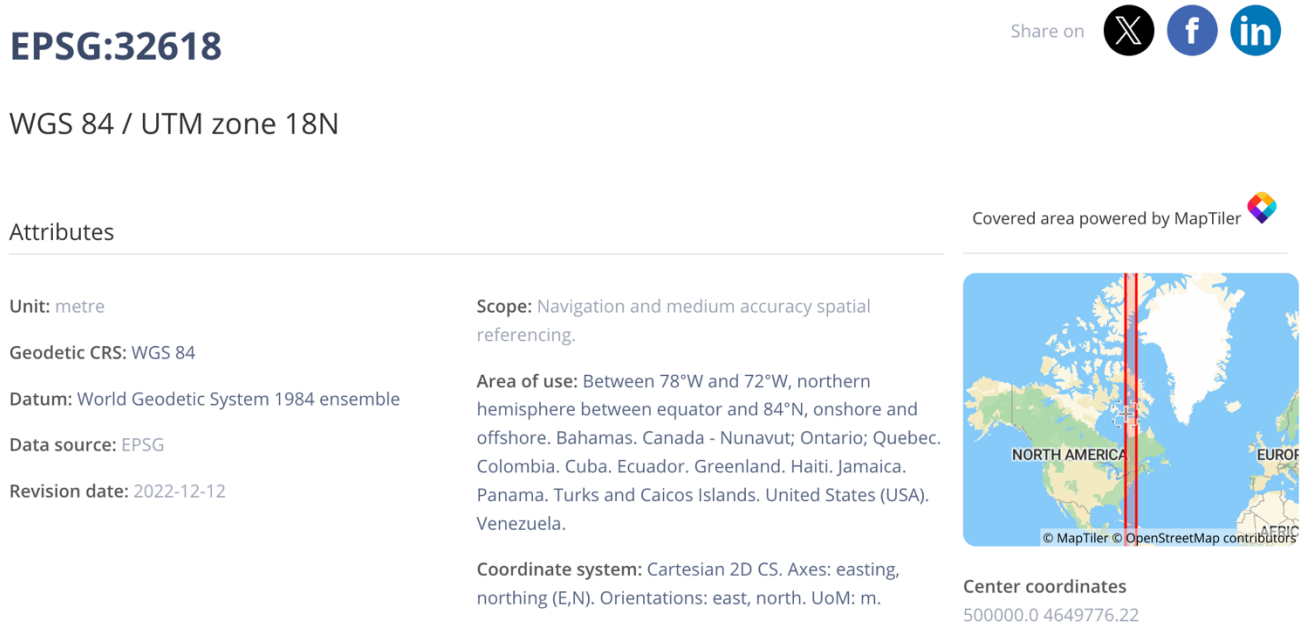
A very commonly used projection is the **Universal Transverse Mercator (UTM)**



Defining a CRS

Each CRS is defined by a unique ***EPSG code*** (or "*proj4string*").

- The most common geographic CRS is WGS84, EPSG code **4326**.
- Projected CRS codes can be looked up here: https://epsg.io/



If using the wrong CRS, the data may be missing or skewed when plotted over the region of interest.

Key takeaways on CRSs

- We must know the CRS to work with spatial data
- For work on a large/global scale, use a **geographic** CRS
  - WGS84 (i.e. standard latitude/longitude) has EPSG code 4326
- When calculating distances or areas, use a **projected** CRS
  - UTM zones for every country
  - Some countries have their own projected CRS

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Working with spatial data in R

Working with spatial data in R

The ***sf*** ("simple features") package allows us to easily work with spatial data in R.

- Integrates with the ***tidyverse***
- Handle multiple types of vector data (points/lines/polygons)
- Quickly create maps with ***ggplot***, ***tmap*** and ***mapview***
- Lots of functionality with spatial operations (calculating distances, drawing buffers, finding overlap, etc.)

Working with spatial data in R

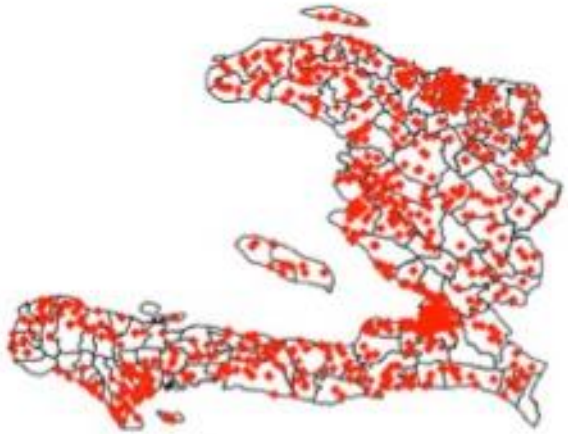
First steps:

