

Dati multi-fonte e analisi territoriali

Marco Tosi, Irene Barbiera, e Federico Gianoli
a.a. 2025/2026

Presentazione del Corso

Ci presentiamo

Marco Tosi

- **Professore Associato** in Statistica Sociale.
- Precedentemente, Postdoc all'Università di Colonia, al Collegio Carlo Alberto (Torino), e London School of Economics and Political Science.
- **MI occupo di:**
 - Relazioni e dinamiche familiari
 - Invecchiamento e salute
 - Dati longitudinali

Progetto di ricerca: <https://sites.google.com/view/kinhealth/home>



Ci presentiamo

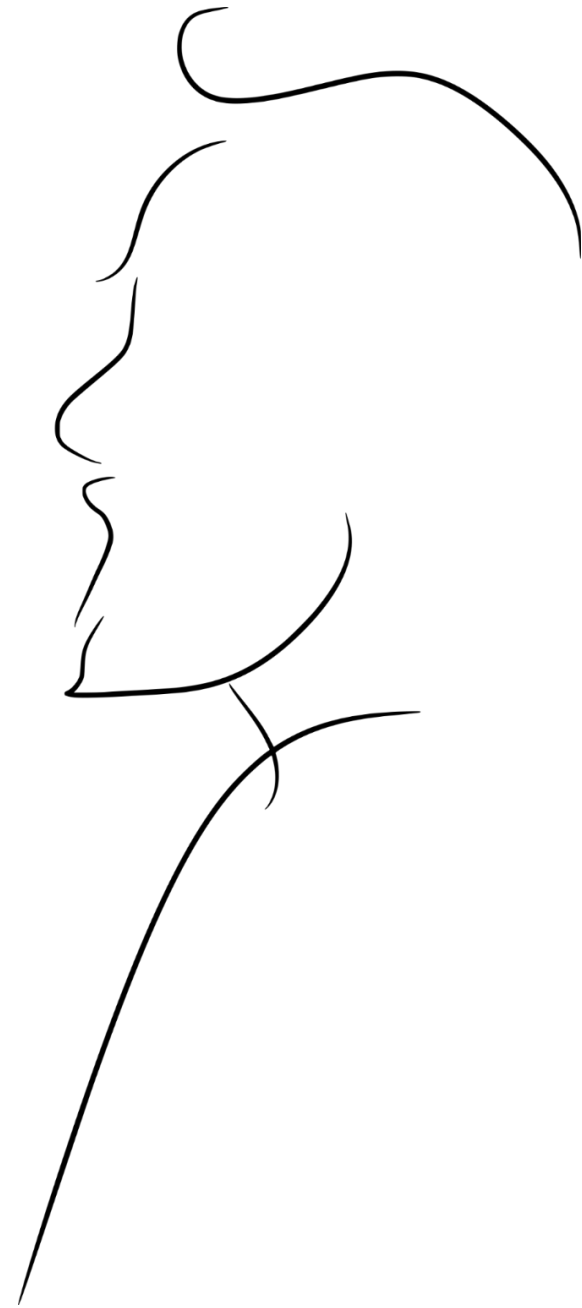
Irene Barbiera

- Professoressa Associata di Demografia
- Mi occupo di demografia storica
 - dinamiche di popolazione e regimi di mortalità nel passato
 - migrazioni nel mondo antico
 - storia di genere, sessualità e sex ratio
- Sono stata post doc e docente presso la Central European University (Budapest-Vienna) e l'Österreichische Akademie der Wissenschaften di Vienna.

Ci presentiamo

Federico Gianoli

- Geografo
- Mi occupo di cartografia, analisi dati spaziali e sviluppo GIS
- Consulente per la Commissione Europea
- Membro della comunità di QGIS Italia e socio di Gfoss.it
- Docente al Master di II Livello in GIScience dell'Università di Padova



IL CORSO di Dati Multi-fonte e analisi territoriale

IL CORSO è LABORATORIALE!!

- E' pensato per essere svolto in presenza (per frequentanti)
- Poca teoria, molta pratica
- 64 ore in lab. ASID 20
- E' diviso in 2 Moduli: (1) Analisi Territoriale & (2) Analisi Multi-Fonte

Struttura del Corso

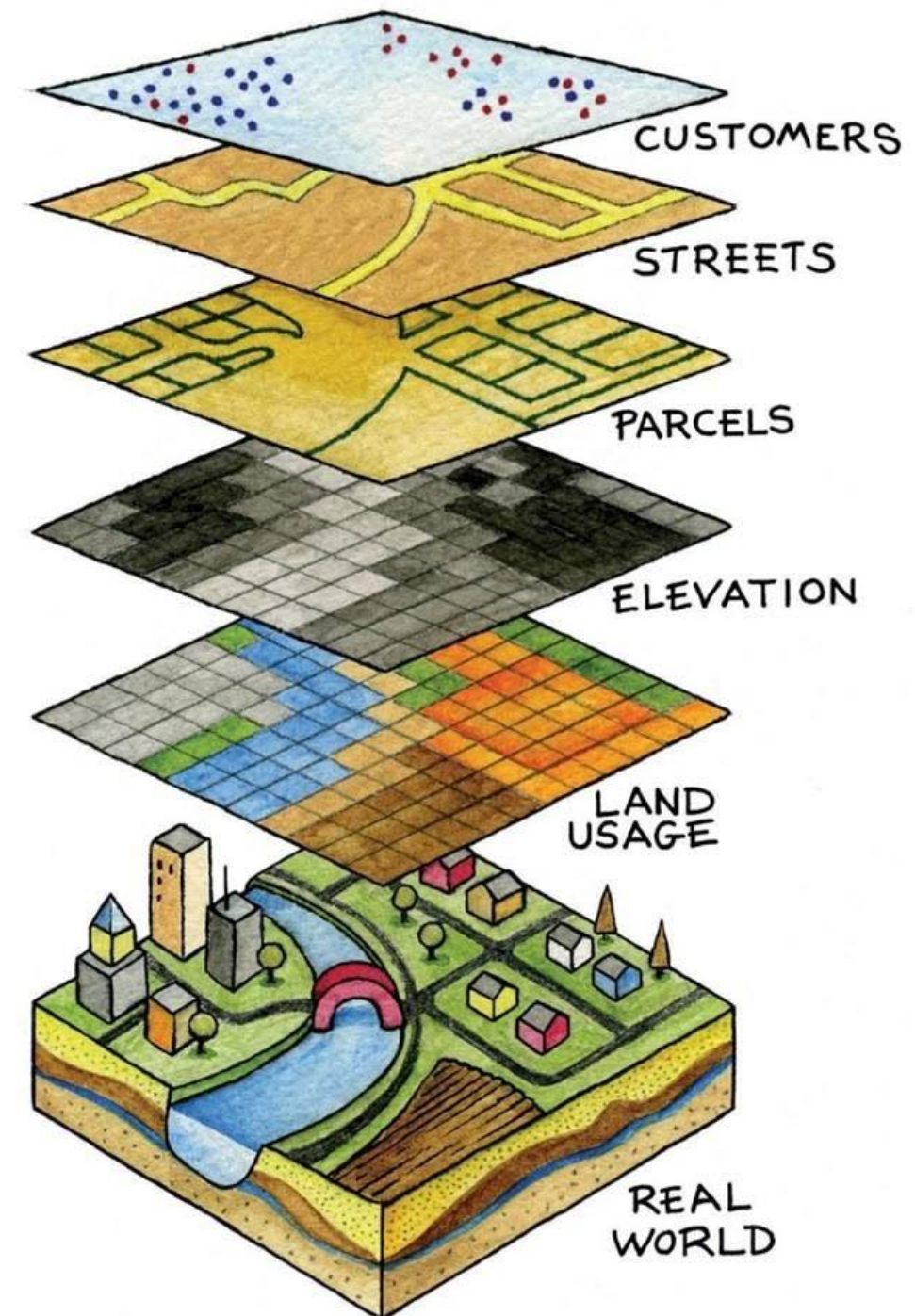
- Settembre: Introduzione
 - *Introduzione ai dati SHARE, STATA software, Alcuni concetti di analisi dati e disegno della ricerca.*
- Ottobre (fino al 20): Analisi Territoriale
 - *Obiettivo:* Analizzare dati territoriali con l'utilizzo di QGIS
 - 6 e 7 Novembre esercitazione di QGIS
- Novembre (fino a circa metà): Record Linkage
 - *Obiettivo:* Abbinare dataset provenienti da fonti diverse.
- Novembre/Dicembre: Imputazione dei dati mancanti
 - *Obiettivo:* Ridurre le distorsioni dovute ai dati mancanti.
- Simulazione di Esame (settimana 15-19 Dicembre?)

MODULO 1.

