

Per la serie di Webinar
*L'importanza dei dati e delle metodologie
statistiche di ricerca: l'emergenza Covid-19*

Strategie di campionamento per il Covid-19

Giuseppe Arbia, Piero Demetrio Falorsi, Vincenzo Nardelli

In collaborazione con Giorgio Alleva ed Alberto Zuliani



Sommario

- Centralità dei dati per monitoraggio e controllo della pandemia
- Incompletezza e inaccuratezza dei dati ufficiali
- Necessità di intervento per migliorare le stime dei parametri della pandemia (Stime indirette e stime campionarie)
- Due proposte di campionamento per dati epidemici

Centralità dei dati

- Il **rito quotidiano** della diffusione dei dati crea confusione fra i cittadini.
- L'insufficienza e talvolta l'ambiguità dei dati alimentano la **distanza delle posizioni** fra i tecnici e nel confronto politico e istituzionale.
- Mancanza di un **disegno organico** che fornisca un **contesto informativo** coerente a sostegno delle decisioni e della consapevolezza dei cittadini.
- Necessità di rilascio di **dati individuali resi anonimi alla comunità scientifica** (le resistenze nascondono un problema di qualità?).
- Promuovere **interdisciplinarietà** e **collaborazione istituzionale**.

Strategie di monitoraggio attualmente in corso

Quattro strategie (Zuliani e Alleva, 2020):

- **Campagne massicce** (indipendentemente dalla presenza di sintomi) effettuate senza seguire un disegno formale di campionamento per identificare soggetti infetti ma senza sintomi o solo lievi sintomi: Corea del Sud e Hong Kong, nonché degli Emirati Arabi Uniti, Australia, Islanda, Veneto e Bolzano
- **Test diagnostici attraverso un campione probabilistico.** Tutte queste indagini campionarie sono trasversali e utili per misurare il tasso di infezione in un istante preciso: Helmholtz Center for Research on Infections in Germany (100.000 individui), campione Bucarest (10.500 persone), Italia, (65.000 intervistati).
- Specifica **indagine web** massiccia raccolta da individui e famiglie che decidono di partecipare su base volontaria. Israelis Weizmann Institute (60.000), Islanda, Estonia e altri paesi.
- **Indagini campionarie pre-esistenti** modificate parzialmente per raccogliere informazioni sull'epidemia. Proposto da Daniel Gros (2020).

Occorre un sistema di monitoraggio

Integrare informazioni sanitarie, raccolte in modo rigoroso, con altre campionarie.

- Campione ripetuto nel corso del tempo per **stimare la prevalenza** e seguirne l'evoluzione ed i **cambiamenti di stato** dei contagiati.
- Monitoraggio costante della situazione nelle **scuole** tra i **di medici di base** e nelle **residenze sanitarie assistite**.
- Valutazione delle **conseguenze di Covid-19** sulle condizioni di salute delle **persone** dopo la guarigione.
- Impiego delle **tecnologie dell'informazione** e dei *Big Data* per stimare gli effetti delle restrizioni adottate.
- Individuazione, attraverso sistemi di tracciamento, dei **luoghi e dei contesti nei quali la trasmissione del SARS-CoV-2 è più frequente**.
- Copertura informativa del **piano vaccinale**, non limitata al solo numero dei vaccinati.

Soggetti interessati

- L'**Invalsi** dovrebbe condurre un progetto specifico per la valutazione dell'effetto differenziale della didattica a distanza rispetto a quella in presenza sul rendimento scolastico degli alunni.
- L'**Istat** potrebbe aggiungere stabilmente alcune domande nei questionari di rilevazione periodica sulle imprese e sulle forze di lavoro relativamente a presenza e intensità dello *smart working*. Altre domande potrebbero riguardare l'adozione delle norme igieniche essenziali e il rispetto delle misure di contenimento via via introdotte da parte delle persone.
- In effetti l'**Istat** dovrebbe svolgere un **ruolo centrale** nella progettazione, realizzazione, analisi e validazione del quadro informativo necessario.
- Il sistema dovrebbe essere in grado di **attrezzarsi rapidamente** per produrre stime affidabili **su questioni rilevanti che emergessero** via via e per future possibili analoghe emergenze.

La nostra proposta: *Alleva et al. (2021) Journal of Official Statistics*

Caratteristiche del piano di campionamento:

- Flessibilità
- Affidabilità statistica
- Minimizzazione dello sforzo di raccolta
- Ripetibilità nel tempo

Obiettivi:

Produrre stime affidabili dei contagiati per l'intera popolazione, in modo da poter poi stimare correttamente **la diffusione del virus su base territoriale, il tasso di contagio, la mortalità e il tasso di letalità.**

Proposta di campionamento (Alleva *et al.*, 2021a, JOS)

Variabile target $y_j = 1$ persona infetta, $y_j = 0$ persona sana

Variabile ausiliaria $v_j = 1$ infetto verificato (es. quarantena o ospedale)

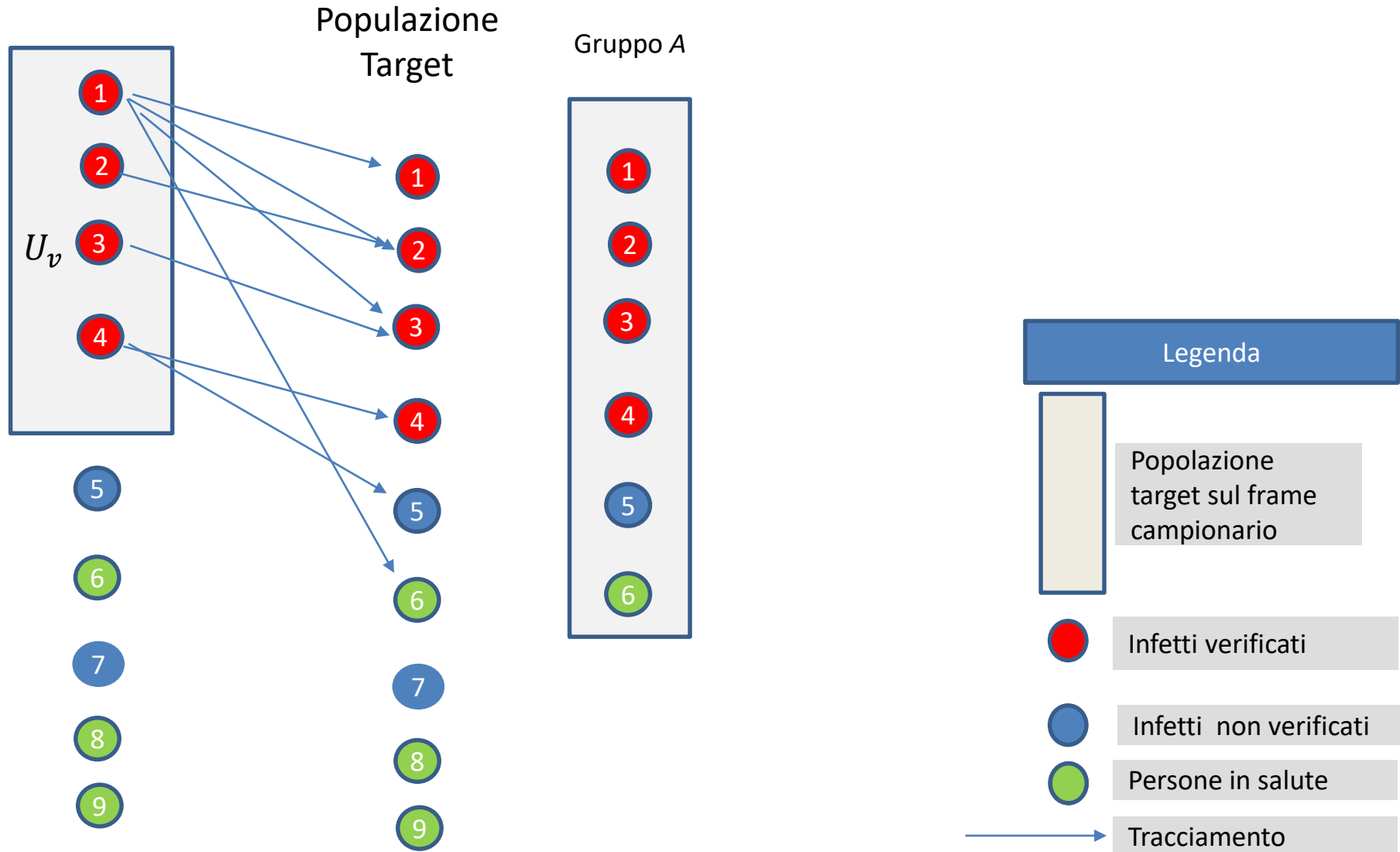
- **Gruppo A:** individui per i quali è stato accertato uno stato di infezione ($v_j = 1$) e persone che hanno avuto contatti con loro nei giorni precedenti.