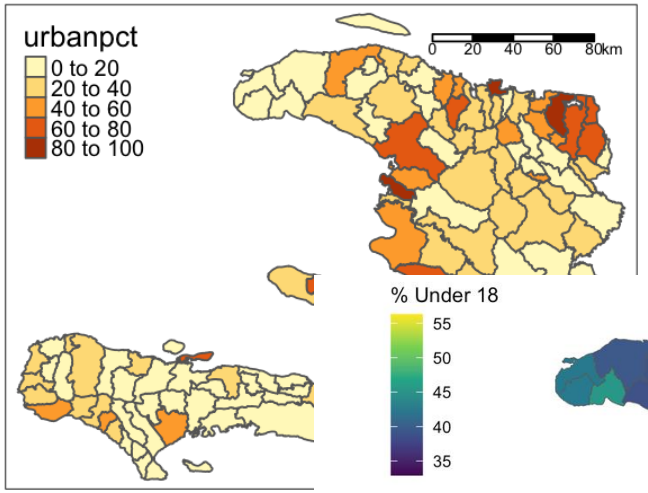
1. Read in data from
    a) shapefile
    b) csv
2. Check (and, if necessary, set) the CRS
3. Reproject the data for plotting/analysis

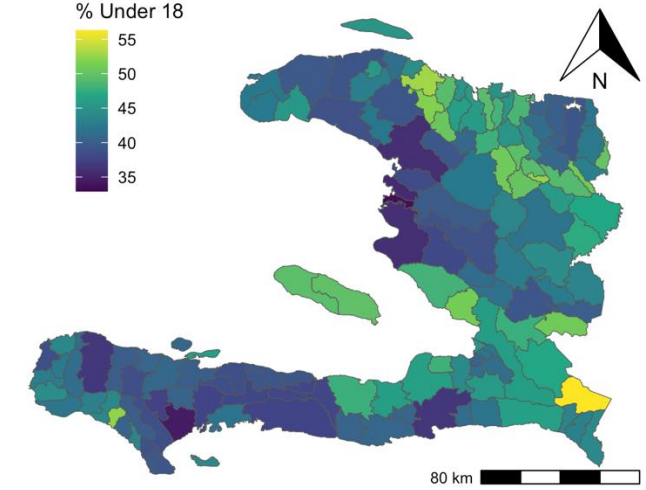
Example:
Mapping health centres in Haiti

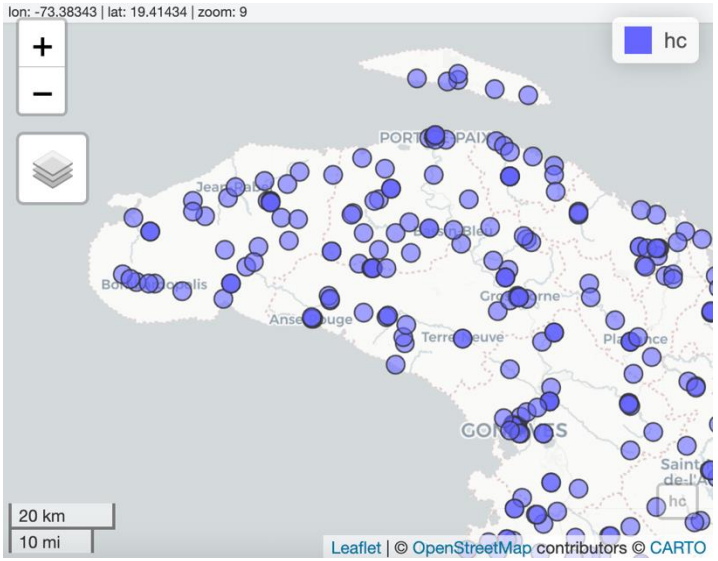
Plotting spatial data in R

- Quick and dirty, checking data: ***base R***
- Presentation-ready: ***ggplot2*** and ***tmap***
- Interactive: ***mapview*** *(also now possible with tmap)*









Exercises

Share your best map with your classmates:
https://padlet.com/emilynightingale2/isair__2025__session3

Summary

Today we've covered:

- Types of spatial data
- Coordinate reference systems and projections
- How to read/write spatial data using the ***sf*** package in R
- Plotting using ***base R***, ***ggplot***, *and* ***tmap*** packages
- Creating interactive maps using ***mapview***

Tomorrow:

- Spatial data operations
- Practical task

Spatial data operations
and bringing it all together

Spatial data operations

The *sf* package has many functions to manipulate spatial data.

