B.STAT.5 Analisi di dati ambientali tramite risorse open source per la data science

Dati georeferenziati, sistemi di riferimento geografici e interpolazione spaziale

Statistica spaziale

In molti contesti è noto il riferimento geografico in cui le osservazioni sono raccolte.

Assunto fondamentale dell'analisi spaziale: ‘‘everything is related to everything else, but near things are more related than distant things” (prima legge della geografia Tobler, 1970).

In altri termini le osservazioni presentano regolarità (correlazione, clustering, trend locali/globali)

I metodi della Statistica Spaziale utilizzano esplicitamente le regolarità spaziali per "migliorare" l'informazione prodotta dall'analisi statistica

Statistica spaziale

John Snow, medico ed epidemiologo inglese del XIX secolo, ha studiato la diffusione dell'epidemia di colera di Londra del 1854, mettendola in relazione con la locazione delle pompe d'acqua nel quartiere di Soho (in particolare quella localizzata nell'attuale Broadwick Street).

Le sue analisi suggerirono l'ipotesi che il colera si diffondesse in modo batteriologico, per contatto (tramite la rete idrica nel caso specifico) e non per via aerea (teoria dei miasmi) come all'epoca si riteneva.

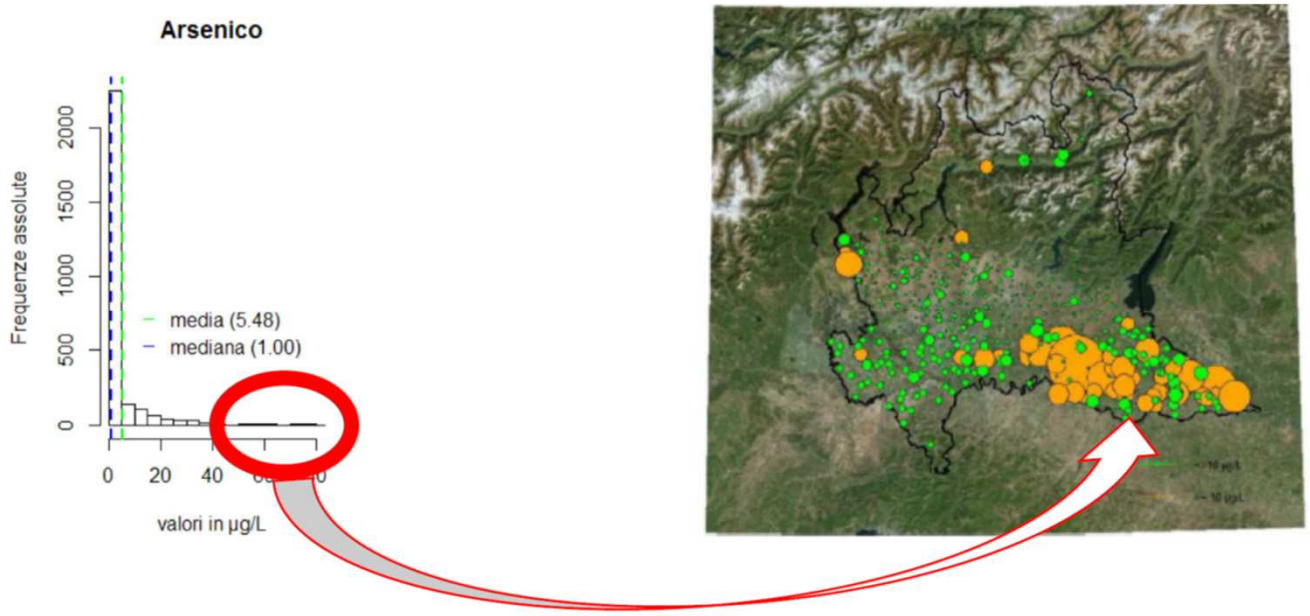
Statistica spaziale

Georeferenziazione - collegamento presente nei dataset tra le variabili oggetto di studio e la collocazione geografica in cui è avvenuta la loro rilevazione. L'osservazione viene indicizzata non in maniera generica, ma secondo un preciso criterio che l'associa al luogo in cui è stata rilevata.

La locazione è potenzialmente rilevante per il fenomeno considerato.

In molti contesti è noto il riferimento geografico in cui le osservazioni sono raccolte.

Laddove disponibile, l'informazione spaziale fornisce una conoscenza più approfondita dei fenomeni.



Georeferenziazione

Collegamento presente nei dataset tra le variabili oggetto di studio e la collocazione geografica in cui è avvenuta la loro rilevazione. L'osservazione viene indicizzata non in maniera generica, ma secondo un preciso criterio che l'associa al luogo in cui è stata rilevata.

