IMT2200

INTRODUCCIÓN A CIENCIA DE DATOS

2022-2

TIPOS Y FORMATOS DE DATOS

CLASE 4

TIPOS DE DATOS

Estructurados

- Esquema predefinido y homogéneo.
- Estructura tabular con filas y columnas.
- Fácil de analizar y modelar.

Semi-estructurados

- No están organizados en filas y columnas.
- Cuentan con llaves y etiquetas que proporcionan una estructura jerárquica a los datos.
- Análisis requiere más trabajo previo.

No estructurados

- No hay un estructura o jerarquía interna clara.
- Existen muchos formatos nativos no estructurados.
- Más difícil de analizar.

```
Year,Make,Model,Price
1997,Ford,E350,3000.00
1999,Chevy,"Venture Extended Edition",4900.00
1999,Chevy,"Venture Extended Edition",5000.00
1996,Jeep,Grand Cherokee,4799.00
```


```
[
  {
      "age_adjusted_death_rate": "7.6",
      "death_rate": "6.2",
      "deaths": "32",
      "leading_cause": "Accidents Except Drug Posioning (V01-X39, X43, X45-X59, Y85-Y86)",
      "race_ethnicity": "Asian and Pacific Islander",
      "sex": "F",
      "year": "2007"
  },
  {
      "age_adjusted_death_rate": "8.1",
      "death_rate": "8.3",
      "deaths": "87",
      ...
```


```
Call me Ishmael. Some years ago—never
mind how long precisely—having little or
no money in my purse, and nothing particular
to interest me on shore, I thought ....
```

DATOS ESTRUCTURADOS

- Data altamente organizada y fácil de descifrar.
- Está en un formato predefinido (tabla o base de datos)
- Formatos comunes: *.csv, .tsv, .txt, .xlsx, SQL*
- Dato (datum) ➔ una observación o abstracción de una entidad real (persona, objeto o evento). Puede estar descrita por uno o más atributos.
- Datos (data/dataset) ➔ conjunto homogéneos de datos relativo a una colección de entidades.

```
Year,Make,Model,Price
1997,Ford,E350,3000.00
1999,Chevy,"Venture Extended Edition",4900.00
1999,Chevy,"Venture Extended Edition",5000.00
1996,Jeep,Grand Cherokee,4799.00
```

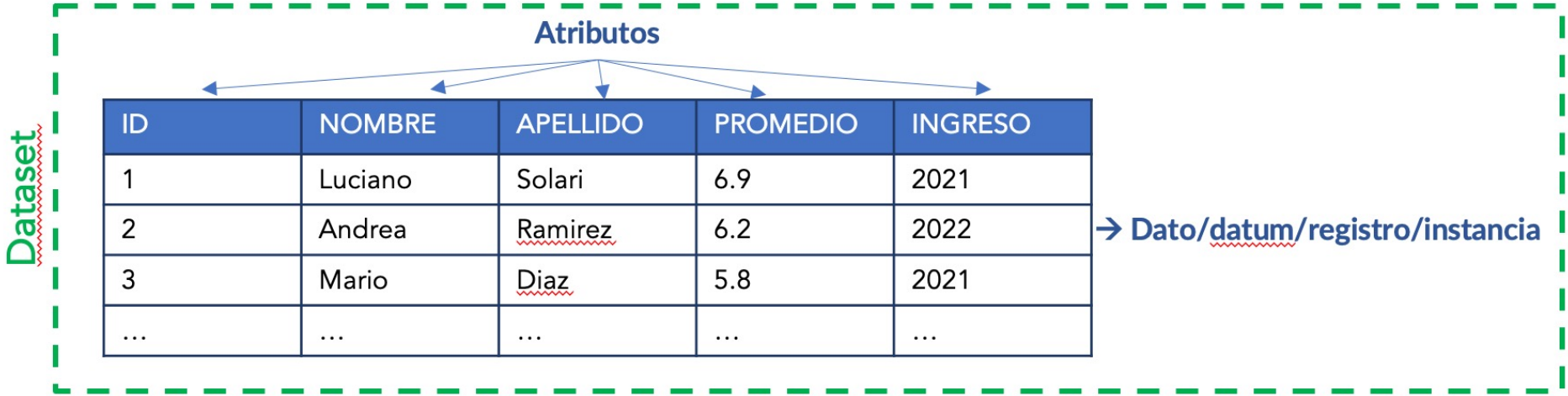
	A	B	C
1	**name**	**age**	**height**
2	Michael	46	5'9"
3	Jim	31	6'0"
4	Pam	29	5'7"
5	Meredith	53	5'6"
6	Dwight	35	5'10"