- Introduzione ai Sistemi Informativi Geografici
 - Modelli di dati Raster e Vettoriali
 - Sistemi di Riferimento
 - Licenze Dati

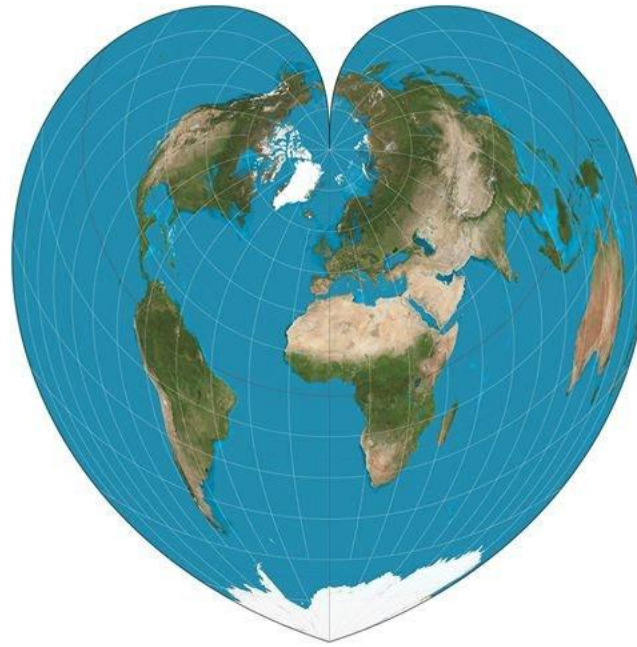
GIS

- Cosa sono i GIS
- Modelli di Dati
 - Raster
 - Vettoriali



GIS

Lavorare coi Sistemi
di Riferimento



Licenze Dati

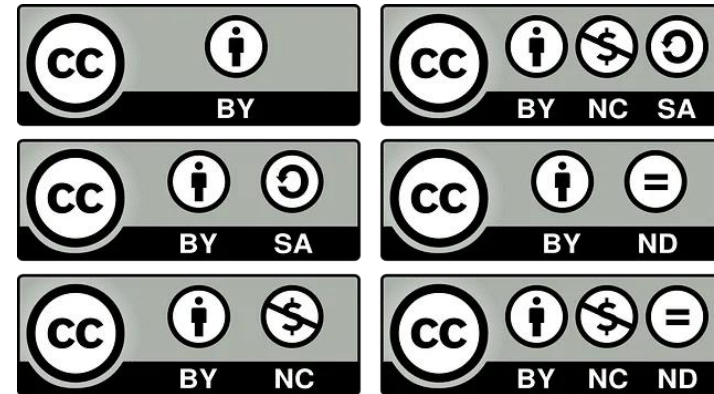
I dati hanno delle
Licenze che ci dicono
come e quando
possono essere
utilizzati



Copyright

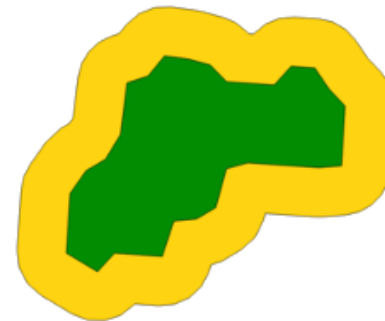
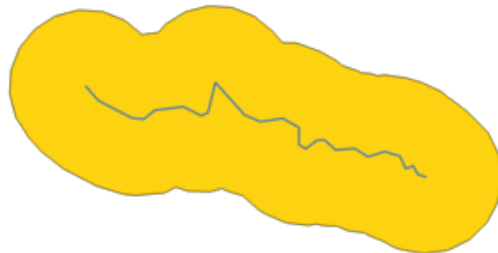
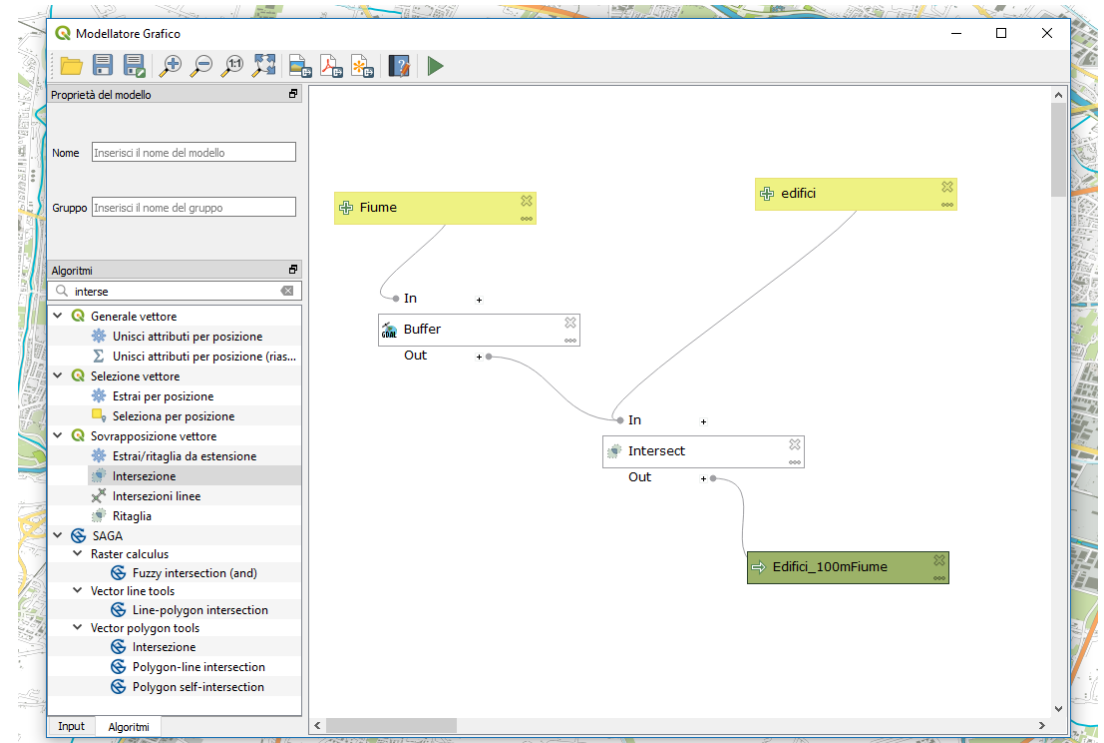