These generally start with "*st_*"

- *Make use of autocomplete and help files in Rstudio to explore the available functions*

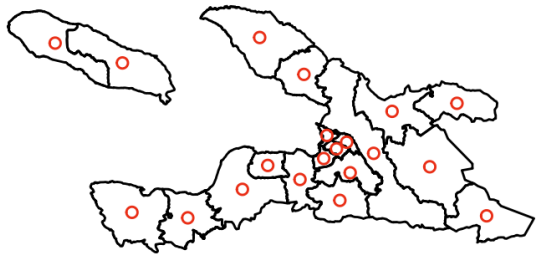
As noted yesterday, your data should be in a projected CRS to apply any operation relating to distances.

Spatial data operations

West Department

Buffer

Centroids



Union

Intersection

Difference

Practical task

Identify and visualise the areas that are further than 5km from a health service.

Share your final maps!
https://padlet.com/emilynightingale2/isair_2025_session4

Preparation for next sessions
Recap of linear regression

Recap: Linear regression modelling

Statistical models allow us to:

- Describe trends
- Test hypotheses
- Predict from data

All models are collections of assumptions.

> *"Essentially, all models are wrong but some models are useful."*
> *George E. Box*

Recap: Linear regression modelling

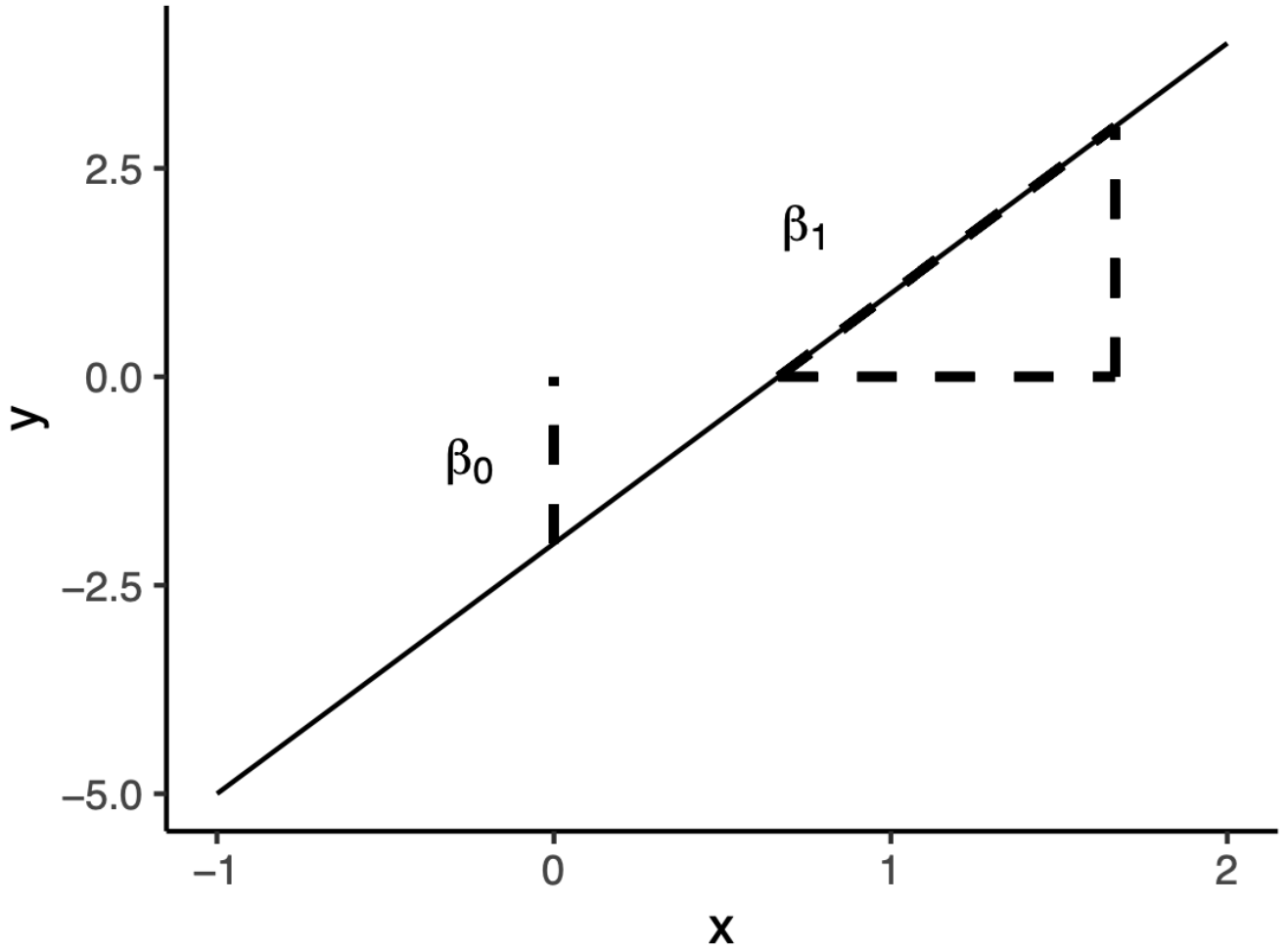
A basic models is the linear model with one predictor variable, i.e.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where