Concentrarsi sull'osservazione dei contatti massimizza il numero di individui infetti osservati nel campione.

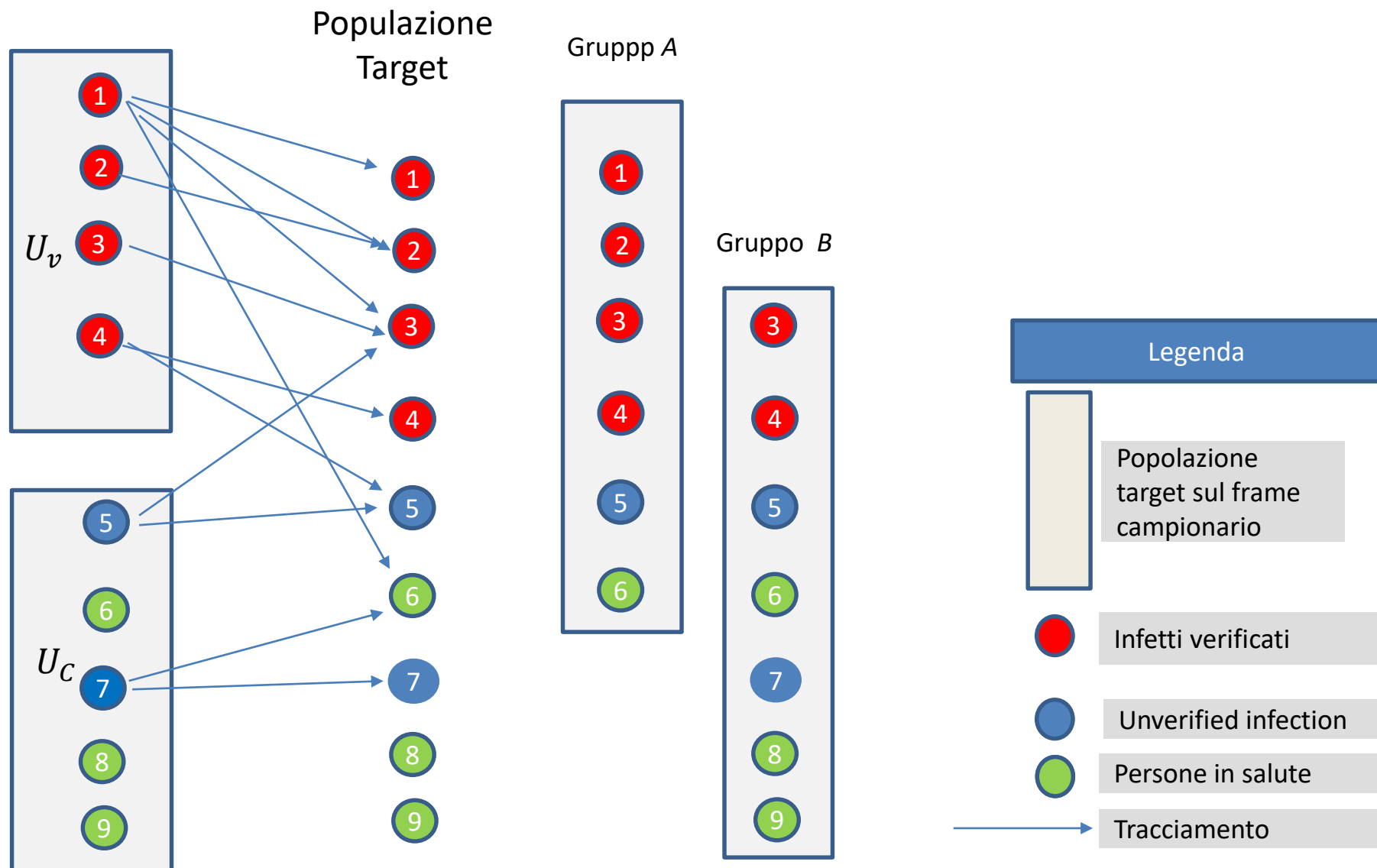
- **Gruppo B:** persone sane ($v_j = 0$), persone infette ma non consapevoli di esserlo ($v_j = 0$) e persone che hanno avuto contatti con loro.

Il Gruppo B consente di produrre stime affidabili per l'intera popolazione, in modo da poter poi stimare correttamente il tasso di contagio e il tasso di letalità.