Copyleft



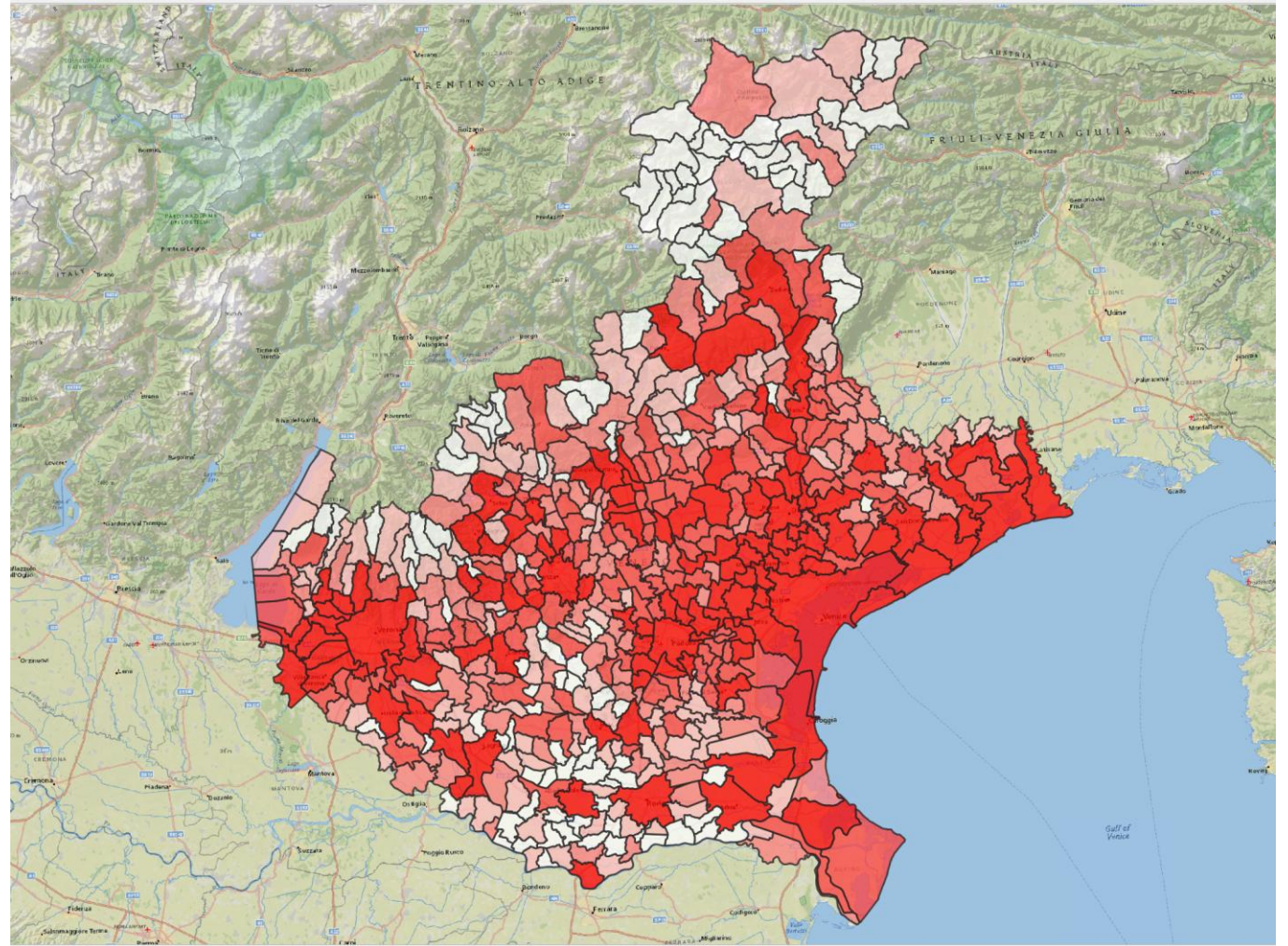
Geoprocessing

Come fare operazioni
spaziali tra dati per
ricavare nuove
informazioni



QGIS

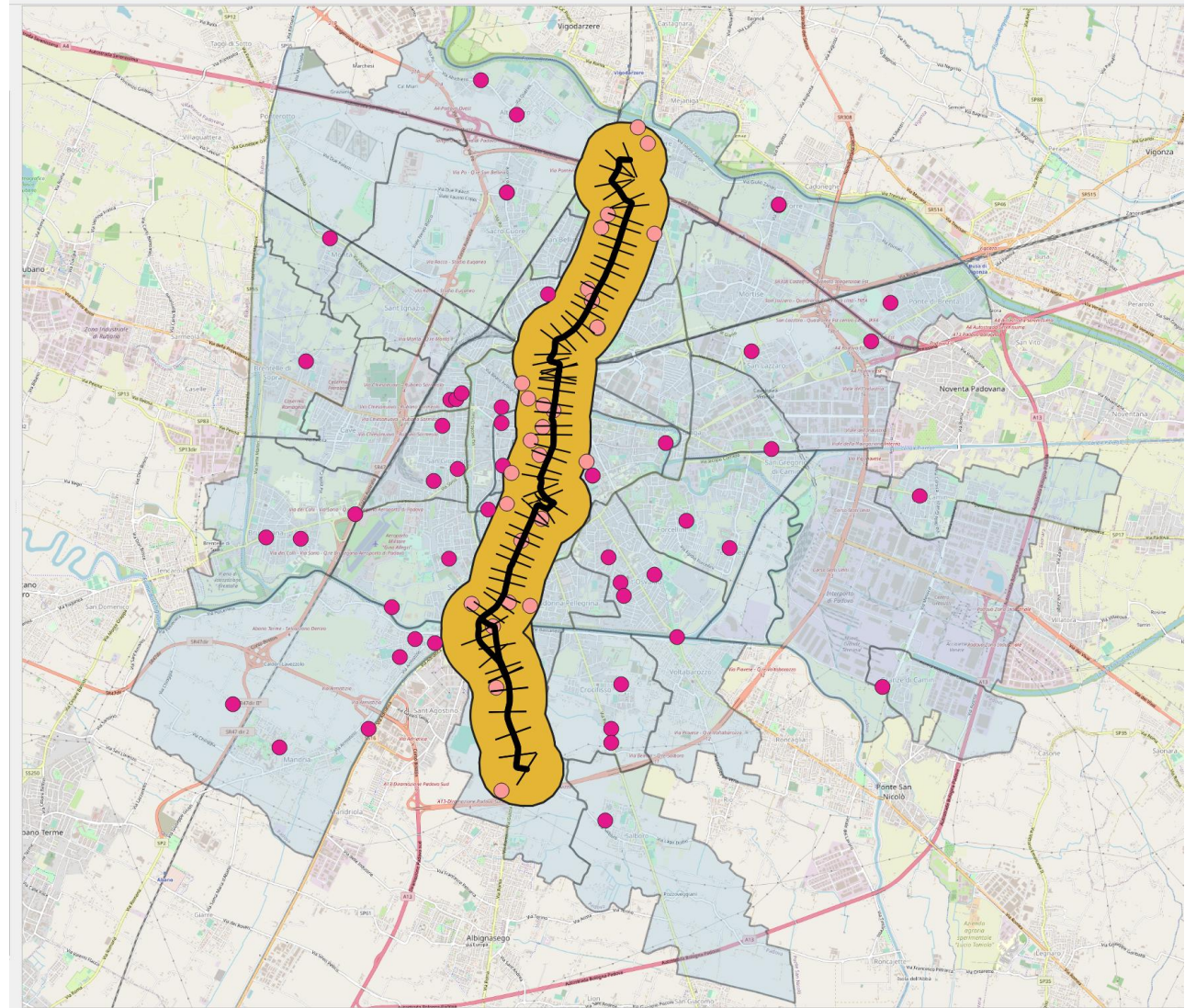
- Il software
- Lavorare coi dati
 - Tabelle di attributi
 - Join
 - Analisi Spaziale



La popolazione nei diversi comuni del Veneto, 2019

Layer e contenuti

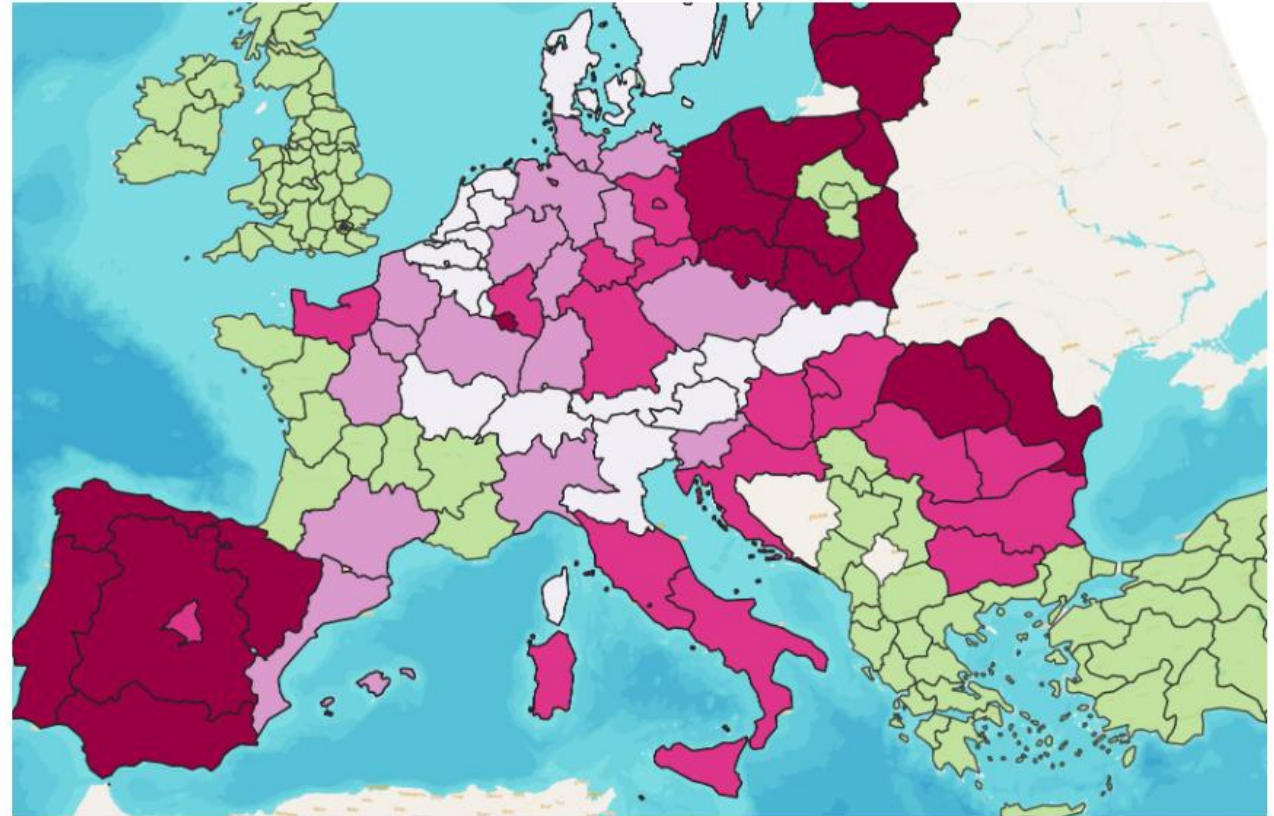
- Relazionare diversi elementi di una mappa
 - Creare uno shapefile
 - Creare layer di dati selezionati
 - creare un buffer



Case in vendita a Padova lungo la linea del tram

Layers e contenuti

- Relazionare diversi elementi di una mappa
 - Creare uno shapefile
 - Creare layer di dati selezionati
 - Creare una relazione tra diversi layers e tabelle



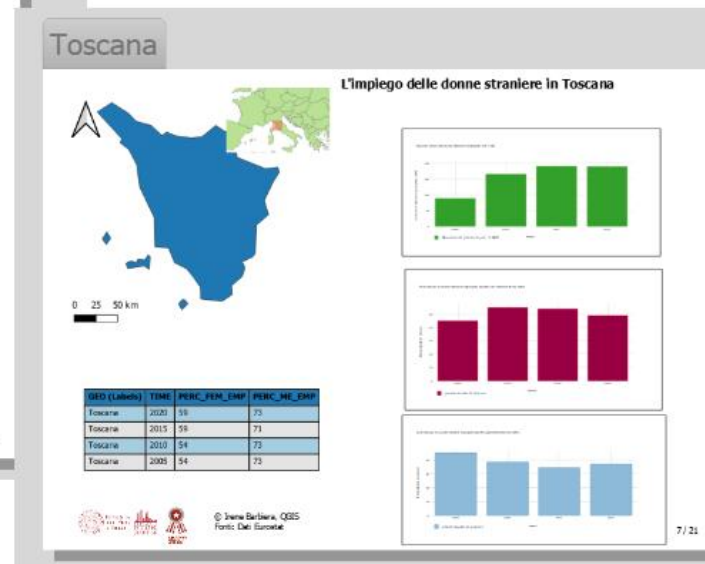
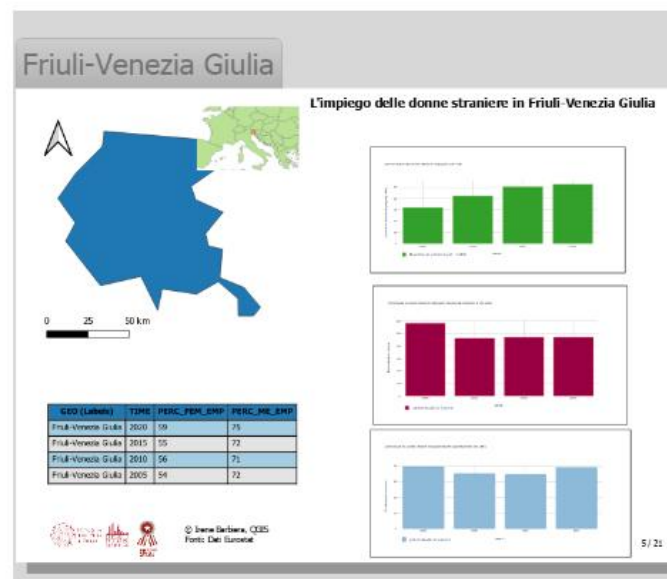
Media regionale di salute auto percepita, wave 8, dati SHARE