Name	Dry/Wet Food	Good Boy (Y/N)
Fido	Dry	Y
Rex	Wet	N
Bubbles	Dry	Y
Cujo	Wet	N

Tag #	Height (in)	Weight (lbs)
1573	15	21
2684	9	7
3795	27	130
4806	6	5

Tag #	Name	Breed	Color	Age
1573	Fido	Beagle	Brown/White	1.5
2684	Rex	Pekingese	White	9
3795	Bubbles	Rottweiler	Black	5
4806	Cujo	Chihuahua	Gold	4

DATOS ESTRUCTURADOS



TIPOS DE DATOS

Estructurados

- Esquema predefinido y homogéneo.
- Estructura tabular con filas y columnas.
- Fácil de analizar y modelar.

```
Year,Make,Model,Price
1997,Ford,E350,3000.00
1999,Chevy,"Venture Extended Edition",4900.00
1999,Chevy,"Venture Extended Edition",5000.00
1996,Jeep,Grand Cherokee,4799.00
```

Semi-estructurados

- No están organizados en filas y columnas.
- Cuentan con llaves y etiquetas que proporcionan una estructura jerárquica a los datos.
- Análisis requiere más trabajo previo.


```
[
  {
      "age_adjusted_death_rate": "7.6",
      "death_rate": "6.2",
      "deaths": "32",
      "leading_cause": "Accidents Except Drug Posioning (V01-X39, X43, X45-X59, Y85-Y86)",
      "race_ethnicity": "Asian and Pacific Islander",
      "sex": "F",
      "year": "2007"
  },
  {
      "age_adjusted_death_rate": "8.1",
      "death_rate": "8.3",
      "deaths": "87",
      ...
```


No estructurados

- No hay un estructura o jerarquía interna clara.
- Existen muchos formatos nativos no estructurados.
- Más difícil de analizar.

```
Call me Ishmael. Some years ago—never
mind how long precisely—having little or
no money in my purse, and nothing particular
to interest me on shore, I thought ....
```

Datos Semi-Estructurados

DATOS SEMI-ESTRUCTURADOS

XML: Extensible Markup Language

- Metalenguaje diseñado para almacenar y transportar datos.
- Diseñado para ser auto-descriptivo.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<websites>
  <website id="133">
    <title lang="en">File Extension Database</title>
    <url>https://www.file-extension.info</url>
    <category>Data Formats</category>
  </website>
</websites>
```

JSON: JavaScript Object Notation (https://www.json.org/json-en.html)

- Formato de texto liviano para intercambio de data, fácil de interpretar por humanos (archivo de texto simple) y de generar y formatear (parse) para máquinas.
- Estándar para envío de data mediante entre servidores y aplicaciones web.
- Estructura auto-descriptiva.
- Semejante a un diccionario de Python, con dos estructuras base:
  - keys : strings
  - valores:
    - 4 tipos de datos atómicos: number, string, boolean, null
    - 2 tipos de datos compuestos: array, object

```json
{
  "departamento":8,
  "nombredepto":"Ventas",
  "director": "Juan Rodríguez",
  "empleados":[
    {
      "nombre":"Pedro",
      "apellido":"Fernández"
    },{
      "nombre":"Jacinto",
      "apellido":"Benavente"
    }
  ]
}
```

DATOS SEMI-ESTRUCTURADOS - JSON

- La estructura puede ser anidada: el valor de un atributo es un nuevo diccionario, o conjunto de pares atributo-valor.

- Para grandes conjuntos de datos, permite evitar repeticiones y campos en blanco ➔ formato más liviano

- Existen varias librerías que permiten trabajar con datos en formato json:

  - json: librería con funciones básicas para leer, escribir y analizar datos en formato JSON.

    https://docs.python.org/3/library/json.html

    - json.loads ➔ leer un archivo .json