Esempio concentrazioni di radon indoor in provincia di Bergamo (dati geostatistici)

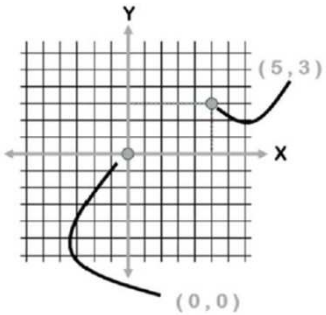
CODICE PUNTO	Riferimento CTR	Est	Nord	Destinazione d'uso dell'edificio	Tipologia edificio	valore misurato (Bq/m^3)
BG001	C4e3	1576261	5082231	LUOGO DI LAVORO	SCUOLA	149.55
BG002	D4a2	1582490	5085329	RESIDENZIALE		250.79
BG003	D4a3	1580416	5084594	LUOGO DI LAVORO	SCUOLA	107.43
BG004	D4a3	1580449	5084615	LUOGO DI LAVORO	SCUOLA	606.20
BG005	D4a2	1582175	5085760	LUOGO DI LAVORO	SCUOLA	252.11
BG006	D4a2	1582449	5085170	LUOGO DI LAVORO	SCUOLA	50.45
BG007	D4a2	1582106	5085308	RESIDENZIALE		978.67
BG008	D4a2	1582775	5086078	RESIDENZIALE		46.84
BG009	C4e3	1576626	5081678	RESIDENZIALE		83.92
...	...	...	...	...	...	...

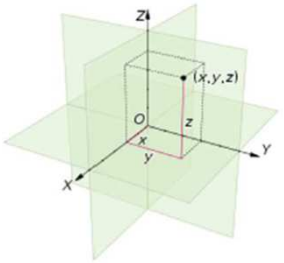
Edifici georeferenziati in base al *sistema di coordinate* Gauss-Boaga.

Per ogni edificio viene riportata la **concentrazione di radon indoor** e alcune caratteristiche dell'edificio.

Proiezioni e sistemi di coordinate

La georeferenziazione posiziona ogni osservazione sulla superficie terrestre. Questa operazione richiede sistemi/modelli di riferimento (*datum geodetico*).





Idea simile al posizionamento usuale di un punto in un piano o in uno spazio 3D

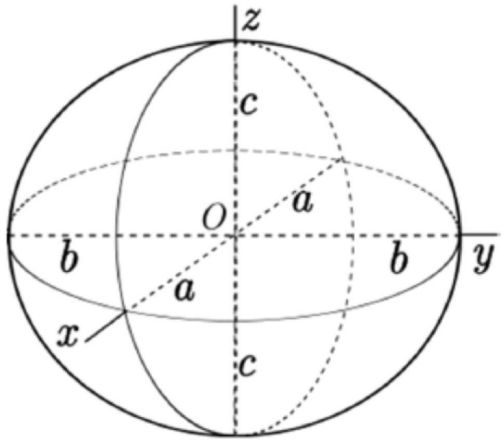
Sistema di riferimento geografico estende l'idea del sistema cartesiano alla superficie della Terra.

Richiede la specificazione di un modello «semplice»

Proiezioni e sistemi di coordinate

Ellissoide: approssimazione matematica della Terra ottenuta tramite una superficie di equazione definita da 2 parametri

$\frac{x^2 + y^2}{a^2} + \frac{z^2}{c^2} = 1$ a: semiasse equatoriale c: semiasse polare



L'ellissoide si ottiene per rotazione di un'ellisse di equazione $\frac{x^2}{a^2} + \frac{z^2}{c^2} = 1$ attorno al suo asse minore (sferoide oblato).

Una scelta $c < a$ induce la tipica forma schiacciata sui poli.

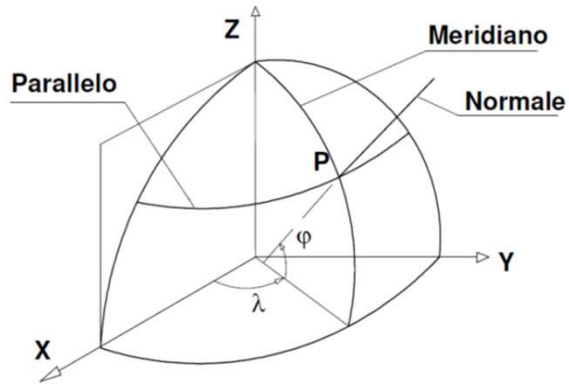
L'ellissoide non è unico: poiché la Terra non è uno sferoide perfetto, si possono avere diversi *sistemi di riferimento geografico*

Coordinate geografiche

Ogni punto P sull'ellissoide può essere espresso tramite una coppia di **coordinate geografiche:**

latitudine (φ): angolo formato dal vettore congiungente P con il centro dell'ellissoide e dal piano passante per l'origine e ortogonale all'asse di rotazione (piano equatoriale)

longitudine (λ): angolo formato dal vettore congiungente P con il centro dell'ellissoide e dal piano identificato dal meridiano di riferimento



Parallelo: sezione (circonferenza) ottenuta intersecando la superficie dell'ellissoide con un piano perpendicolare all'asse di rotazione. Il parallelo posizionato sul piano equatoriale è detto **equatore**

Meridiano: sezione (ellisse) ottenuta secando l'ellissoide con piani uscenti dall'asse di rotazione. Meridiano **di riferimento** (o fondamentale) passa dall'Osservatorio di Greenwich (Londra)

Sistema WGS84 (World Geodetic System 1984)

È un sistema di riferimento di coordinate geografiche globale tra i più usati nel mondo.