«Mentire» con le mappe

- Come presentare una mappa
 - Scelta dei contenuti
 - Scala
 - Leggenda
- Creare una atlante
 - Mappa e panoramica
 - Grafici e tabelle

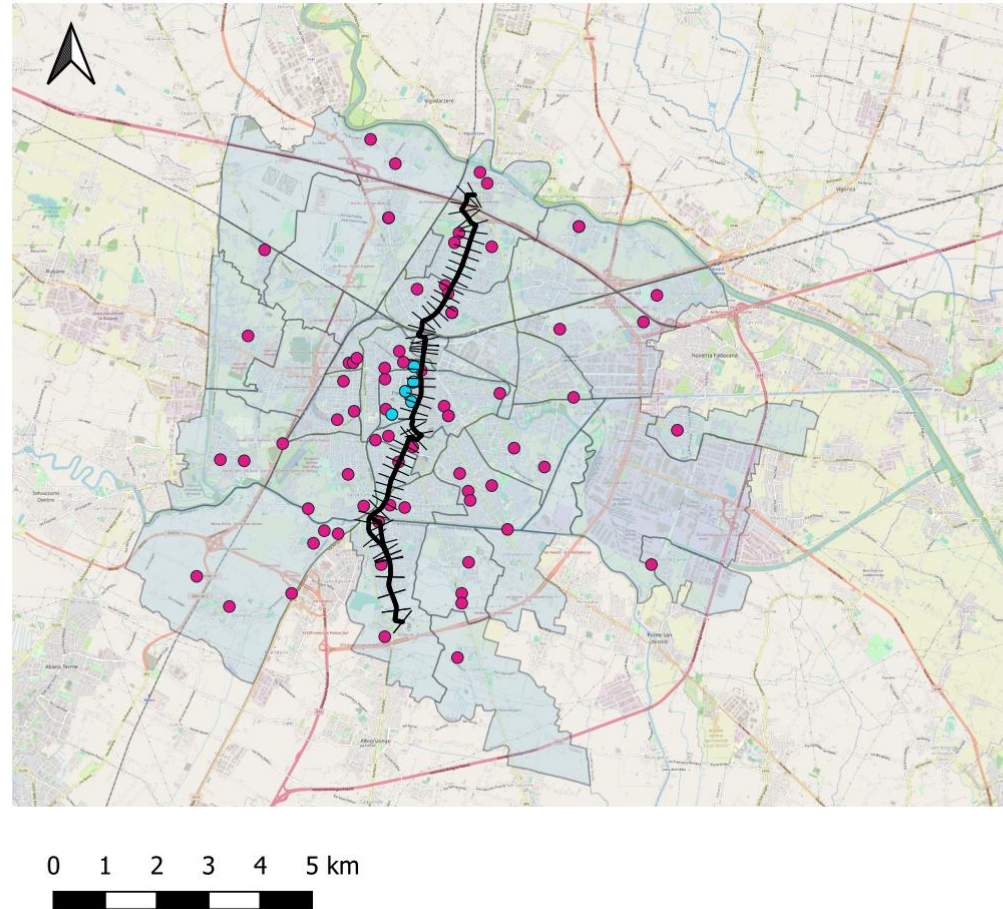


Figura 2. Percentuale di donne straniere impiegate in Italia nel 2020 per macro regione



«Mentire» con le mappe

- Come presentare una mappa
 - Scelta dei contenuti
 - Scala
 - Leggenda

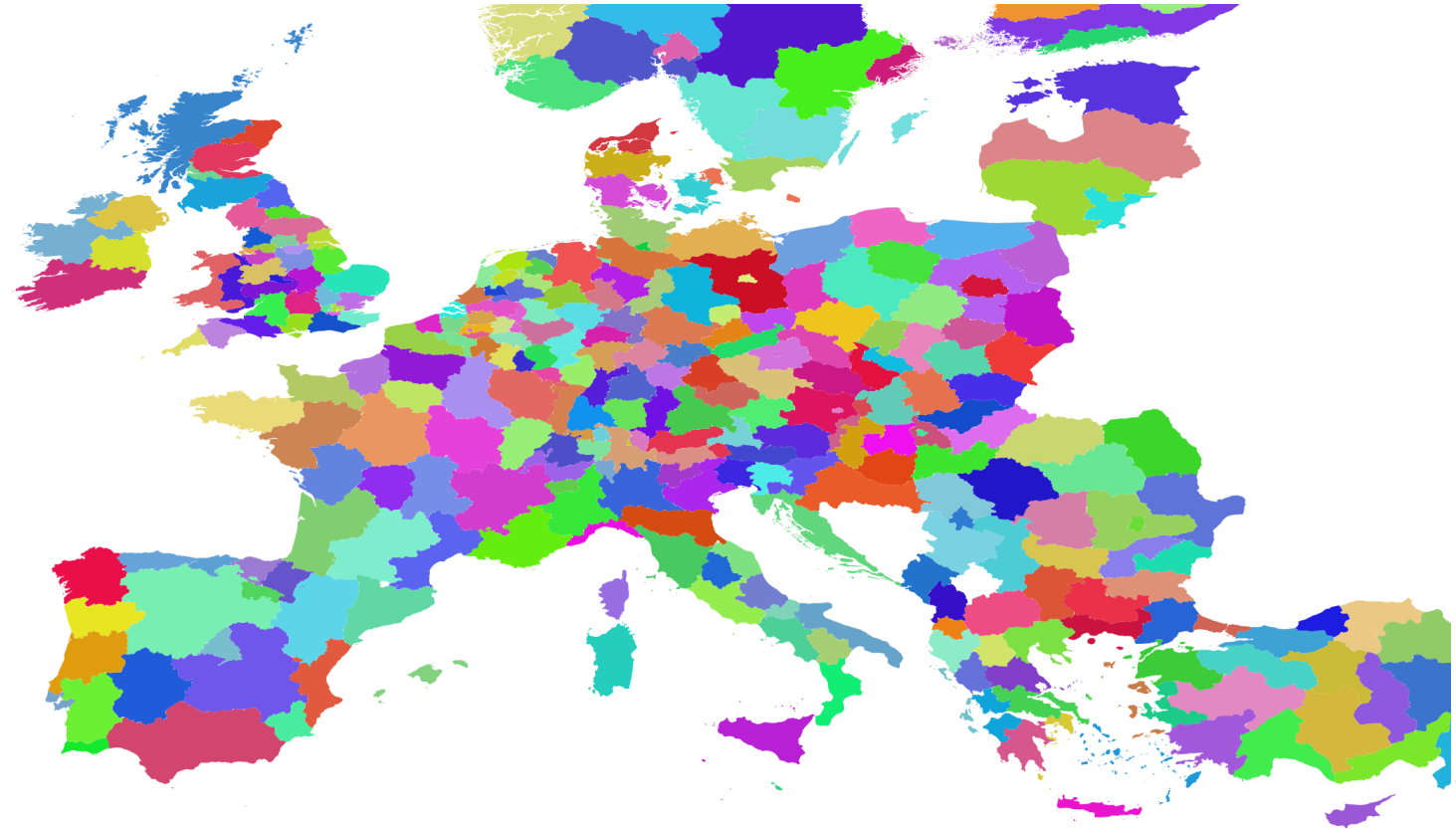


Leggenda

- Linea tram 1
- CASE vicino al tram in zona piazze
- case in vendita a Padova
- Quartieri di Padova