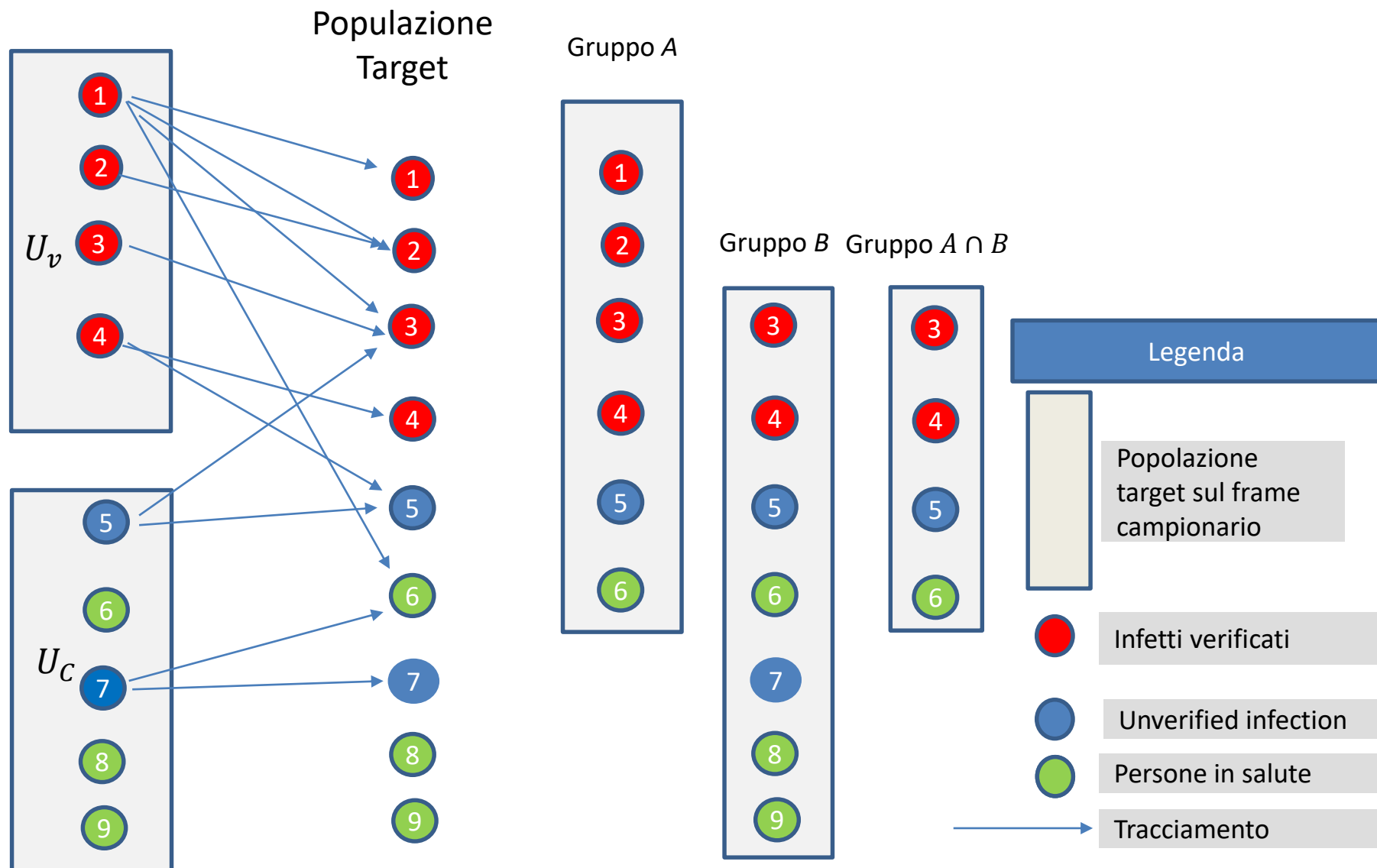
Metodologia di campionamento



Metodologia di campionamento



Metodologia di campionamento



Numero totale di infetti e la suddivisione

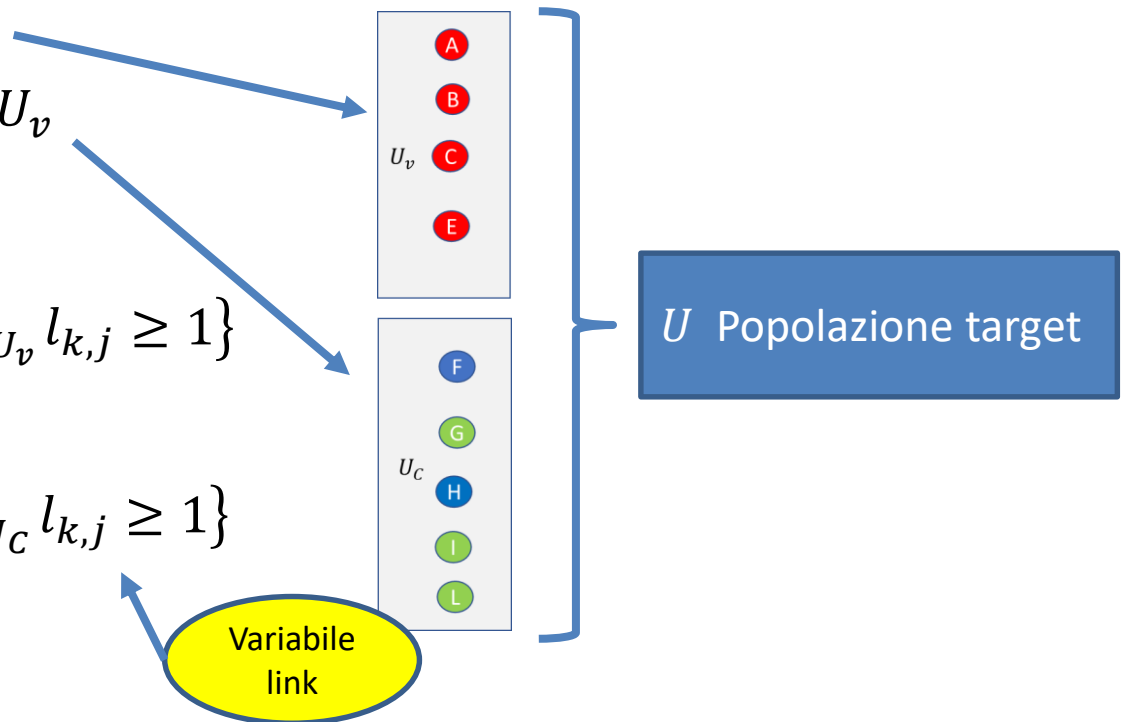
$$\sum_{j \in U} y_j = Y = Y_A + Y_B - Y_{AB}$$

U_v Infetti verificati

$U_c = U \setminus U_v$ Complemento di U_v

Gruppo A : $U_A = \{j \in U : \sum_{k \in U_v} l_{k,j} \geq 1\}$

Gruppo B : $U_B = \{j \in U : \sum_{k \in U_c} l_{k,j} \geq 1\}$



Campionamento per il Gruppo A

Ci sono due opzioni per selezionare S_v in base alla struttura dei dataset di U_v :

Opzione 1. I dati di U_v sono disponibili in un unico set di dati **centralizzato**

Proposta: Campionamento bilanciato in modo che il campione selezionato riproduca i totali noti di alcune variabili ausiliarie (ad esempio, distribuzione per sesso ed età, posizione lavorativa, distribuzione geografica, ecc.)

$$\sum_{k \in S_v} \frac{\mathbf{x}_k}{\pi_{vk}} = \sum_{k \in U_v} \mathbf{x}_k$$

Opzione 2. I dati di U_v sono disponibili solo ad un **livello decentralizzato** in modo che ogni istituto sanitario abbia la propria lista.