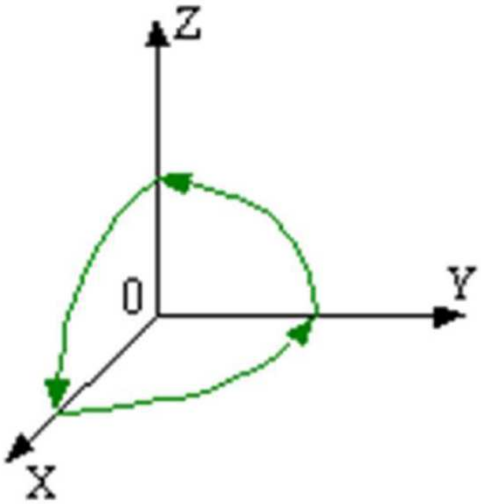
Origine: centro della massa terrestre

Asse Z: asse di rotazione terrestre (passante per il Polo Nord)

Assi X e Y posizionati sul piano equatoriale

Asse X: identificato dal meridiano di riferimento

Asse Y: identificato da una rotazione antioraria di X per un osservatore posizionato su Z



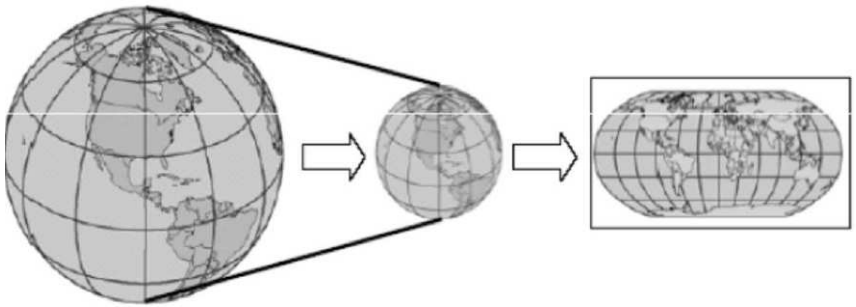
Parametri dell'ellissoide: $a = 6378137\,m$ e $e^2 = 0.00\,669438$ (eccentricità)

I sistemi di rappresentazione cartografica

Permettono il calcolo delle coordinate piane (cartografiche) (x,y), espresse in unità di lunghezza, a partire dalle coordinate geografiche (λ, ϕ). Esistono vari metodi di proiezione dell'ellissoide.

La rappresentazione planare comporta necessariamente delle **deformazioni**

Carta geografica: rappresentazione grafica simbolica, riprodotta in scala della superficie terrestre



Proiezione planare della superficie

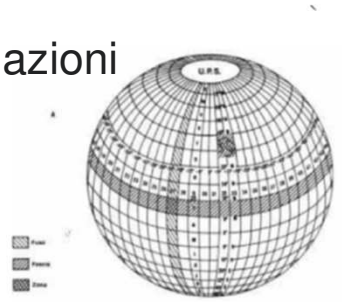
Riduzione degli oggetti rappresentati

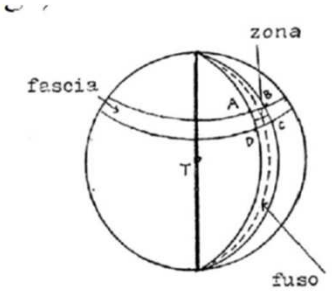
Scala della carta: rapporto tra le dimensioni degli oggetti rappresentati e reali

Es. scala 1:10000 (1 a 10000): un centimetro sulla carta equivale a 10000 cm nella realtà

Il sistema UTM (Universal Transverse Mercator)

Proiezione locale della superficie terrestre utile per limitare le deformazioni





Il globo terrestre è suddiviso in

20 fasce di latitudine di 8°, identificate da una lettera

60 fusi di longitudine di 6°, identificati da un numero

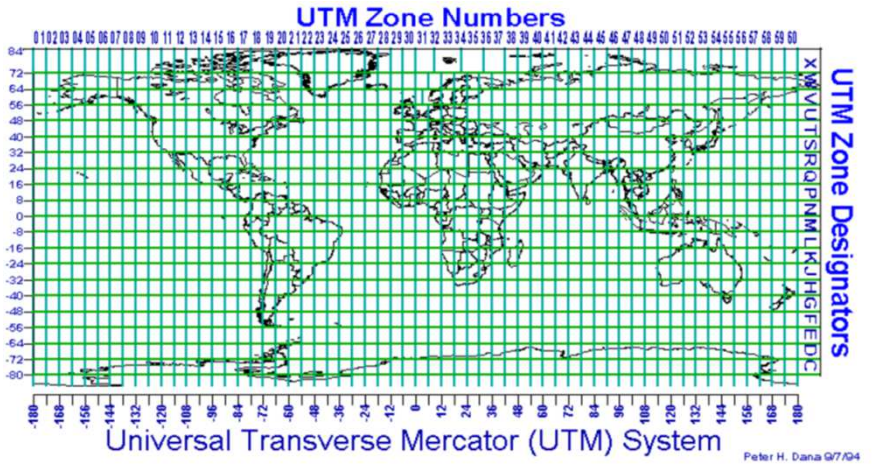
La numerazione dei fusi inizia dall'antimeridiano di riferimento passante per Greenwich.

Una **zona** è identificata dall'incrocio di un fuso e una fascia.