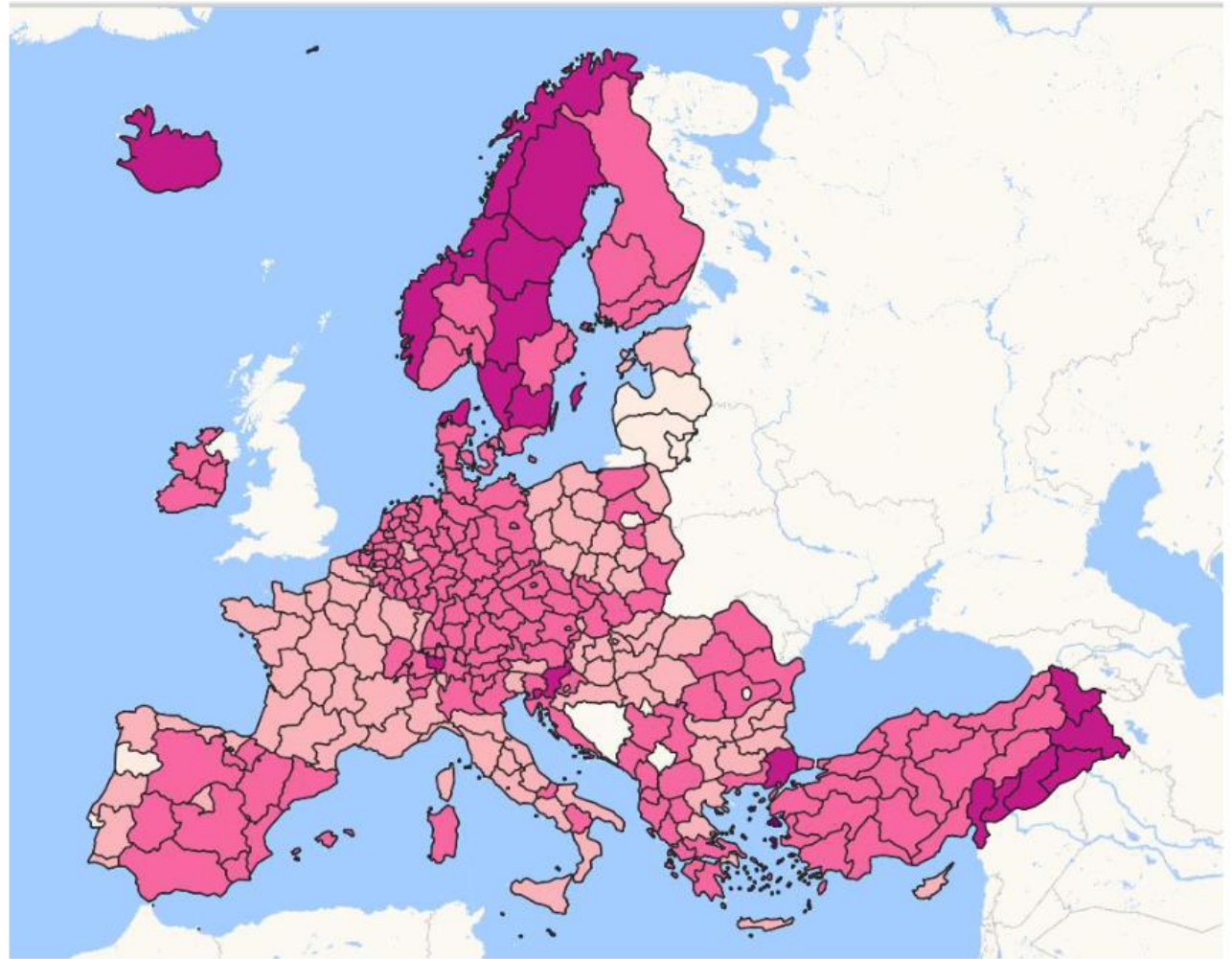
Laboratorio GIS

- Esplorare i dati presenti sul portale di Eurostat e aggregarli per regione NUTS
- Rappresentare le variabili su base spaziale



Scegliere ed elaborare i dati

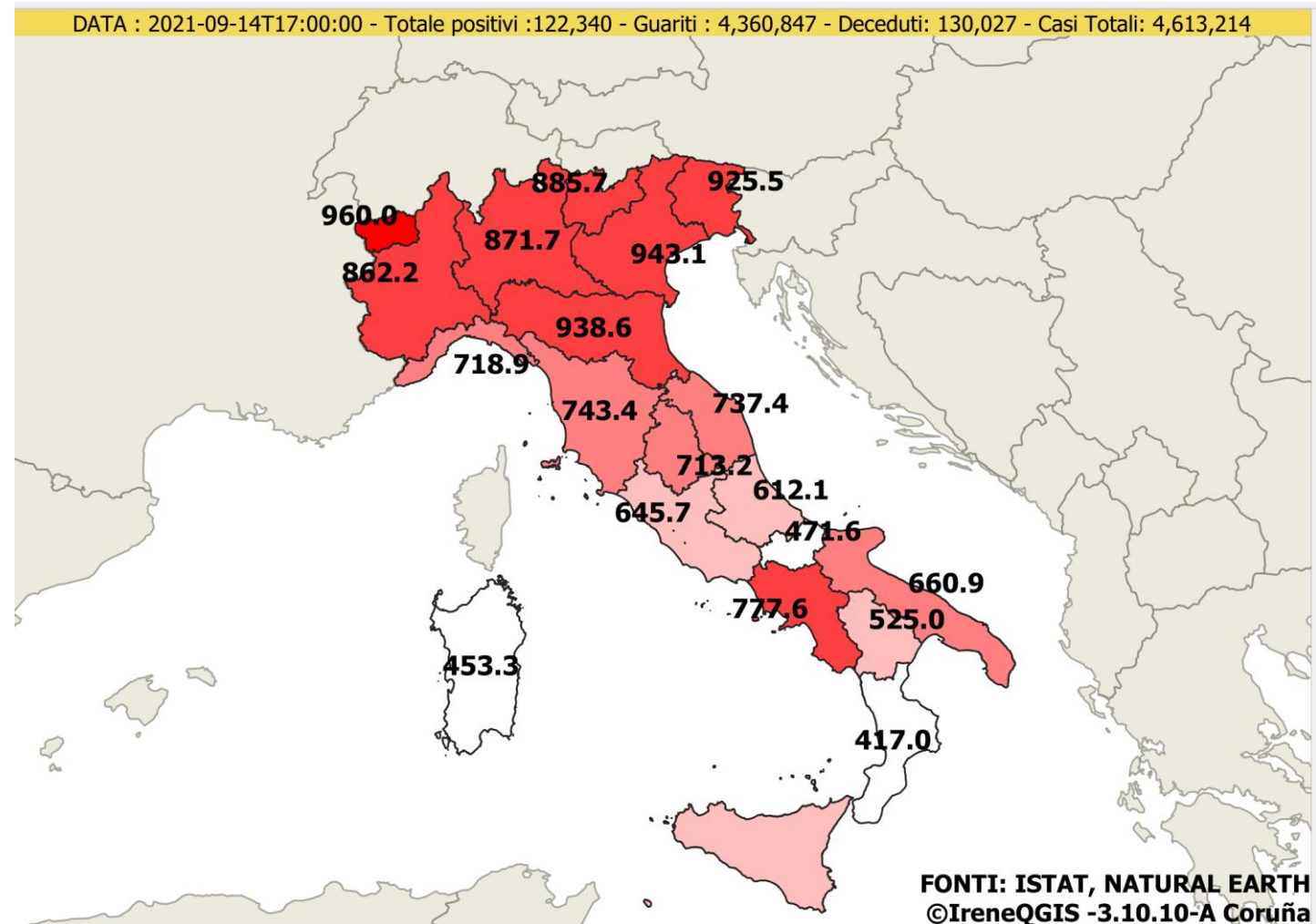
- Lavorare coi dati
 - Tabelle di attributi
 - Join
 - Analisi Spaziale



Sex ratio in Europa per micro-regione (Nuts 2)

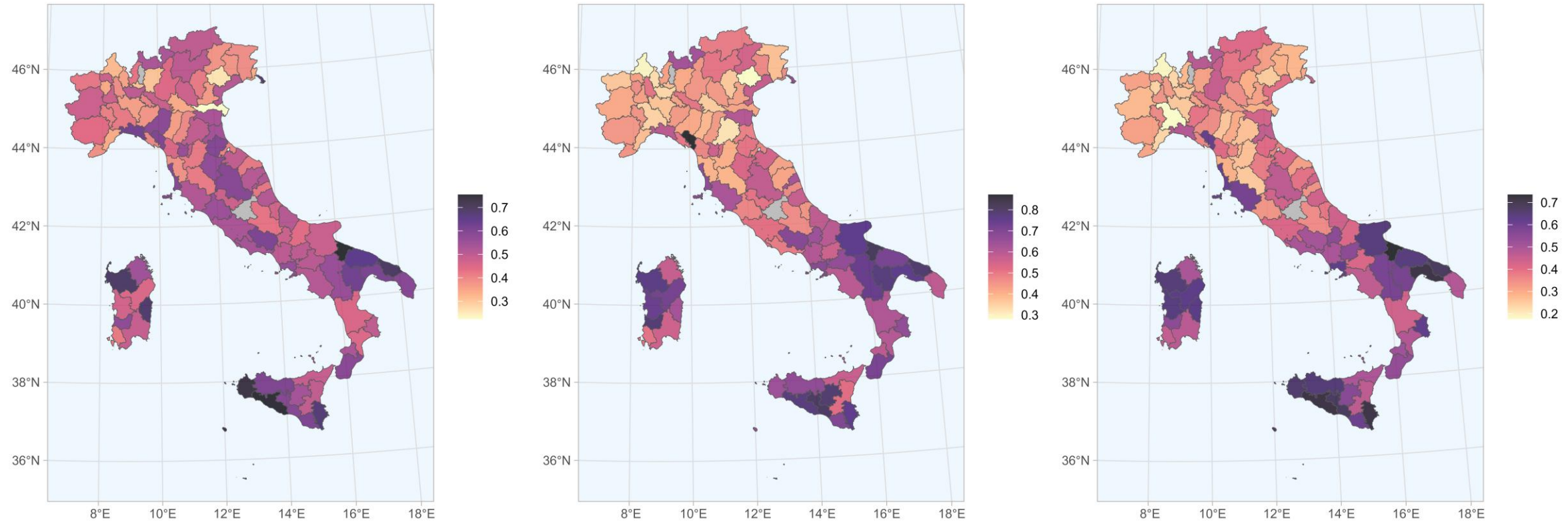
Analisi spaziale dei dati Covid in Italia

- Creare una mappa dinamica usando i dati della protezione civile aggiornati in tempo reale
 - Calcolatore di campi
 - Plugin Data Plotly



Casi totali per 10.000 abitanti, dati per regione

Modulo 2. Geografia, missing e fonti



Probabilità di vivere in un raggio di 15km dai figli (destra), fratelli (centro) e genitori (sinistra) nella popolazione italiana tra i 50 e 65 anni, stimate da un modello logistico a tre livelli: rispondente-parente (genitori, figli e fratelli), rispondente, e provincia. *ISTAT Famiglie e Soggetti Sociali 2016*.

! ATTENZIONE alla scala, ma anche ATTENZIONE ai missing, ATTENZIONE al link tra le fonti !

2.1 Record linkage

- La variabilità che abbiamo osservato tra le province italiane può riflettere differenze tra Nord e Sud Italia date da dati demografici (densità), economici (tassi di disoccupazione) o culturali (norme tradizionali verso la famiglia).