Proposta: two stage sampling epsem in domains:

- 1- PPS balanced sampling delle istituzioni sanitarie
- 2- SRSWOR di infetti (o sistematici in base al momento dell'accesso all'istituzione)

Campionamento per il Gruppo B

1. Per prima cosa, un campione senza reimmissione S_C è selezionato da U_C , dove le probabilità di inclusione sono π_{Ck}
2. Le persone nel panel es prendono regolarmente un test diagnostico. (ad esempio 1/4 del campione ogni settimana)
3. Se il membro k del pannello ha un risultato positivo del test (i.e., $y_k = 1$), tutti i suoi contatti U_k vengono tracciati.
4. Se $y_k = 1$, un campione S_{Ck} è selezionato da U_k senza reimmissione e con uguale probabilità di inclusione $\pi_{2C|k}$.
5. $S_B = S_C \cup_{k=1:y_k=1}^{\#S_C} S_{Ck}$ è formato con un meccanismo di **campionamento indiretto**, che comprende persone provenienti da S_C (persone per le quali i loro stati di infezione non sono noti) e $\cup_{k=1:y_k=1}^{\#S_C} S_{Ck}$

Stima

$$\hat{Y} = \hat{Y}_A + \hat{Y}_B - \hat{Y}_{AB}$$

$$\hat{Y}_{AB} = \alpha \hat{Y}_{AB}^A + (1 - \alpha) \hat{Y}_{AB}^B,$$

$$\alpha^* = \frac{V(\hat{Y}_B)}{V(\hat{Y}_A) + V(\hat{Y}_B)} \quad (\text{Hartley, 1974})$$

$$\hat{Y}_A = \sum_{k \in S_v} \frac{1}{\pi_{vk}} \sum_{j \in S_{vk}} \frac{1}{\pi_{2v|k}} \frac{1}{L_{vj}} l_{k,j} y_j$$

$$\hat{Y}_B = \sum_{k \in S_c} \frac{1}{\pi_{ck}} y_k \sum_{j \in S_{ck}} \frac{1}{\pi_{2c|k}} \frac{1}{L_{cj}} l_{k,j} y_j$$

$$\hat{Y}_{AB}^A = \sum_{k \in S_v} \frac{1}{\pi_{vk}} \sum_{j \in S_{vk}} \frac{1}{\pi_{2v|k}} \frac{1}{L_{vj}} l_{k,j} y_j \mathbb{I}(L_{cj} \geq 1)$$

$$\hat{Y}_{AB}^B = \sum_{k \in S_c} \frac{1}{\pi_{ck}} \sum_{j \in S_{ck}} \frac{1}{\pi_{2c|k}} \frac{1}{L_{cj}} l_{k,j} y_j \mathbb{I}(L_{vj} \geq 1)$$

GWSM estimators

Proprietà formali

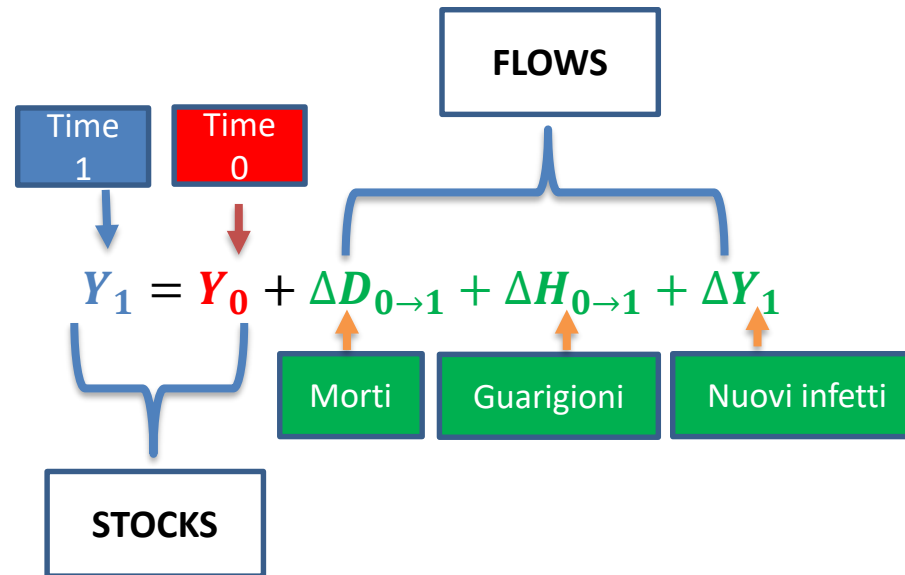
In Alleva *et al.* (2021) è stato provato formalmente che il metodo proposto garantisce stimatori:

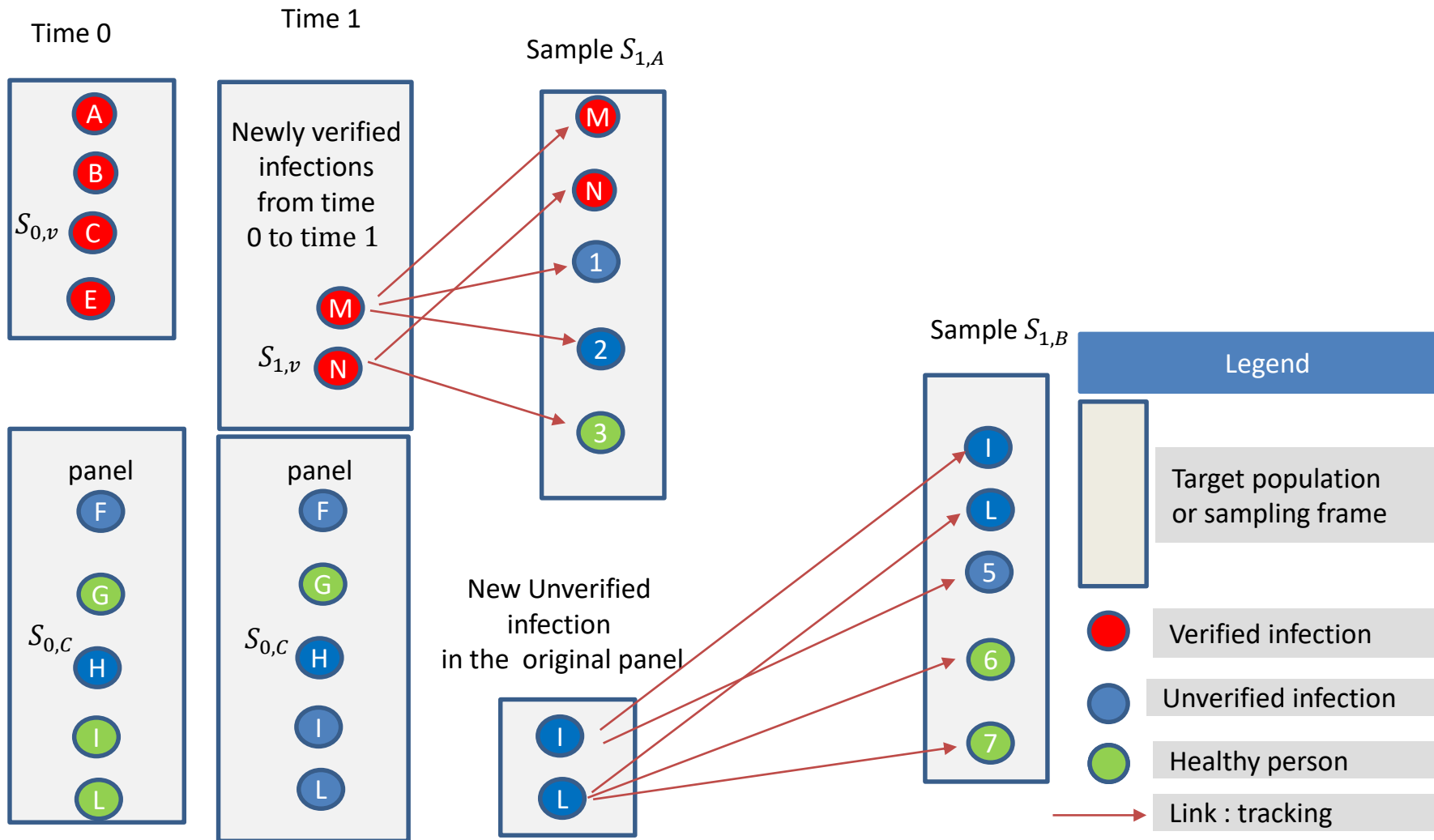
non distorti → in media fornisce il risultato corretto

più efficienti rispetto al campionamento casuale semplice → fornisce la stessa precisione con una dimensione campionaria inferiore