```
{
  "departamento":8,
  "nombredepto":"Ventas",
  "director": "Juan Rodríguez",
  "empleados":[
    {
      "nombre":"Pedro",
      "apellido":"Fernández"
    },{
      "nombre":"Jacinto",
      "apellido":"Benavente"
    }
  ]
}
```

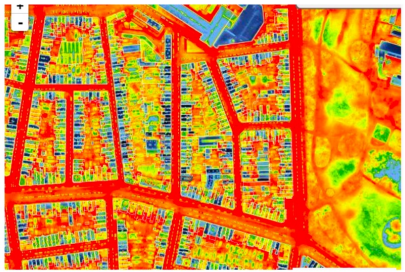
Lectura y Manejo de Datos en Python

Notebook Clase 3

DATOS GEOESPACIALES

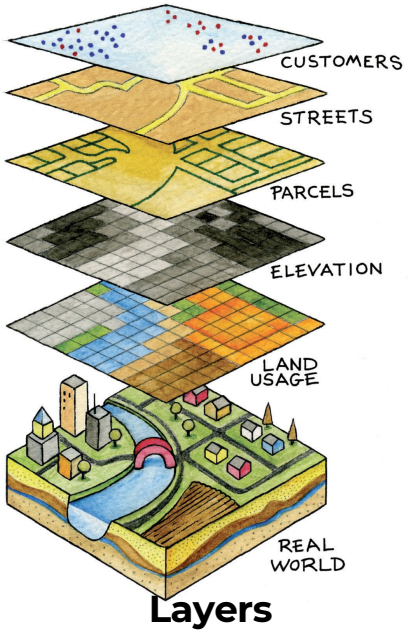
Otro tipo de datos estructurados: **datos geoespaciales**

- La superficie terrestre se representa mediante una **superposición de capas o layers** de información.
- Existen dos tipos de datos geoespaciales digitales, que corresponden a distintas estrategias de muestreo de la superficie.
  - **Raster:** codifican la información como una superficie continua representada por una cuadrícula, como los pixeles de una imagen.
  - **Vectoriales:** representan el mundo como una colección de objetos discretos usando puntos, líneas y polígonos.

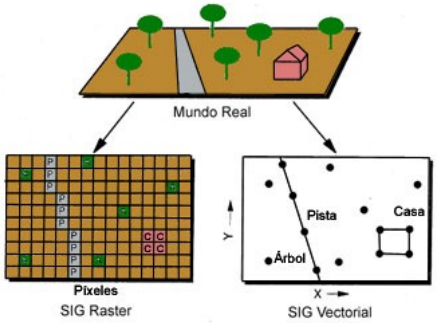


Raster

Vector

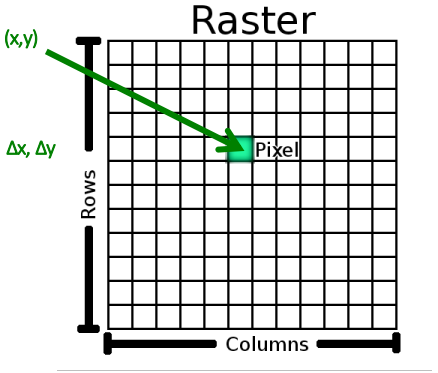


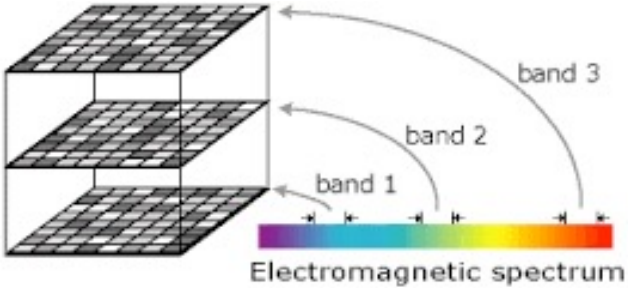
Layers



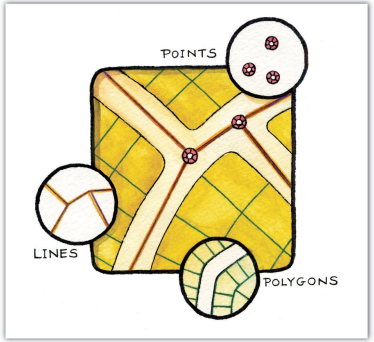
DATOS RASTER

- Se samplean atributos de la superficie en intervalos fijos.
- Cada muestra representa una celda en una cuadrícula o matriz de valores.
- **Georreferenciación:** cada pixel tiene asociadas coordenadas geográficas.
- Resolución espacial: está dada por el tamaño del pixel, que depende del sensor.
- Imágenes raster pueden contener múltiples bandas: RGB, IR, multiespectral.
- **Formatos:** .tiff, geotiff, jpeg, hgt, xyz, asc
  - Lista exhaustiva: `https://gdal.org/drivers/raster/index.html`



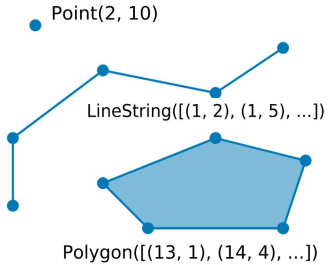


DATOS VECTORIALES



Son apropiados para representar **entidades discretas con bordes bien definidos**, como carreteras, terrenos, límites políticos o administrativos, etc.

- Se almacenan como una serie de geometrías con atributos asociados que describen el elemento.
  - **Atributos** ➔ pueden corresponder a cualquier característica cuantitativa o cualitativa del elemento.