➤ Necessità di utilizzare diverse fonti

$$\sigma_{between}^2 = \frac{1}{N-1} \sum_{j=1}^N (\bar{y}_{.j} - \bar{y}_{..})^2$$

	Model 2		Model 4	
	Coef.	S.E.	Coef.	S.E.
Type of kin (ref. Children)				
Sibling	-0.44**	(0.13)	-0.43**	(0.12)
Parents	0.12	(0.23)	0.12	(0.22)
Population density per 100 km ²			-0.00	(0.01)
Traditional family norms			0.47	(0.36)
Unemployment rates			0.05**	(0.01)
Constant	-8.65	(6.08)	-9.23	(6.06)
<i>Random Parameters at individual level:</i>				
Intercept variance	4.76**	(0.22)	4.73**	(0.22)
<i>Random Parameters at province level:</i>				
Intercept variance	0.35**	(0.08)	0.17**	(0.06)
Slope variance for Sibling	0.31**	(0.08)	0.22**	(0.05)
Slope variance for Parents	0.15*	(0.07)	0.10	(0.05)
N. of anchor-kin dyads	20,911		20,911	
N. of anchor-respondents	6,403		6,403	
N. provinces	105		105	

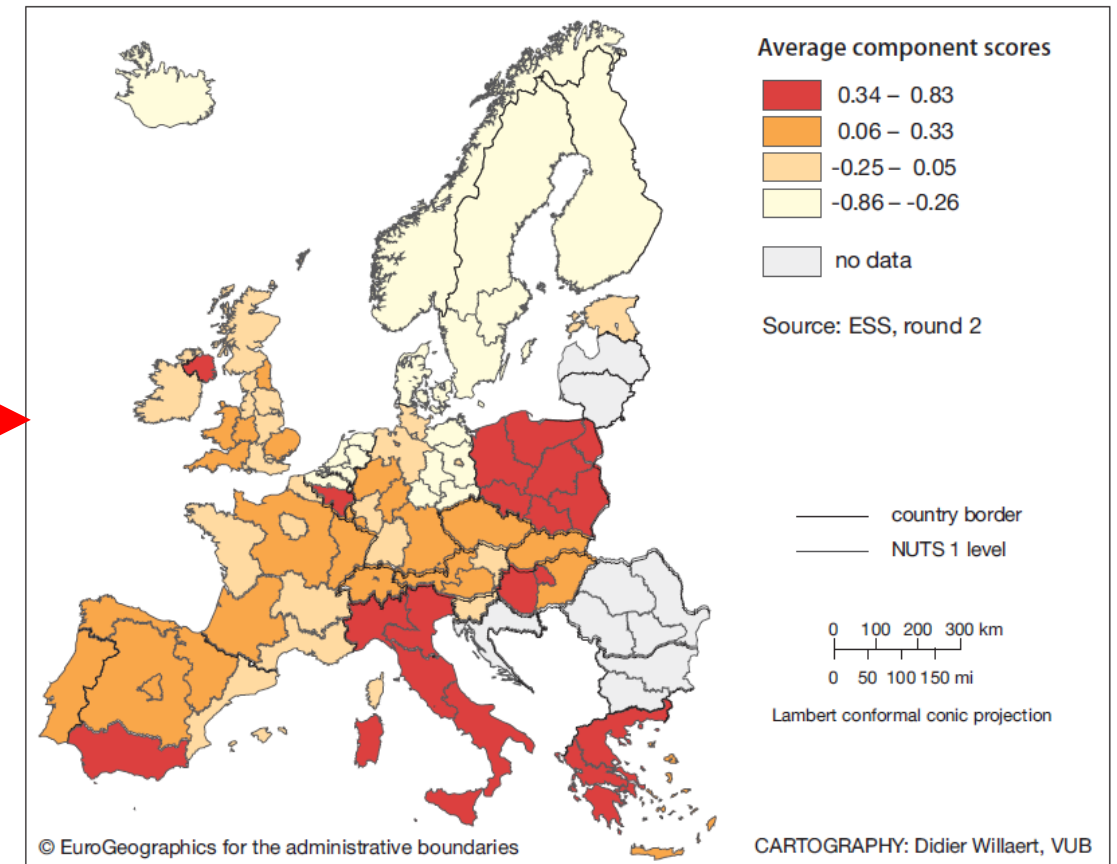
Record linkage: dati micro e macro

- Dati (e comportamenti) individuali abbinati a variabili regionali o nazionali (es, pil).

Table 1: Countries included in the analysis and number of mothers aged 20-54 with children 0-12 in the household

Country	N	Country	N
Austria	148	Ireland	189
Belgium	128	Iceland	64
Switzerland	144	Italy	46
Czech Republic	149	Luxembourg	90
Germany	166	Netherlands	142
Denmark	145	Norway	178
Estonia	109	Poland	106
Spain	64	Portugal	170
Finland	127	Sweden	130
France	147	Slovenia	93
Greece	107	Slovakia	77
Hungary	96		
		2815	

Map 1: Average scores for conservative family norms



Record Linkage deterministico

- Collegare informazioni da fonti diverse con strutture diverse

Anagrafe			
Nome	Nascita	Sesso	...
<u>Aluccio</u>	RE	1	
Ines	RE	2	
Renzo	RE	1	
Monica	PA	2	
Marco	BO	1	
Irene	BO	2	
Leo	BO	1	

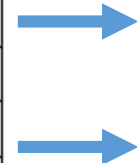
Survey su genitori-figli							
Nome	Nascita	Sesso	Padre	Madre	Figlio 1	Figlio 2	Figlio 3
Renzo	RE	1	<u>Aluccio</u>	Ines	Marco	Irene	
Monica	PA	2			Marco	Irene	Leo

Incidenza del Covid 19				
Città	Incidenza	Data 1	Data 2	Data 3
RE				
PA				
BO				

Record Linkage probabilistico

- Formulazione del problema: ID individuo cambia o è errato nei diversi database
- Spiegazioni logiche su queste variazioni

Anagrafe			
Nome	Nascita	Sesso	...
<u>Aluccio</u>	RE	1	
Ines	RE	2	
Renzo	RE	1	
Monica	PA	2	
Marco	BO	1	
Irene	BO	2	
Leo	BO	1	



Registro parrocchiale			
Nome	Nascita	Sesso	...
Lucio	RE	1	
Ines	RE	2	
Lorenzo	RE	1	
Monica	PA	2	
Marco	BO	1	
Irene	BO	2	
Leo	BO	1	



Survey			
Nome	Nascita	Sesso	...
<u>Missing</u>			Morto
Ines	<u>Missing</u>	<u>Missing</u>	94enne
Renzo	RE	1	
Monica	PA	2	
Marco	BO	1	
Irene	<u>Missing</u>	<u>Missing</u>	Vive all'estero
Leo	BO	1	

Record Linkage nelle serie Netflix

- *Indian matchmaker* programma in cui un intermediario cerca di accoppiare delle persone attraverso l'uso di cataloghi (cartacei)



Viene anche chiamato Assortative Mating, ossia la propensione umana ad accoppiarsi con persone simili (in termini probabilistici? Nella serie Netflix è a ^casaccio^)

2.2 Gestire le risposte mancanti

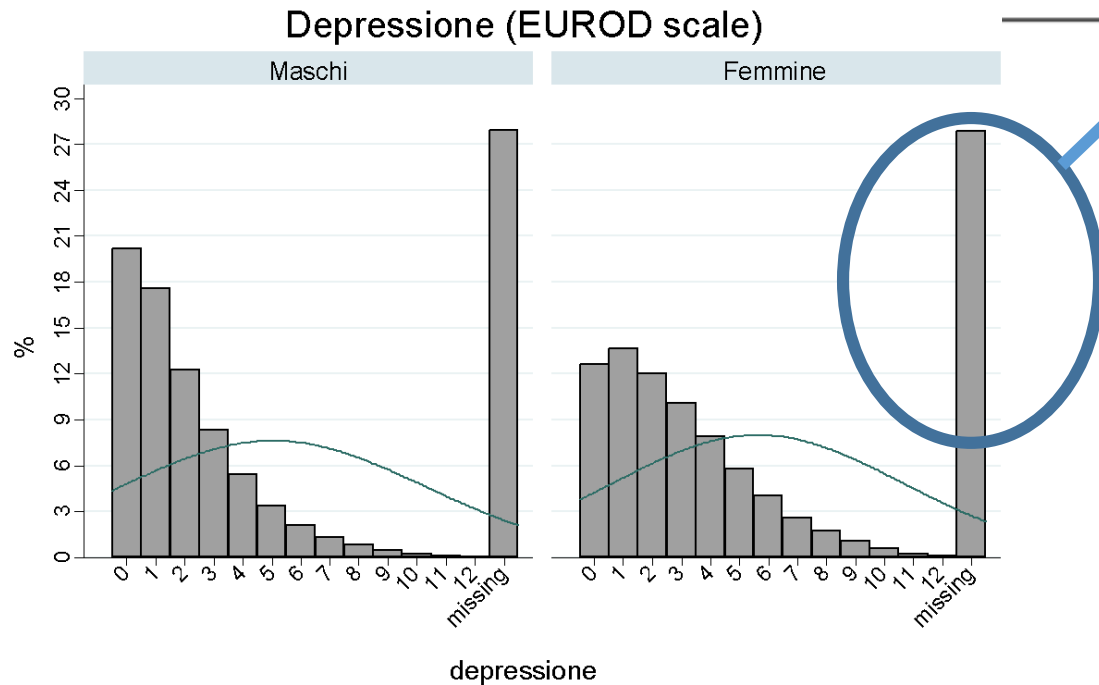
- Missing per costruzione dei dati
- Missing «veri»
 - Approcci tradizionali
 - Esclusione dei casi mancanti
 - Sostituire con la media (o media di gruppo)
 - Creare indicatori
 - Campionamento aleatorio
 - Approcci «nuovi»
 - Imputazione singola vs. imputazione multipla
 - Approcci per i dati longitudinali panel
 - Attrition vs. death