L'Italia è compresa nei fusi 32-33-34 e nelle fasce S e T

(zone 32S-32T-33S-33T-34S-34T)

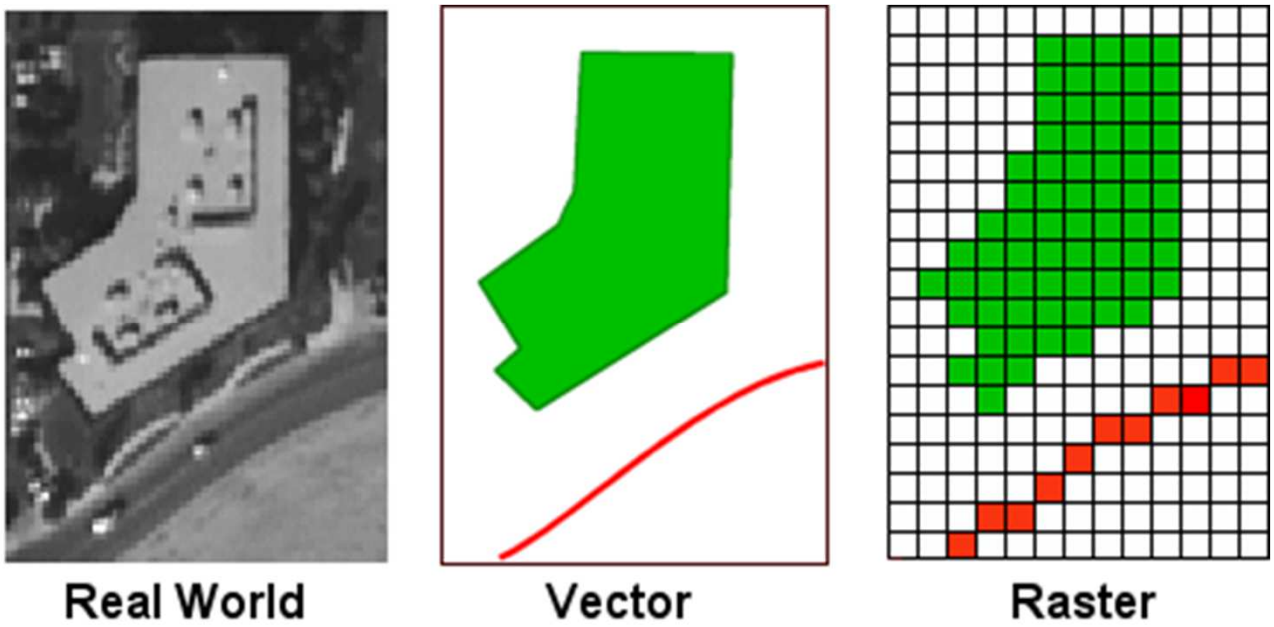
Le coordinate cartografiche (x,y) (espresse in metri) sono riferite all'equatore e al meridiano centrale.



I formati dei dati spaziali

Formato raster e formato vettoriale.

Entrambi i formati possono essere usati per rappresentare oggetti spaziali, ma presentano differenze nella struttura usata per la digitalizzazione e nella visualizzazione dell'immagine

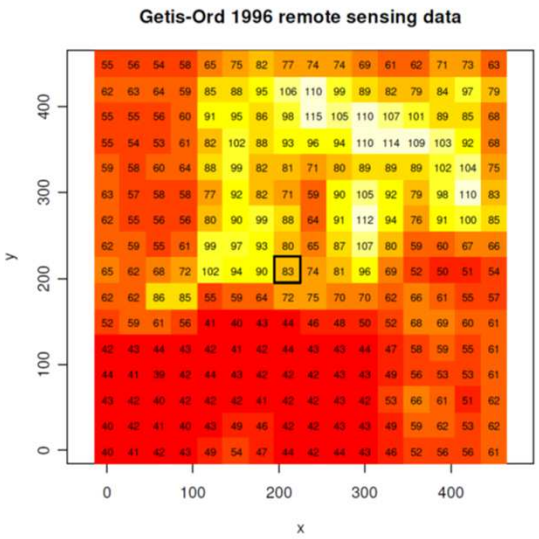


Il formato raster

Dati raster: rappresentazione di fenomeni attraverso una matrice di celle (pixel: picture element).

Ogni raster ha un'origine che è posizionata in uno dei quattro vertici della griglia.

rilevazione	x	y	val
1	0	0	40
2	30	0	41
3	60	0	42
4	90	0	43
…	…	…	…
17	0	30	40
18	30	30	42
19	60	30	41
	…	…	…



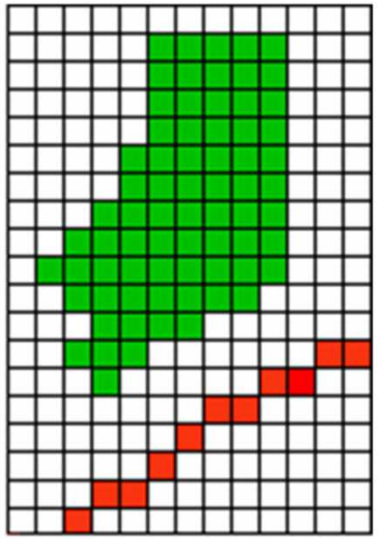
Ogni pixel è identificato dalle coordinate di un suo punto rappresentativo (per esempio centroide o vertice) espresso nel sistema di riferimento scelto.

La dimensione del pixel è legata alla precisione del dato.