- y are the observations of the outcome/response/dependent variable
- x are the observations of the predictor/covariate/independent variable
- β_0, β_1 are the coefficients of the model
- ϵ are the assumed *errors*, which follow a normal distribution $\epsilon_i \sim N(0, \sigma^2)$

Recap: Linear regression modelling



Recap: Linear regression modelling

We can extend to a linear model with multiple predictors, e.g. for m predictors $x_1, x_2, \ldots, x_m$,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \ \ldots + \beta_m x_m + \epsilon_i$$

We can also write this in matrix notation,

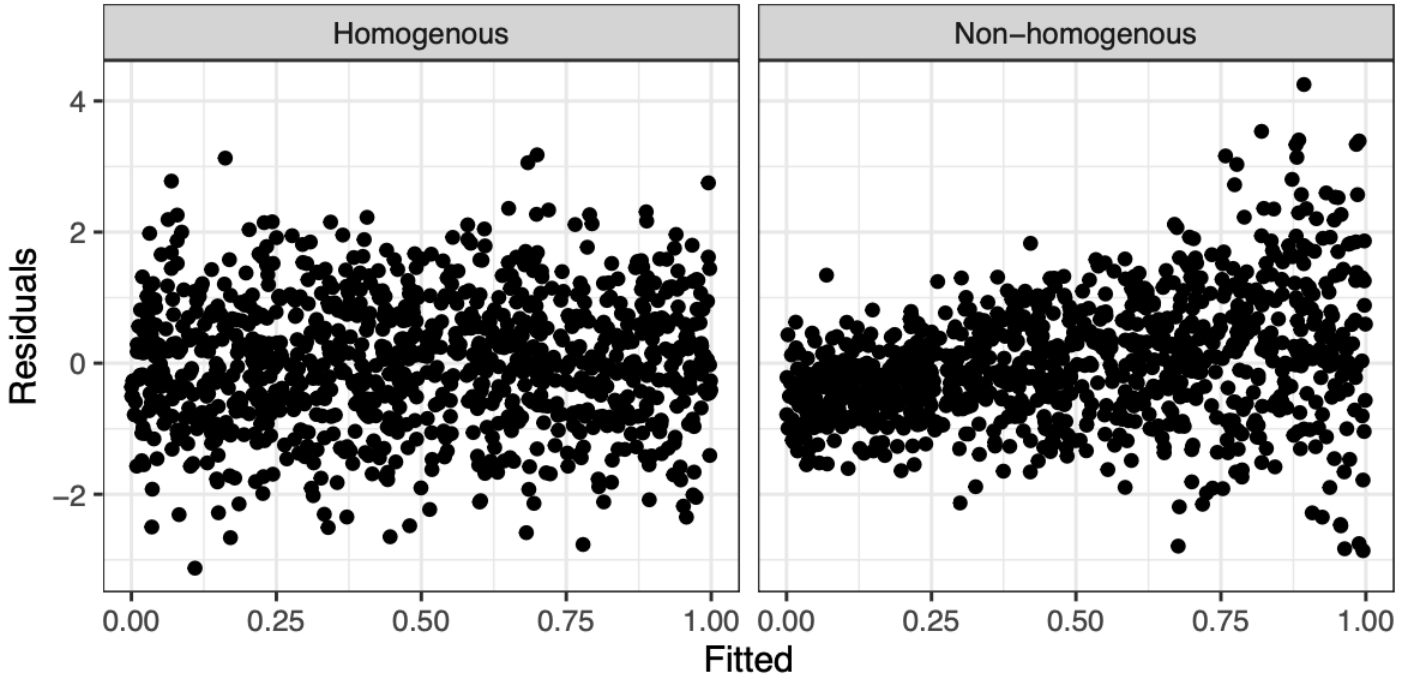
$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$\epsilon_i \sim N(0, \sigma^2)$$

where $X = [x_1|x_2|\ldots|x_m]$.

Recap: Linear regression modelling

Assumptions of a linear model:

1. **Linearity** Every predictor x_k has a linear effect on $\mathbb{E}(y)$
2. **Normality of errors** Any unexplained variability follows a normal distribution
3. **Homogeneity of errors** All errors are described by the *same* normal distribution



Recap: Linear regression modelling

If this isn’t feeling familiar or you’re a bit rusty, you may wish to review the basics of linear models before tomorrow’s session.