Risposte mancanti

I dati SHARE sulla salute

- Diversi tipi di dati mancanti:
Veri e Per costruzione

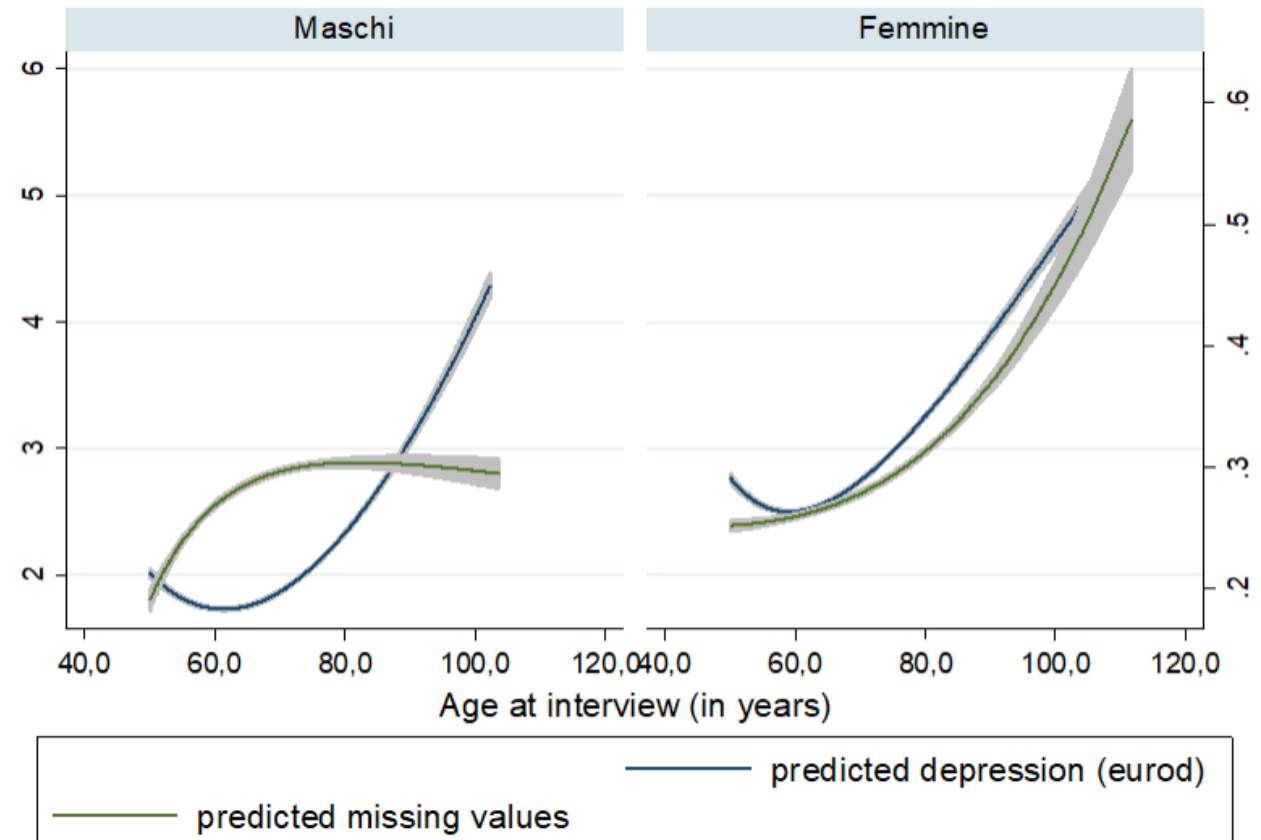
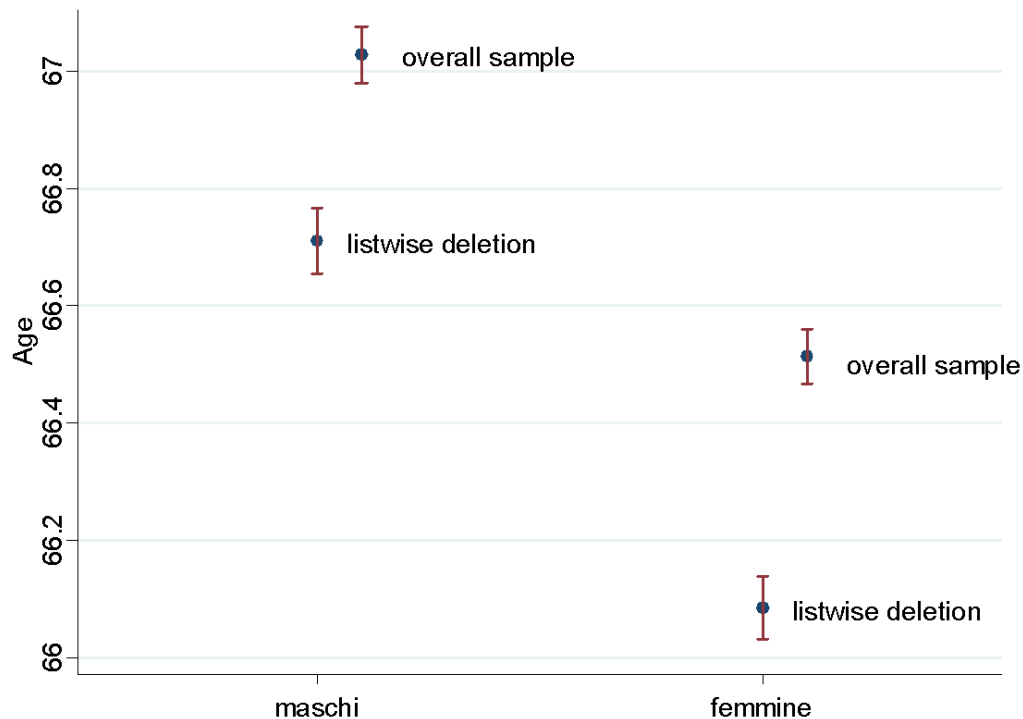
Depression scale EURO-D - high is depressed	Freq.	Percent	Cum.
-15. no information	10,212	10.01	10.01
-13. not asked in this wave	28,472	27.92	37.93
-10. SHARELIFE interview	63,304	62.07	100.00
Total	101,988	100.00	



- Possiamo ignorare i valori mancanti se assumiamo che siano *completamente Random*, ossia il meccanismo generativo non dipende da variabili osservate o non osservate.

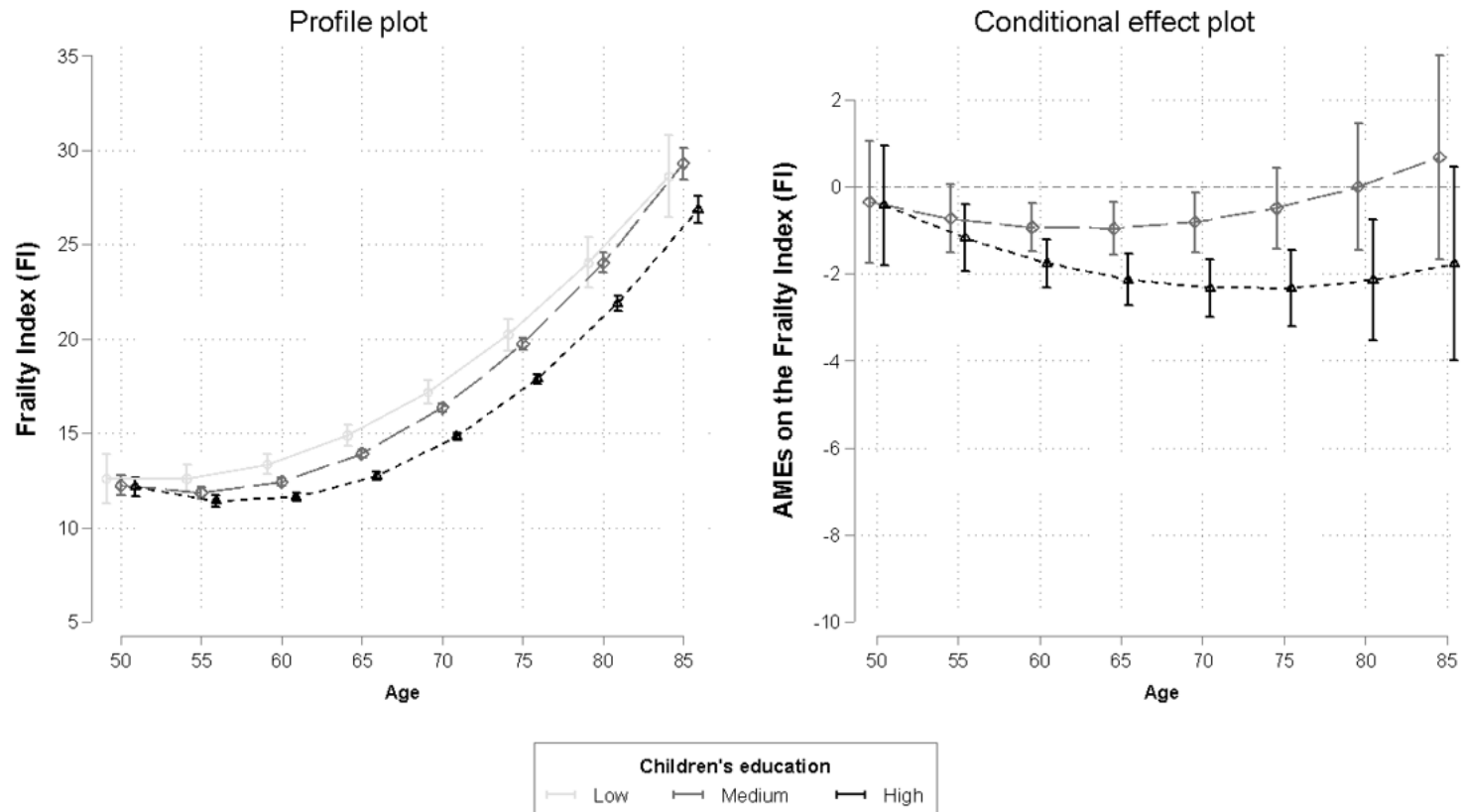
Risposte mancanti

- Tuttavia in molti casi le risposte mancanti non sono casualmente generate.
- Sia la depressione sia la probabilità di non rispondere aumentano con l'età -
Il meccanismo generativo dei missing è l'età?