Il formato raster (2)

I dati raster rappresentano gli oggetti tramite una matrice di celle:

- gli elementi puntuali sono rappresentati da singole celle
- gli elementi lineari da una serie di celle adiacenti caratterizzate da un attributo comune (per esempio il colore)
- gli elementi poligonali da regioni di celle adiacenti caratterizzate dalla medesima modalità di un attributo.

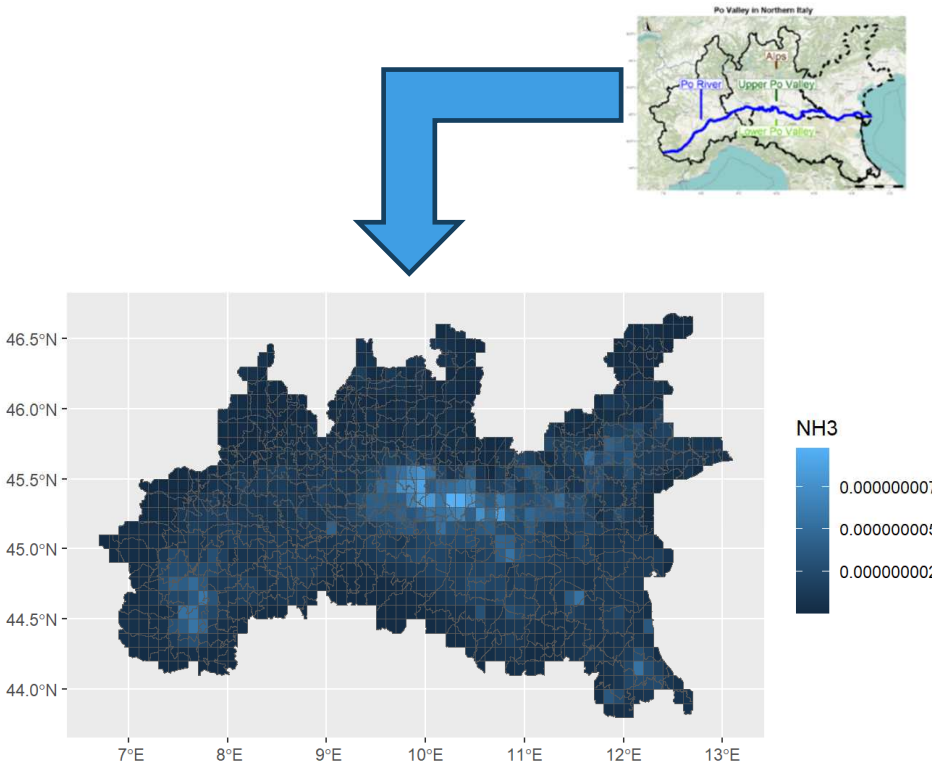


Raster

Il formato raster (3)

Un raster può essere ottenuto da fotografie/immagini da aerei, satelliti o sonde (telerilevamento o remote sensing), griglie a maglia regolare, ….

La cartografia raster è adatta a rappresentare fenomeni con natura continua, per esempio una carta di acclività di un versante o una superficie di concentrazione di inquinante.



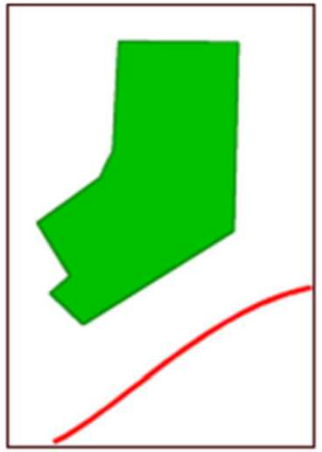
Emissioni di NH3 nel bacino del Po (dati EEA 2023)

Il formato vettoriale

I dati vettoriali sono costituiti da elementi semplici: punti, linee e poligoni, codificati e memorizzati sulla base delle loro coordinate.

Un punto è individuato attraverso le sue coordinate (x, y);

Una linea o un poligono attraverso la posizione dei suoi nodi $[(x_1, y_1); (x_2, y_2); \cdots; (x_N, y_N)]$.



Vector

La cartografia vettoriale è adatta a rappresentare dati che variano in modo discreto e caratterizzati da forme e bordi ben precisi, per es. l'ubicazione di centraline meteo, la rappresentazione delle strade di una città …

Il formato vettoriale (2)

Il formato vettoriale è adatto a rappresentare dati che variano in modo discreto e caratterizzati da forme e bordi ben precisi, per es. l'ubicazione di centraline, i contorni di un'area (provincia,…)…

ID	Est	North	Building Use	Building Type	Rn Concetration (Bq/m³)
BG001	1576261	5082231	LUOGO DI LAVORO	SCUOLA	149.55
BG002	1582490	5085329	RESIDENZIALE		250.79
BG003	1580416	5084594	LUOGO DI LAVORO	SCUOLA	107.43
BG004	1580449	5084615	LUOGO DI LAVORO	SCUOLA	606.20
BG005	1582175	5085760	LUOGO DI LAVORO	SCUOLA	252.11
BG006	1582449	5085170	LUOGO DI LAVORO	SCUOLA	50.45
BG007	1582106	5085308	RESIDENZIALE		978.67
BG008	1582775	5086078	RESIDENZIALE		46.84
BG009	1576626	5081678	RESIDENZIALE		83.92
…	…	…	…	…	…

Il dato vettoriale è ricavato da rilevazioni GPS, digitalizzazione di mappe cartacee, misure a terra o importato da file grafici di scambio (per esempio AutoCAD).