  - **Geometría** ➔ conjuntos de vértices (x,y)
    - Valores de x,y dependen del sistema de coordenadas utilizado.



- Se utilizan 3 tipos de geometrías:
  - **Punto:** única ubicación (x,y)
  - **Línea:** grupo de puntos conectados
  - **Polígono:** línea cerrada que encierra un área

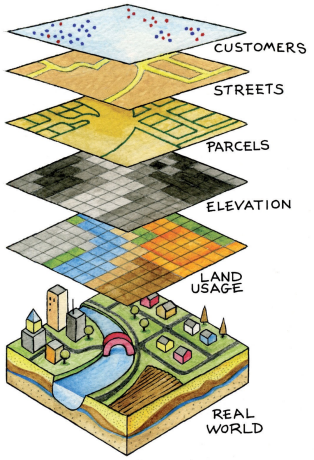
```
[27]: data.head()
```

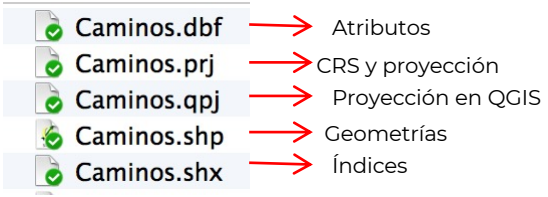
[27]:

	GROUP	CLASS	geometry	area
0	64	32421	POLYGON ((379394.248 6689991.936, 379389.790 6...	76.027392
1	64	32421	POLYGON ((378980.811 6689359.377, 378983.401 6...	2652.054186
2	64	32421	POLYGON ((378804.766 6689256.471, 378817.107 6...	3185.649995
3	64	32421	POLYGON ((379229.695 6685025.111, 379233.366 6...	13075.165279
4	64	32421	POLYGON ((379825.199 6685096.247, 379829.651 6...	3980.682621

DATOS VECTORIALES

- Una capa vectorial contiene sólo un tipo de geometría.
- Existen múltiples tipos de formatos de datos vectoriales:
  - **ESRI Shapefile (.shp):** muy usado, formato multi-archivo
  - **GeoJSON (.gjson)**: formato liviano, basado en JSON. Archivo de texto con colección de diccionarios con geometría y atributos.
  - **GeoPackage (.gpkg):** alternativa más liviana y moderna a shapefile, único archivo.
  - **Keyhole Markup Language (.kml):** lenguaje basado en XML para representar datos geográficos, usado en Google Earth.





FORMAT	SHAPEFILE	GEOJSON	GEOPACKAGE
Age (years)	30	10	5
Compatibility	GIS	GIS, any text editor	GIS, SQL
Relative size	1.00	2.26	1.30
Compression ratio	4.79:1	12.08:1	4.53:1
QGIS performance	Good	Bad	Good
Use case	Old standard	Web, small data sets	New standard

GEOPANDAS

https://geopandas.org/

- Librería que facilita el trabajo con datos vectoriales en Python.

- **Extiende** los tipos de datos usados por pandas para permitir manipulación y operaciones espaciales con datos geométricos.
  - DataFrame➔ GeoDataframe
  - Series➔ GeoSeries

- Se basa en otras librerías pre-existentes:
  - fiona: acceso a formatos de datos geoespaciales
  - shapely : operaciones geométricas