Attrition o cadute

- Il problema delle non risposte si estende alle cadute nei dati longitudinali che introducono serie distorsioni negli studi sulla salute (ad esempio): effetti più piccoli ad età più elevate potrebbero essere dovuti alle stesse determinanti che influiscono sia sulla salute che su mortalità e sulle non-risposte (ospedalizzazione?).



un esempio di Record linkage, missing imputation & analisi territoriale:



Home

About Megan's Law

Summary of Megan's Law

Penalties for Misuse

Announcements

Additional Laws

Education & Prevention

About Sex Offenders

Search Offenders

AQ

Summary of Megan's Law

California's Megan's Law was enacted in 1996 [Penal Code § 290.46](#). It mandates the California Department of Justice (CA DOJ) to notify the public about specified registered sex offenders. Megan's Law also authorizes local law enforcement agencies to notify the public about sex offender registrants found to be posing a risk to public safety. Megan's Law is named after seven-year-old Megan Kanka, who was raped and killed by a known child molester who had moved across the street from the family without their knowledge. In the wake of that tragedy, the Kankas sought to have local communities warned about sex offenders in the area. All states in the U.S. now have some form of Megan's Law.

The California Sex and Arson Registry is the source of sex offender information displayed on this website. This database contains registration information provided by the offender to local law enforcement agencies. This website indicates that some of the registrants are currently in violation of their registration requirements. Any information you may have on these individuals should be reported to your local law enforcement agency.

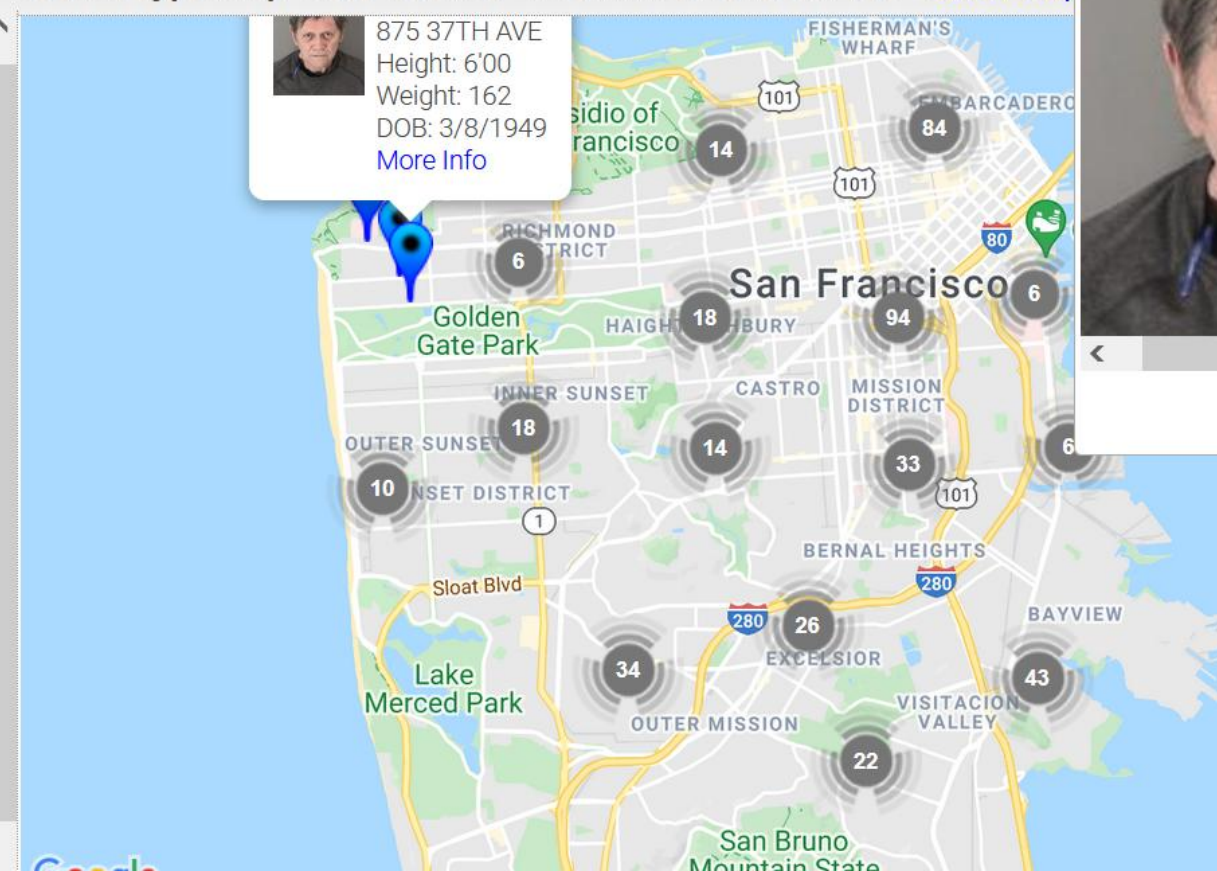
Record linkage e analisi territoriale

Sex offenders in California (Megan's Law)
<https://www.meganslaw.ca.gov/Search.aspx>

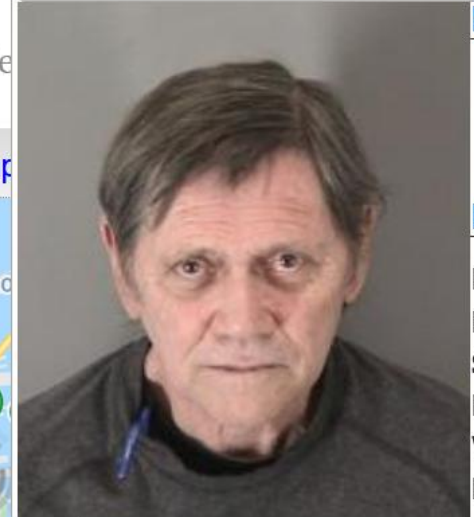
California Megan's Law Website

State of California Department of Justice Office of the Attorney General

Search Type: City: San Francisco Postable Offenders: 796 Show Map



Offender Profile



Known Aliases


- KNIGHT, MAX
- KNIGHT, CHA

Description

Date Of Birth: 3/8/1949
Sex: MALE
Height: 6'00
Weight: 162
Eye Color: BROWN

Record Linkage e analisi territoriale

- Record linkage può essere complicato se una persona ha più «Alias», cambia nome, o i middle-names vengono riportati diversamente.
- Record Linkage per ricostruire una storia di crimini sessuali dal 1979 al 2011

Offender Profile			
REILLY, KEVIN MICHAEL		Offense Code:	261.2
	Known Aliases <ul style="list-style-type: none">STEPHENS, KEVIN MICHAELRILEY, KEVINBUSKA, GARY JOHNREILLY, KEVIN MBUSKA, GARYREILLY, KEVINREILLY, MICHAELMICHAEL, REILLY KEVIN	Description:	PRIOR CODE: RAPE BY FORCE
		Year of Last Conviction:	1985
		Year of Last Release:	1991
		Offense Code:	288
Sexually Violent Predator	Description	Description:	PRIOR CODE: CRIMES AGAINST CHILDREN/LEWD OR LASCIVIOUS
		Year of Last Conviction:	1979
		Year of Last Release:	
		Offense Code:	289
		Description:	SEXUAL PENETRATION WITH FOREIGN OBJECT
		Year of Last Conviction:	1985
		Year of Last Release:	1991
		Offense Code:	626.81(a)
		Description:	SEX OFFENDER ENTERING SCHOOLS W/O LAWFUL BUSINESS (NOT REG OFFENSE)
		Year of Last Conviction:	2011
		Year of Last Release:	

MODULO 2. in sintesi

- Introduzione ai dati SHARE, comandi di base di STATA, e concetti di base del disegno della ricerca.
- ***Record linkage***
 - Deterministico vs. probabilistico
- ***Gestione dei dati mancanti***
 - Metodi di Imputazione
- Applicazioni pratiche (Lab. ASID 17)
 - Utilizzo della base-dati SHARE (salute degli over-50)
 - Fonti esterne: Ocse, Google Trends, Eurostat, COVID-19 Government Response Tracker (Oxford).. etc..
 - Utilizzo del Software STATA 13
- Materiale fornito su Moodle
- Prova finale
 - L'ultimo giorno di lezione 17 Dicembre, oppure alla data di appello

Materiali per il corso

- Pagina Moodle
 - Link utili
 - Slides delle lezioni

DATI MULTI-FONTE E ANALISI TERRITORIALI 2025-2026 -
SCQ0093539

Modifica in massa 

Corso

Impostazioni

Partecipanti

Valutazioni

Report


Altro 




DATI MULTI-FONTE E ANALISI TERRITORIALI 2025-2026 - PROF. MARCO TOSI 

Minimizza tutto



Annunci 



Pagina dell'offerta Formativa 



Prova Finale

- **Modulo 1-** Esercizi in Laboratorio (max 20 punti)
 - *Abbinamento (linkage) deterministico vs. probabilistico:* riuscire ad abbinare 2 data-set di grandi dimensioni.
 - *Gestione delle risposte mancanti:* imputare le risposte mancanti nel miglior modo possibile (relativamente ad un specifico obiettivo).
- **Modulo 2-** Produrre un cartogramma tematico (max 11 punti) da consegnare una settimana prima della data d'esame
 - *Mappa della popolazione residente nelle diverse province e dei dati covid per provincia (totale casi positivi per 10.000 abitanti) di una regione italiana a scelta.*