Sistema di monitoraggio e follow-up dell'indagine

FLows EQUATION

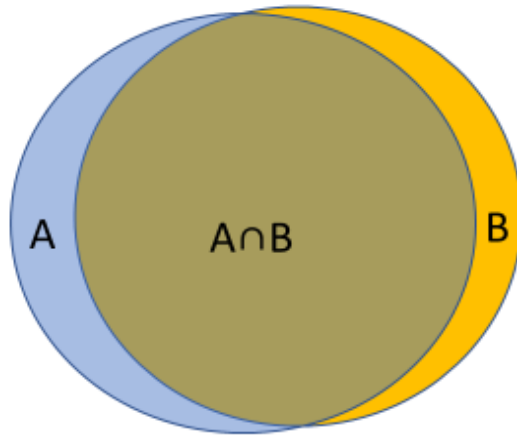




Efficienza della strategia proposta

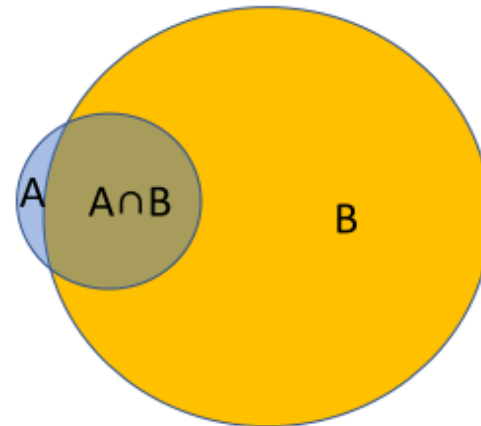
CASE 1

Groups A and B have the same sizes with a strong intersection between the two groups



CASE 2

Group A is much smaller than Group B with a strong intersection between the two groups



1. L'efficacia della strategia è al massimo nel caso 1.
2. L'efficacia è al massimo quando si campiona da U_v

Una buona strategia potrebbe essere di sovracampionare da U_v e ottenere un piccolo campione da U_c

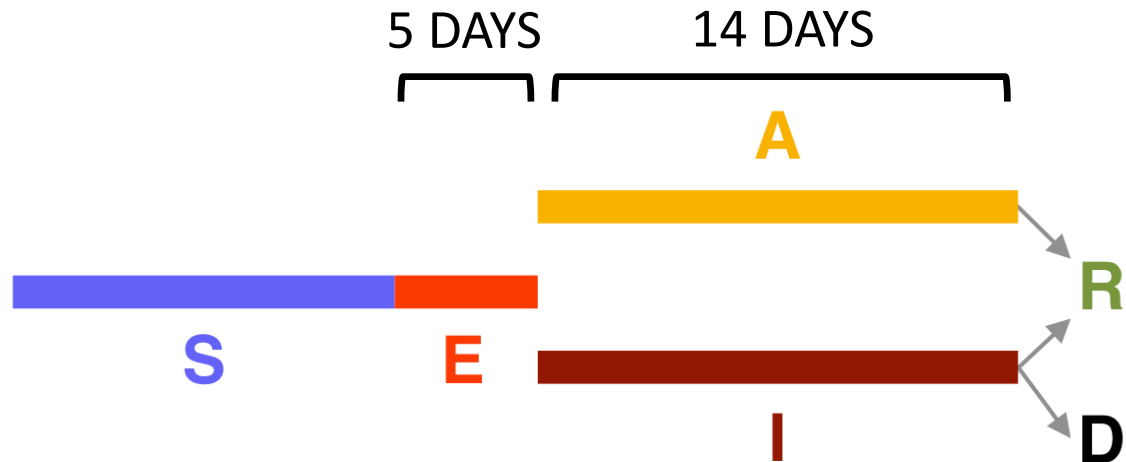
Generazione dei dati

Abbiamo simulato una popolazione artificiale considerando un modello costituito da un sistema di sei equazioni differenziali che, in ogni momento, descrivono sei categorie di individui, vale a dire **persone suscettibili (S)**, quelle **esposte al virus (E)**, quelle **infettate con sintomi (I)**, quelle **asintomatiche (A)** e quelle che vengono rimosse dalla popolazione perché **guarite (R)** o sono **morte (D)**.

Questo schema di modellazione è il risultato del contributo seminale di Hamer (1906), Kermack e McKendrick (1927) e Soper (1929), ed è spesso indicato come il "modello SIR"



Generazione dei dati



Probabilità

$S \rightarrow E$???

Dipende dalla
mobilità e
dall'**interazione
sociale**!

$E \rightarrow A$ 0.75

$E \rightarrow I$ 0.25

$A \rightarrow R$ 1.00

$I \rightarrow R$ 0.85

$I \rightarrow D$ 0.15

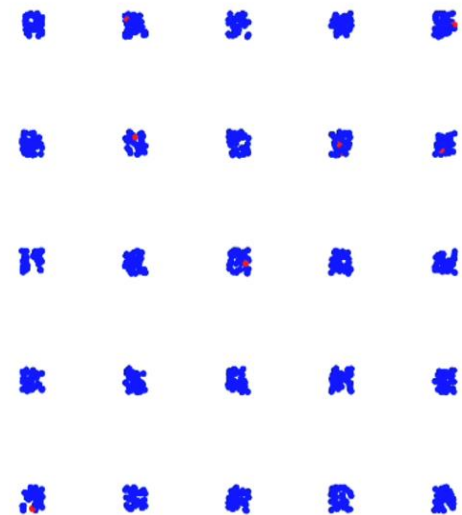
Generazione dei dati

Generazione delle mappa

Per il processo di generazione delle mappe, abbiamo considerato una popolazione distribuita su 25 unità spaziali posate su una normale griglia reticolare quadrata 5 per 5.

Ogni quadrato della griglia contiene un certo numero di individui disegnati casualmente da una distribuzione uniforme compresa tra 800 e 1.000. Dopo aver eseguito un esercizio di simulazione con questi valori di parametro, abbiamo ottenuto una popolazione artificiale con un totale di 22.217 individui.

Questa rappresentazione geografica è molto generale in quanto la mappa generata in questo modo può rappresentare, ad esempio, una città divisa in blocchi, una regione divisa in unioni spaziali più piccole o qualsiasi altra partizione geografica significativa.