```
[27]: data.head()
```

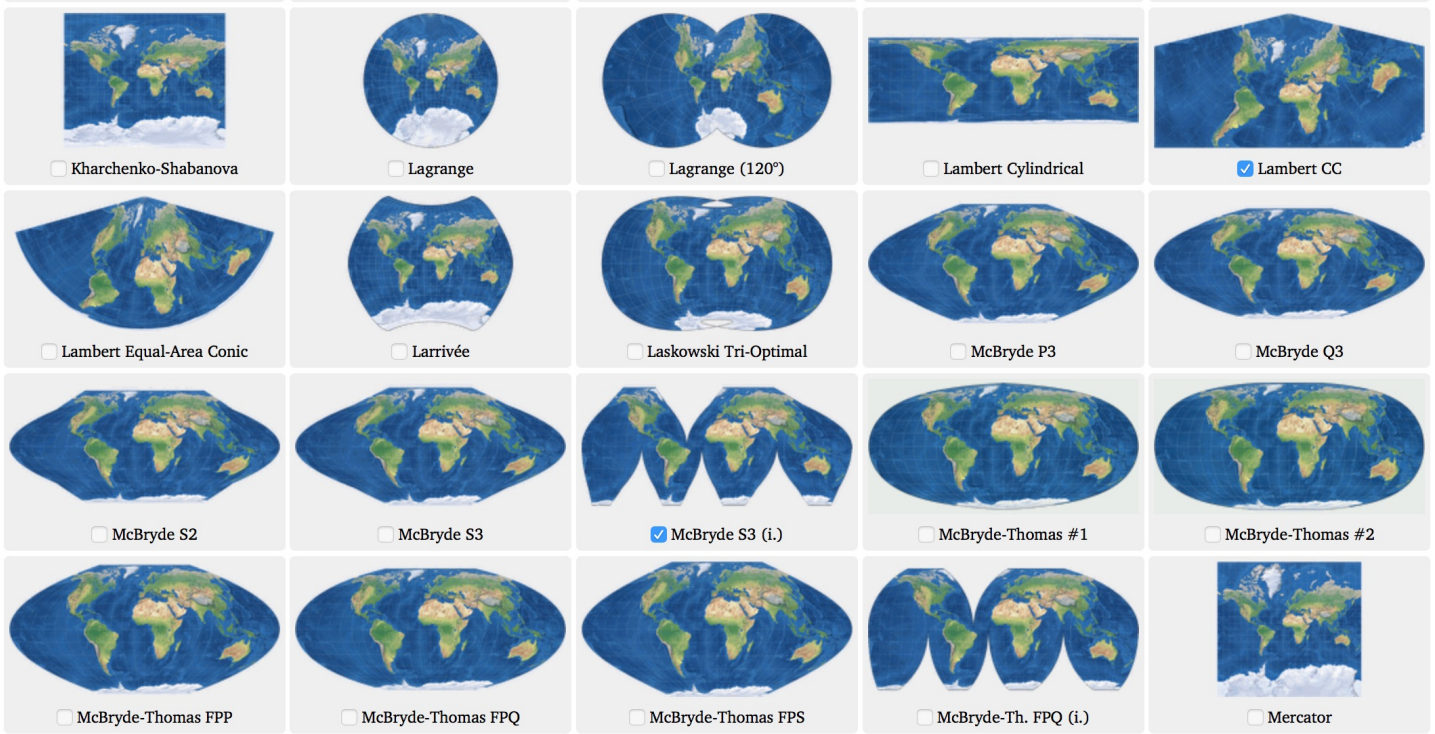
[27]:

	GROUP	CLASS	geometry	area
0	64	32421	POLYGON ((379394.248 6689991.936, 379389.790 6...	76.027392
1	64	32421	POLYGON ((378980.811 6689359.377, 378983.401 6...	2652.054186
2	64	32421	POLYGON ((378804.766 6689256.471, 378817.107 6...	3185.649995
3	64	32421	POLYGON ((379229.695 6685025.111, 379233.366 6...	13075.165279
4	64	32421	POLYGON ((379825.199 6685096.247, 379829.651 6...	3980.682621

DATOS VECTORIALES - CRS

¿Cómo ubicamos objetos o lugares sobre la superficie terrestre?

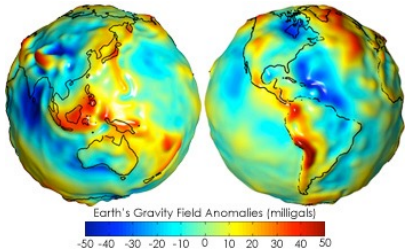
Proyecciones cartográficas: algunos ejemplos

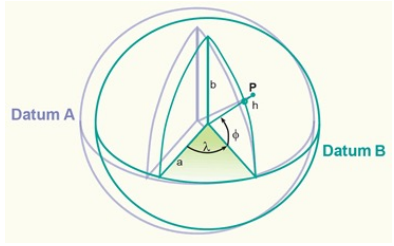


El tipo de proyección a utlizar depende de las características y objetivos del análisis deseado.

DATOS VECTORIALES - CRS

Sistema Coordenado de Referencia (CRS) ➔ define cómo se relaciona un conjunto de coordenadas 2-D con lugares reales en la Tierra (esférica).



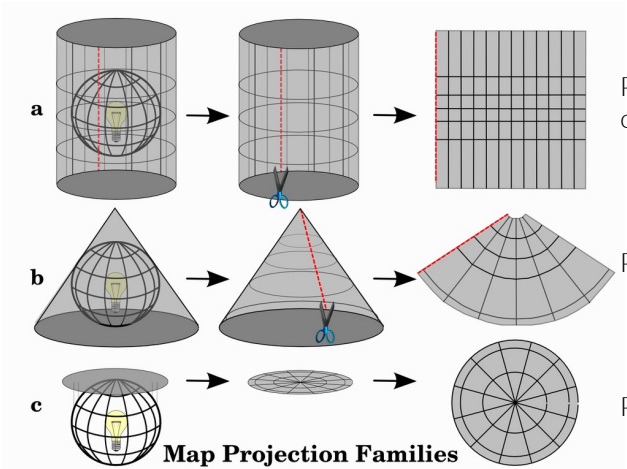


La Tierra no es 100% esférica: existen distintos modelos globales y locales para ajustar su forma exacta (datum).

Un CRS tiene varios componentes clave:

- **Sistema de coordenadas:** grilla ***x-y*** donde se ubican los datos