Gli shape file

Sono un formato standard per i dati vettoriali, sviluppato da ESRI, fornitore leader a livello mondiale di sistemi e applicazioni per la gestione di basi di dati georeferenziate.

Collezione di file con tipica struttura **nome.estensione**

I file obbligatori sono 3 e hanno estensione .shp, .shx, .dbf. File addizionali forniscono informazioni aggiuntive.

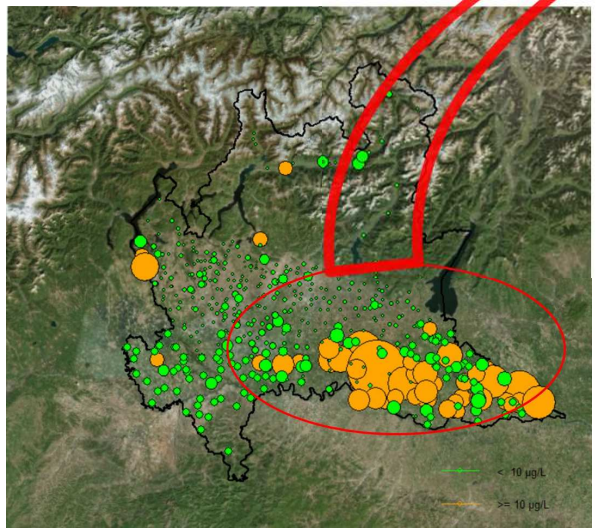
.shp conserva le geometrie. Ogni record codifica un'entità spaziale tramite dei vertici

.**shx** conserva l'indice delle geometrie

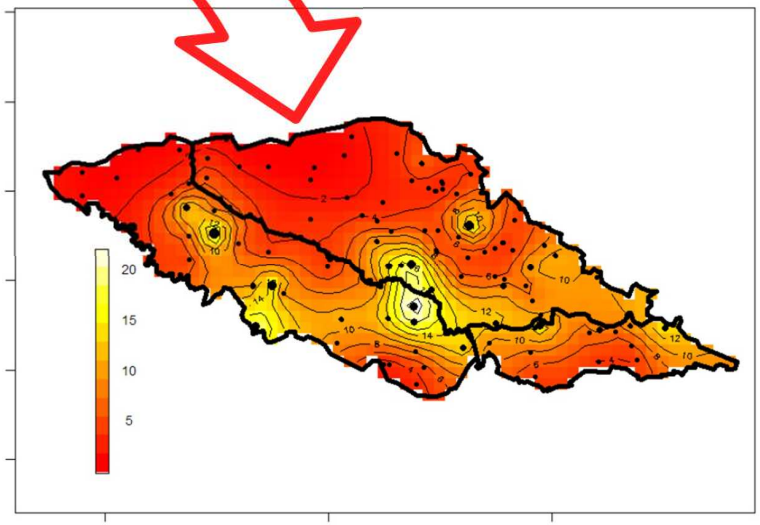
.**dbf** tabella dati degli attributi relativi alle geometrie. Ogni riga corrisponde a un oggetto codificato nel file shp.

Stima delle superfici spaziali

Uno dei principali obiettivi in contesto ambientale è realizzare delle mappe rappresentanti l'andamento delle concentrazioni di una sostanza nel terreno, nell'aria o nell'acqua in una certa porzione di spazio utilizzando l'informazione raccolta in alcuni punti di monitoraggio



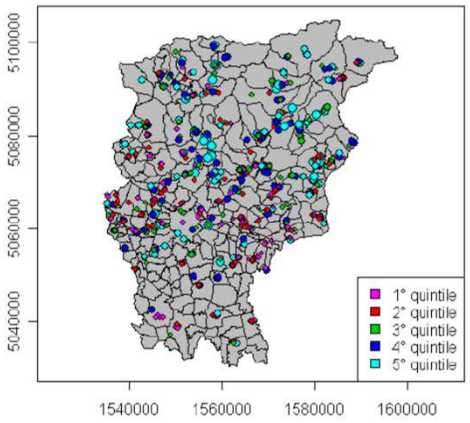
Concentrazioni in $\mu g/L$ di arsenico negli acquiferi lombardi (2009-2012)

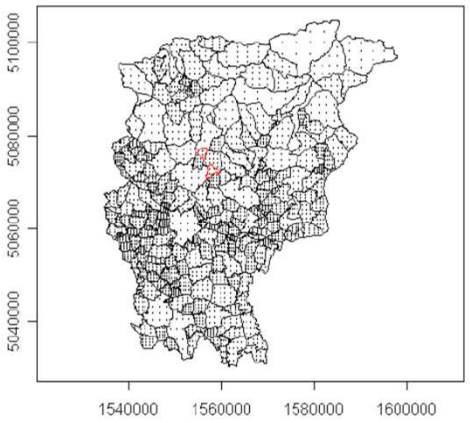


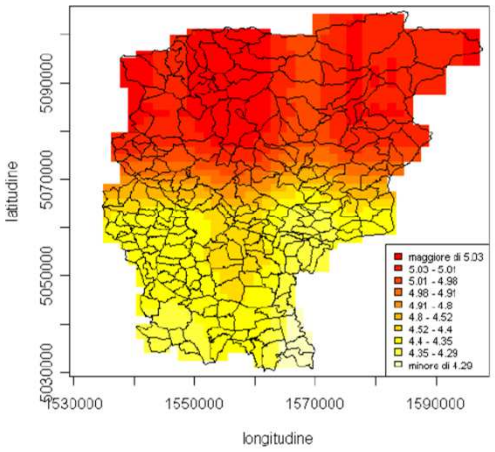
Stima delle superfici spaziali

Quello che occorre fare è imparare a farlo su un punto.