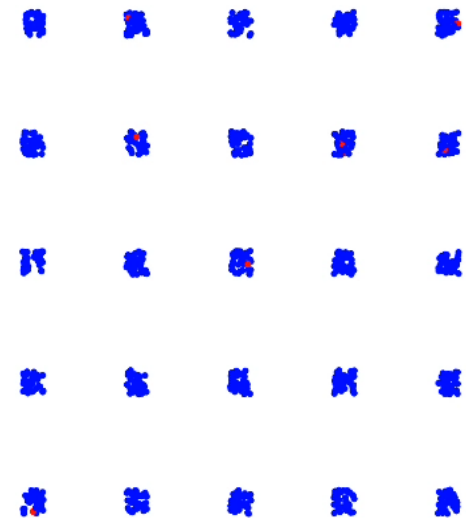


Generazione dei dati

Mobilità

Il meccanismo di contagio è favorito dalla mobilità delle persone. In questa simulazione, abbiamo supposto che in qualsiasi momento una certa percentuale m della popolazione possa spostarsi tra i quadrati. In questo modo è possibile distinguere diverse fasi epidemiche come il periodo di free-to-move e il lockdown.

Il pendolarismo durante il periodo di blocco non è limitato solo dal numero di persone che si spostano, ma anche dall'entità dei loro movimenti. Questo è un ulteriore parametro di simulazione che è generato da una distribuzione uniforme che va da $-s$ a s .



Generazione dei dati

Interazioni sociali

Il contagio è determinato dall'interazione sociale e dalle opportunità di contatto in ogni quadrato della griglia.

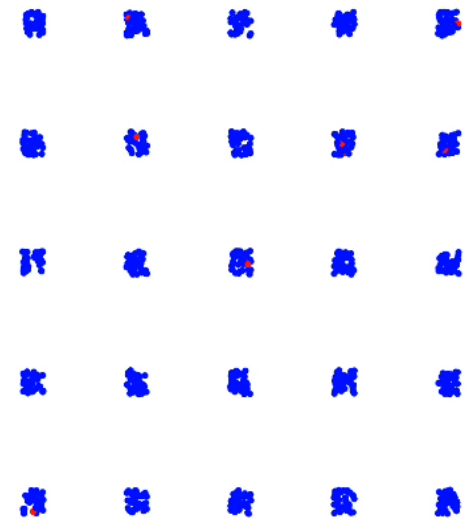
Assumiamo una serie di contatti c_i

$$c_i \sim \text{Pois}(\mathbf{cn})$$

in cui sono coinvolte c_{ip} persone

$$c_{ip} \sim \text{Pois}(\mathbf{cp})$$

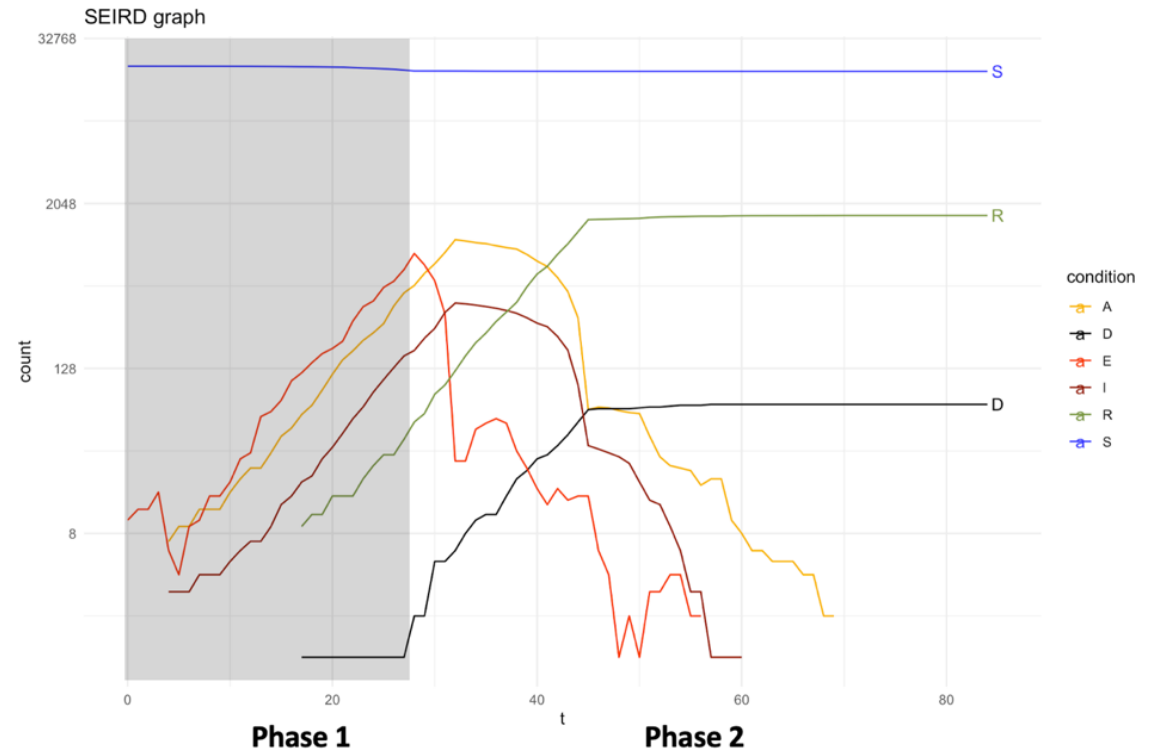
Se in una riunione è presente almeno una persona asintomatica o esposta, gli im suscettibili saranno infettati muovendosi nello stato dell'esposto.



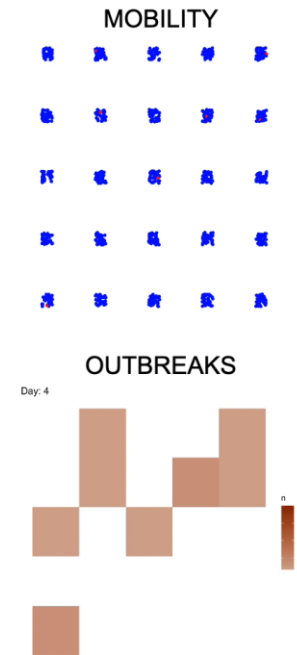
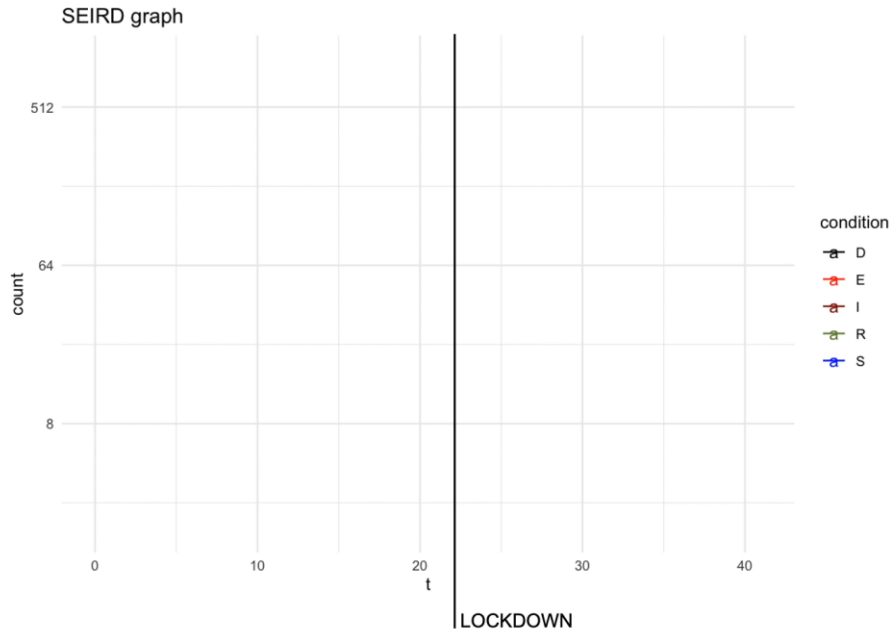
Generazione dei dati