- **Unidades:** horizontales y verticales.
- **Datum:** modelo de la forma de la Tierra que define el origen del sistema de coordenadas en el espacio.
  - Global ➔ **WGS84** (World Geodetic System) ➔ GPS
  - Local ➔ **SIRGAS 2000**, PSAD56
- **Proyección:** modelo matemático usado para proyectar la superficie de la Tierra sobre una esfera o plano.
- Cada CRS tiene un código identificador único definido por el **European Petroleum Survey Group (EPSG)**.

Proyección esférica



Proyección cilíndrica

Proyección cónica

Proyección polar

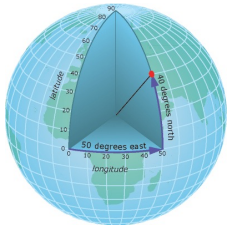
DATOS VECTORIALES - CRS

- Cada conjunto de datos vectoriales tiene un CRS definido.
  - `GeoDataFrame.crs`
- Para operar distintos conjuntos de datos, **hay que convertirlos al mismo CRS.**
  - `gpd.to_crs(crs=4326)`
- Algunos CRS son más convenientes para uno u otro análisis:
  - Medir áreas, distancias ➔ proyección cilíndrica (UTM)

Algunos CRS comunes en Chile

EPSG	Coordenadas	Unidades	Proyección	Datum	Zona
4326	(lat,lon).	grados (°)	Esférica (GCS)	WGS84	Global
32719	(E,N)	Metros (m)	Cilíndrica (UTM)	WGS84	19 S (Chile)
4674	(lat,lon)	grados (°)	Esférica (GCS)	SIRGAS2000	Américas
5361	(E,N)	Metros (m)	Cilíndrica (UTM)	SIRGAS2000	Américas

GCS



UTM