La procedura deve «solo» essere replicata su un grigliato di valori (approssimazione discreta) da trasformare in un raster







Concentrazioni di Rn indoor in provincia di Bergamo (campagna di monitoraggio ARPA Lombardia 2003-2004)

Stima delle superfici spaziali - una formalizzazione «dolce»

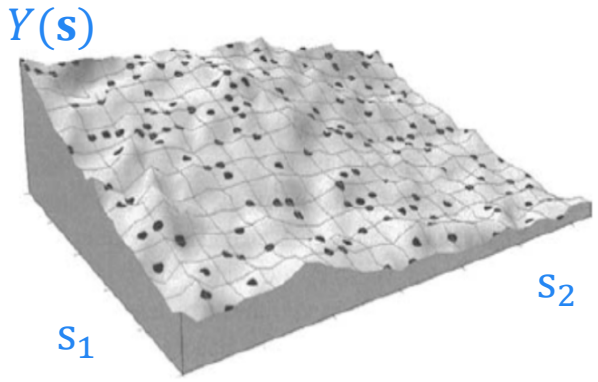
Sia Y la variabile a cui siamo interessati (per es. la concentrazione di PM10 nell'aria nella valle del Po)

Assumiamo che questa variabile sia una funzione dello spazio: $Y(\mathbf{s})$

- $\mathbf{s} = (s_1, s_2) = (\text{longitudine}, \text{latitudine})$
- $\mathbf{s} \in W$, W rappresenta la regione d'interesse (per esempio: la valle del Po)

Siano

- $\mathbf{s_1}, \cdots \mathbf{s_n}$ punti dello spazio su cui abbiamo osservato Y
- $Y_1 = Y(\mathbf{s_1}), \cdots, Y_n = Y(\mathbf{s_n})$ i corrispondenti valori di Y



Stima del valore della superficie $Y(s)$ in un punto

Sia $\mathbf{s_0} = (s_{0,1}, s_{0,2}) \in W$ un generico punto in cui Y non è misurata

Vogliamo ottenere un'approssimazione di $Y_0 = Y(\mathbf{s_0})$ sfruttando l'informazione campionaria $Y_1 = Y(\mathbf{s_1}), \cdots, Y_n = Y(\mathbf{s_n})$

Indichiamo con $\hat{Y}_0 = \hat{Y}(\mathbf{s_0})$ questa *stima*

Come possiamo ottenere $\hat{Y}_0$?

L'idea è di usare i dati disponibili dando importanza maggiore alle misurazioni di Y ottenute in prossimità di $\mathbf{s_0}$

Stima del valore della superficie $Y(s)$ in un punto (2)

Esistono vari modi per ottenere la stima desiderata. Un metodo di interpolazione relativamente semplice è denominato **Inverse distance weighting** (IDW)

Definizione dell'interpolatore su $\boldsymbol{s}_0$: $\hat{Y}_0 = \sum_{i=1}^{n} Y_i \, \omega_i(\boldsymbol{s}_0)$

$\{\omega_i(\boldsymbol{s}_0), i = 1, \cdots, n\}$ è uno schema di ponderazione che dà importanza maggiore alle misurazioni di Y ottenute in prossimità di $\boldsymbol{s_0}$

Schema di ponderazione dell'IDW

$\omega_i(\boldsymbol{s}_0) = \frac{g(d(\boldsymbol{s}_i,\boldsymbol{s}_0))}{\sum_{i=1}^{n} g(d(\boldsymbol{s}_i,\boldsymbol{s}_0))}$ per ogni $i = 1, \cdots, n$

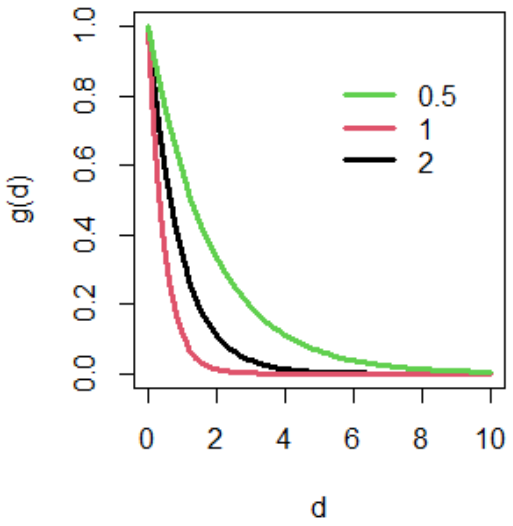
- $d(\boldsymbol{s}_i, \boldsymbol{s}_0)$: funzione di distanza tra i due punti $\boldsymbol{s}_i$ e $\boldsymbol{s}_0$

  per esempio la distanza euclidea $d(\boldsymbol{s}_i, \boldsymbol{s}_0) = \sqrt{(s_{i1} - s_{01})^2 + (s_{i2} - s_{02})^2}$