Simulazione delle politiche di lockdown

	Parameter	Phase 1 4 weeks	Phase 2 8 weeks
Mobility	m	3%	1%
	s	4	1
Social interaction	cn	20	3
	cp	5	3
	im	3	2



Generazione dei dati



R package

epidsampler

Generare un dataset personalizzato

Sviluppo in corso (BETA)

Presto disponibile sul CRAN

Download, documentation and examples
<https://vincnardelli.github.io/epidsampler/>



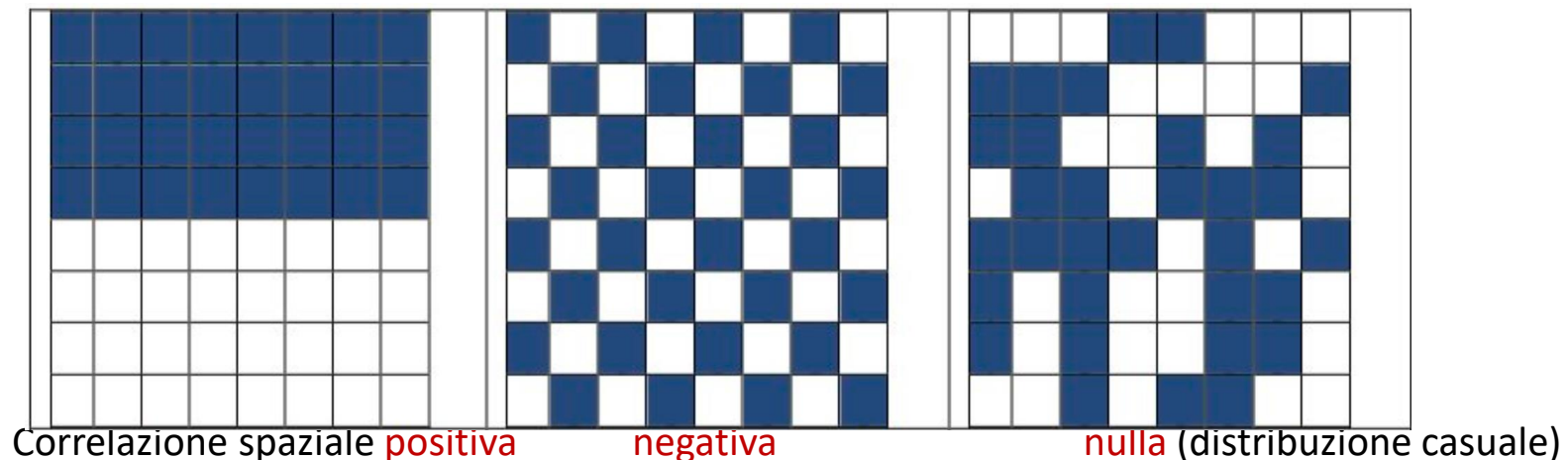
Risultati della simulazione: distorsione ed efficienza

Giorni da inizio epidemia	Percentuale individui infetti nella popolazione	Distorsione relativa della stima	Numerosità campione con tracciamento equivalente a campione tradizionale con 1000 testati
15	0.58%	0.0019	72
25	3.94%	0.0000	61
35	6.54%	0.0045	190

Uno sviluppo ulteriore (Allewa et al. 2021b, sottomesso)

Correlazione spaziale: La tendenza dei dati ad assumere valori simile se prossimi geograficamente.

Prima legge della geografia: «*Tutto è correlato con tutto ma le cose vicine sono più correlate delle cose lontane*» Waldo Tobler, 1970



I dati epidemici presentano la caratteristica intrinseca di essere positivamente **spazialmente correlati** (presenza di focolai e cluster)-

Uno sviluppo ulteriore (Alleva et al. 2021b, sottomesso)

La presenza di correlazione spaziale positiva consente di congegnare piani di campionamento maggiormente efficienti:

- Si possono economizzare unità campionarie perché unità tra loro prossime geograficamente hanno un contenuto informativo simile.

Campionamento spaziale: processo attraverso il quale le osservazioni vengono raccolte in una superficie bi- (o tri-) dimensionale con un'attenzione specifica alla loro localizzazione (Arbia, 1993; Mueller, 2007; Tillè, 2020).

Campionamenti presi in considerazione

Local cube (inf)	Campionamento bilanciato sugli infetti verificati
Local cube (inf-sp)	Campionamento bilanciato sugli infetti e sulla distribuzione geografica delle UP
Local cube (sp)	Campionamento bilanciato sulla distribuzione geografica delle UP
Local pivotal	Campionamento spaziale semplice
Prop sampling	Campionamento semplice a probabilità variabile

Struttura campionaria

- M unità primarie (UP) di campionamento (es. comuni, ASL) $U_1, \dots, U_i, \dots, U_M$. L'UP U_i ha N_i persone
- $\mathcal{V}_i = \sum_{j=1}^{N_i} v_{ij}$ **numero di infetti conosciuti** nell'UP i
- Selezione di m UP con probabilità $\pi_{Ii} = m \frac{N_i}{N}$
- Selezione SRS di $\bar{n}$ persone $\pi_{IIi} = \frac{\bar{n}}{N_i}$
- Campione autoponderante

$$\pi_{ij} = m \frac{N_i}{N} \frac{\bar{n}}{N_i} = m \frac{\bar{n}}{M}.$$

- Stima HT

$$\hat{Y} = \sum_{i=1}^m \sum_{j=1}^{\bar{n}} y_{ij} \frac{1}{\pi_{ij}}$$

Tecniche di campionamento prese in considerazione

Modello di superpopolazione

$$y_{ij} = m(\mathbf{x}_{ij}; \boldsymbol{\beta}) + u_{ij}$$

$$E_M(y_{ij}) = m(\mathbf{x}_{ij}; \boldsymbol{\beta}), V_M(y_{ij}) = \sigma_u^2$$

$$\text{Cov}_M(y_{ij}, y_{\ell k}) = \sigma_u^2 \rho_{ij, \ell k}$$

Varianza anticipata

$$E_P E_M(\hat{Y} - Y)^2 = \sum_{i=1}^M \frac{1}{\pi_{Ii}} \sigma_i^2 - F$$

$$\sigma_i^2 = \underbrace{\eta_i^2 + \sigma_{\tilde{y}_i}^2 N_i \left(\frac{N_i - \bar{n}}{\bar{n}} \right)}_{\text{Campionamento bilanciato azzera questa componente}} + N_i^2 \frac{\sigma_u^2}{\bar{n}} + \rho_{i,i} \sigma_u^2 + \sum_{\ell \neq i}^M \frac{\pi_{Ii, I\ell}}{\pi_{I\ell}} \sigma_u^2 \rho_{i, \ell}$$

Campionamento **spaziale**
azzerà questa
componente

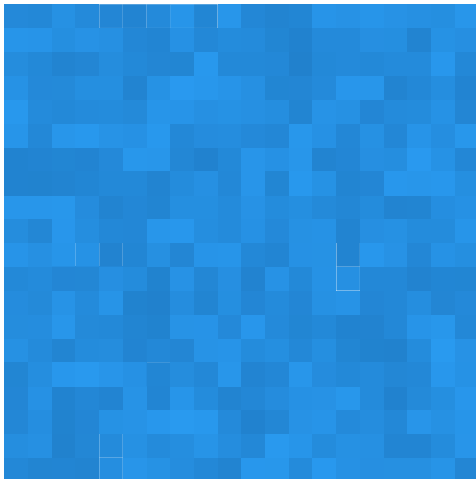


Campionamento
bilanciato azzera questa
componente

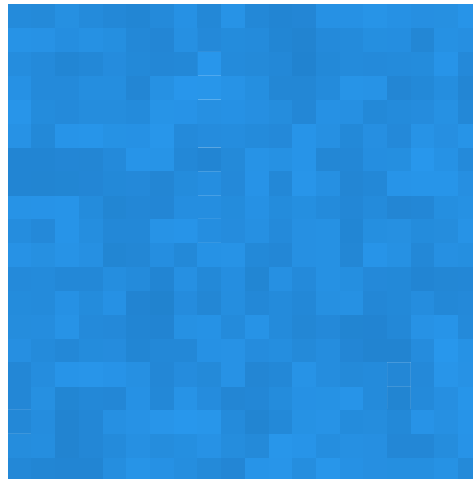
Generazione dei dati

Generazione delle mappa con dipendenza spaziale

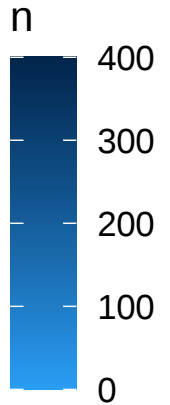
Rho= 0.3



Rho= 0.5



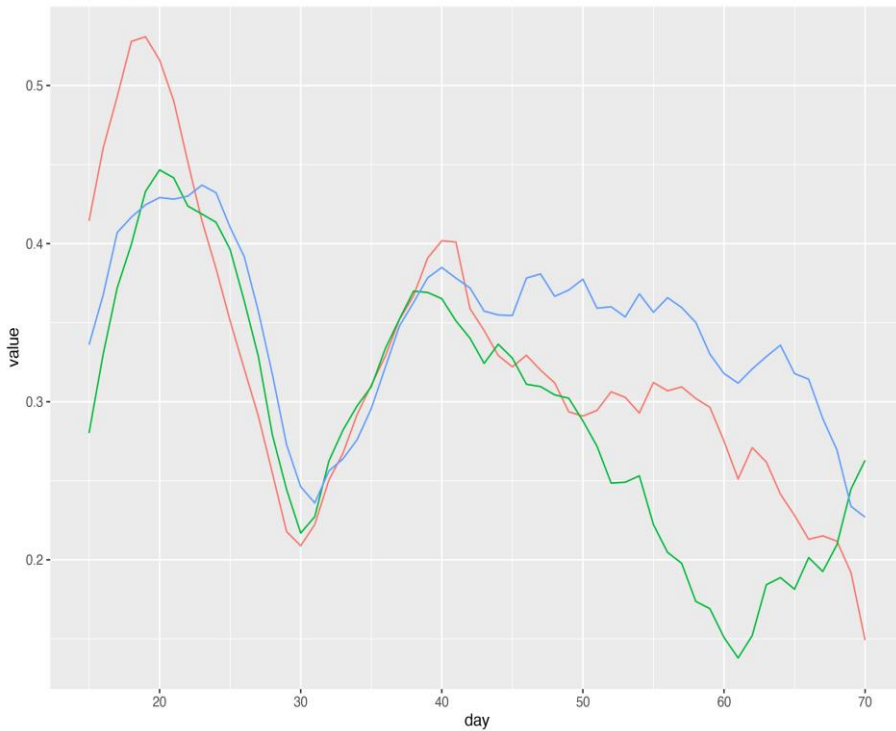
Rho= 0.7



Generazione dei dati

Indice di Moran

Moran Index - A + I



Moran Index - I



Campionamento

Simulazione – Standard Error

Day	15	29	43
Local cube (inf)	0.24	0.55	0.38
Local cube (inf-sp)	0.31	0.60	0.41
Local cube (sp)	0.32	0.63	0.42
Local pivotal	0.32	0.63	0.42
Prop sampling	0.36	0.67	0.45

Campionamento

Simulazione – Standard Error

Screening	Sampling method	SE
homogeneous	Local cube (inf)	0.047
homogeneous	Local cube (sp)	0.053
homogeneous	Local cube (inf-sp)	0.052
heterogeneous	Local cube (inf)	0.054
heterogeneous	Local cube (sp)	0.054
heterogeneous	Local cube (inf-sp)	0.052

Conclusioni

- Le nostre proposte tendono a fornire uno strumento agile e per la raccolta di dati in situazioni emergenziali.
- I piani di campionamento proposti sono più efficienti di quelli convenzionali (stessa accuratezza con meno unità campionarie)
- Sono pensate per essere ripetute nel tempo in un sistema quasi *real-time* di **monitoraggio, sorveglianza e controllo** di fenomeni epidemici senza eccessivo aggravio per le strutture sanitarie.
- Devono **affiancare** e **non sostituire** altre fonti di dati (tamponi, dati sanitari ufficiali, Big Data etc.).

Bibliografia

- ALLEVA G., ARBIA G., FALORSI P.D., NARDELLI V., ZULIANI A., 2021 a, A sample approach to the estimation of the critical parameters of the SARS-CoV-2 epidemics: an operational design.
<http://arxiv.org/abs/2004.0606>, accettato da *Journal of Official Statistics (JOS)*.
- ALLEVA G., ARBIA G., FALORSI P.D., NARDELLI V., ZULIANI A., 2021 b, A spatial sampling design to improve the efficiency of the estimation of the critical parameters of the SARS-CoV-2 epidemic, *Submitted*
- Alleva G., Falorsi P. D., Petrarca F., Righi F. (2021), Measuring The Accuracy Of Aggregates Computed From A Statistical Register, *Journal of Official Statistics*. To be published.
- Arbia, G. (1993) The use of GIS in spatial surveys", *International Statistical Review*, 61, 2, 1993, 339-359.
- Grafström A (2012) Spatially correlated Poisson sampling, *Journal of Statistical Planning and Inference* 142, 1, 139–147, DOI: 10.1016/j.jspi.2011.07.003
- Grafström A, Lundström NL, Schelin L. Spatially balanced sampling through the pivotal method (2012). *Biometrics*, 68, 2, 514-20. doi: 10.1111/j.1541-0420.2011.01699.x. Epub 2011 Oct 31. PMID: 22040115.
- Grafström, A. and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals". *Environmetrics*, 24, 120–131.
- Hamer W. H. (1906), *Epidemic diseases in England*, Lancet, 1.
- Lavallée, P., L. P. Rivest. 2012. Capture–Recapture Sampling and Indirect Sampling. *Journal of Official Statistics*, Vol. 28, No. 1, 2012, pp. 1–27.
- Kermack, W.O. and McKendrick, A. G. (1927) A contributions to the mathematical theory of epidemics" *Proceedings of the Royal society London, series A*, 115, 700-721.
- Tillé, Y. (2020). *Sampling and estimation from finite populations*. John Wiley & Sons.