- $g(d) = \mathbb{R}^+ \to \mathbb{R}^+$ monotona decrescente quindi
  - ✓ $0 < \omega_i(\boldsymbol{s}_0) < 1$
  - ✓ $\sum_{i=1}^{n} \omega_i(\boldsymbol{s}_0) = 1$

  Tipica scelta $g(d) = d^{-\lambda}$ con $\lambda > 0$ parametro esogeno

  (ogni funzione monotona decrescente potrebbe ndare bene)



Distanza sì … ma quale?

La funzione di distanza più appropriata dipende dal contesto

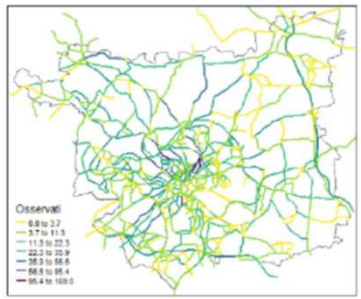
- Modellizzazione delle traiettorie di migrazioni delle rondini

  Occorre tenere conto della curvatura terrestre ➔ distanza geodetica

- Modellizzazione di incidenti stradali

  Occorre tenere conto delle distanze sui percorsi stradali ➔

  distanza su grafo

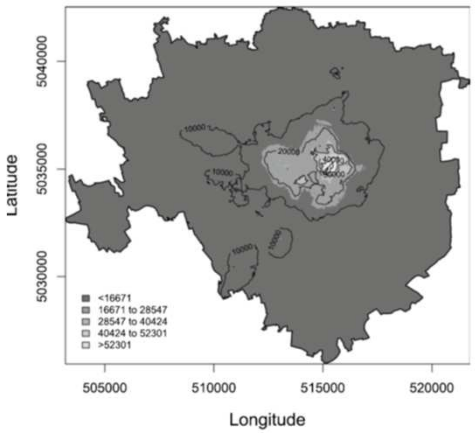


- … …

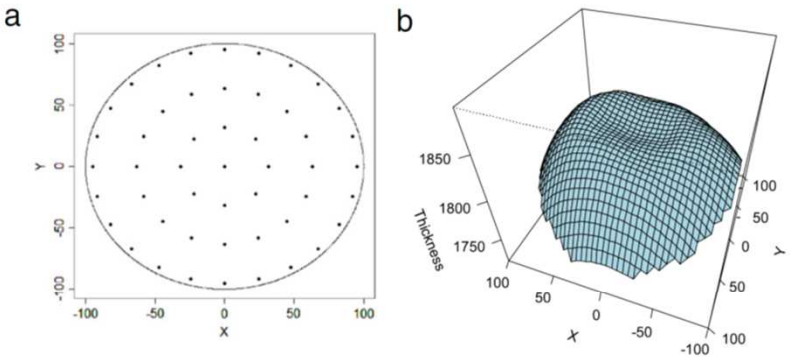
La scelta della metrica modifica i valori dei pesi, ma la procedura di interpolazione rimane comunque invariata

La spatial datascience non è solo «per l'ambiente» …

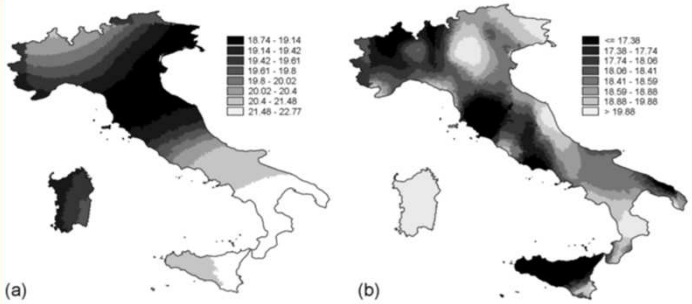
Le tecniche di statistica spaziale e della spatial datascience non hanno una rilevanza solo in contesti ambientali ma in molte discipline diverse dalle «hard sciences» all'economia e alle scienze sociali fino al contesto aziendale e industriale



Prezzi degli immobili nel comune di Milano 2004-2010



Monitoraggio superfici di SiO2 nella realizzazione di chip in microelettronica



Età mediana al primo rapporto sessuale degli studenti universitari italiani (1995-96)

Verso lo spazio-tempo…

Le dinamiche spaziali dei fenomeni possono modificarsi nel tempo: le concentrazioni di PM di notte sono diverse da quelle diurne, quelle estive da quelle invernali, quelle del fine settimana da quelle dei giorni feriali, ….

Questo è legato al fatto che i fenomeni ambientali interagiscono con il contesto che li circonda (aspetti antropici e urbanizzazione, aspetti climatici quali umidità, velocità del vento, temperatura,…)

Occorre quindi introdurre, laddove rilevante, questa ulteriore dimensione.

Ovviamente questo ha un costo in termini di reperimento dei dati, modelli e metodi di elaborazione, risorse di calcolo, metodi di rappresentazione dei risultati, ….