Nowcasting Temporal Trends Using Indirect Surveys

Ajitesh Srivastava[1], Juan Marcos Ramírez[2],
Sergio Díaz Aranda[2,3], José Aguilar[2], Antonio Ortega[1],
Antonio Fernández Anta[2], Rosa Elvira Lillo[3]

[1]USC [2]IMDEA Networks [3]UC3M

El País, March 13th, 2020

España ya es el segundo país de la UE con más contagiados: 3.142 casos

Official number of confirmed cases on March 14th, 2020 (as of now):
6,391 ~ 0.0136% [OWID]
Is this close to the real number?


EL PAÍS

www.elpais.com

EL PERIÓDICO GLOBAL

VIERNES 13 DE MARZO DE 2020 | Año XLV | Número 15.581 | EDICIÓN MADRID | Precio: 1,70 euros

ALEMANIA El servicio secreto vigila al ala más radical de la ultraderechista AfD P2

EE UU Detenidos 200 nar cartel heredero de El C

EL AVANCE DEL CORONAVIRUS DESAFÍA AL ESTADO

España, en emergencia

El Ejecutivo inyecta 14.000 millones en la economía y 3.800 millones en sanidad

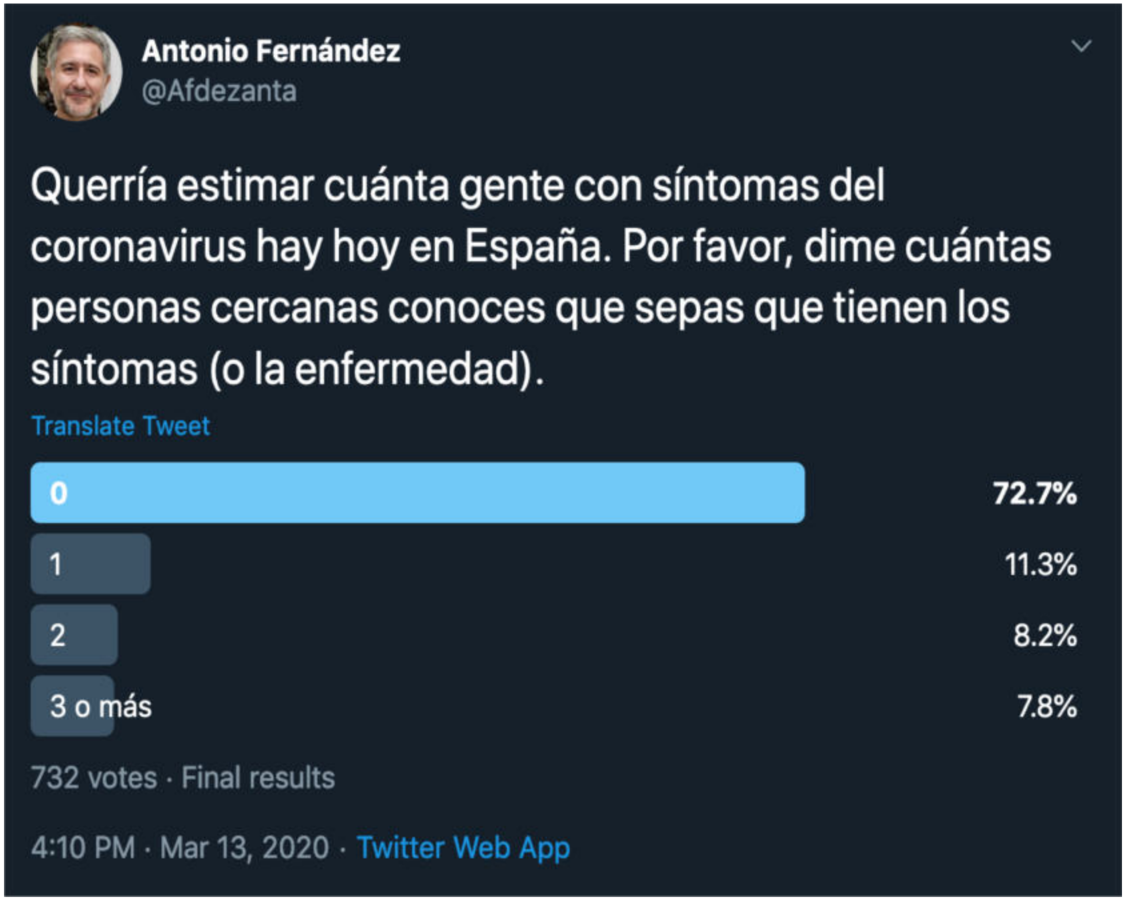
España ya es el segundo país de la UE con más contagiados: 3.142 casos

Sánchez propone un Presupuesto "extrasocial" y Arrimadas ofrece su voto

Casado critica p insuficiente el p apoyará en el C

COVID-19, March 14th, 2020

Official # cases: 6,391 ~ 0.0136%



732 responses
report 374 cases
&
know ~36,600 persons
(Dunbar # of 50 friends)
▼
480,000 cases ~ 1%

14 days onset to death
1.38% Case Fatality Ratio
5,982 deaths (March 28th)
▼
433,478 cases ~ 0.92%

Aggregated Relational Data

* **How many people do you know personally in this geographical area?**

Include only those whose health status you are likely to be aware of.

123 R_v

* **How many of the above have been diagnosed or have had symptoms compatible with COVID-19, to the best of your knowledge?**

Include those who had the symptoms and have recovered. Common symptoms include fever, tiredness, dry cough. Other symptoms include shortness of breath, aches and pains, sore throat, and very few people will report diarrhoea, nausea or a runny nose. (From the WHO webpage.)

123 C_v

Aggregated Relational Data (ARD): Data collected from a survey of indirect questions:

- Privacy is preserved
- Each response reports the status of many individuals
- Responses can be biased

Network Scale-up Metod (NSUM)

* How many people do you know personally in this geographical area?

Include only those whose health status you are likely to be aware of.

123 R_v

* How many of the above have been diagnosed or have had symptoms compatible with COVID-19, to the best of your knowledge?

Include those who had the symptoms and have recovered. Common symptoms include fever, tiredness, dry cough. Other symptoms include shortness of breath, aches and pains, sore throat, and very few people will report diarrhoea, nausea or a runny nose. (From the WHO webpage.)

123 C_v

The prevalence f is estimated with the Network Scale-up Method (NSUM) [Bernard et al, 1991; Laga et al, 2021]:

Mean of Ratios (MoR): $\hat{f} = \frac{1}{|S|} \sum_{v \in S} \frac{C_v}{R_v}$

Applications

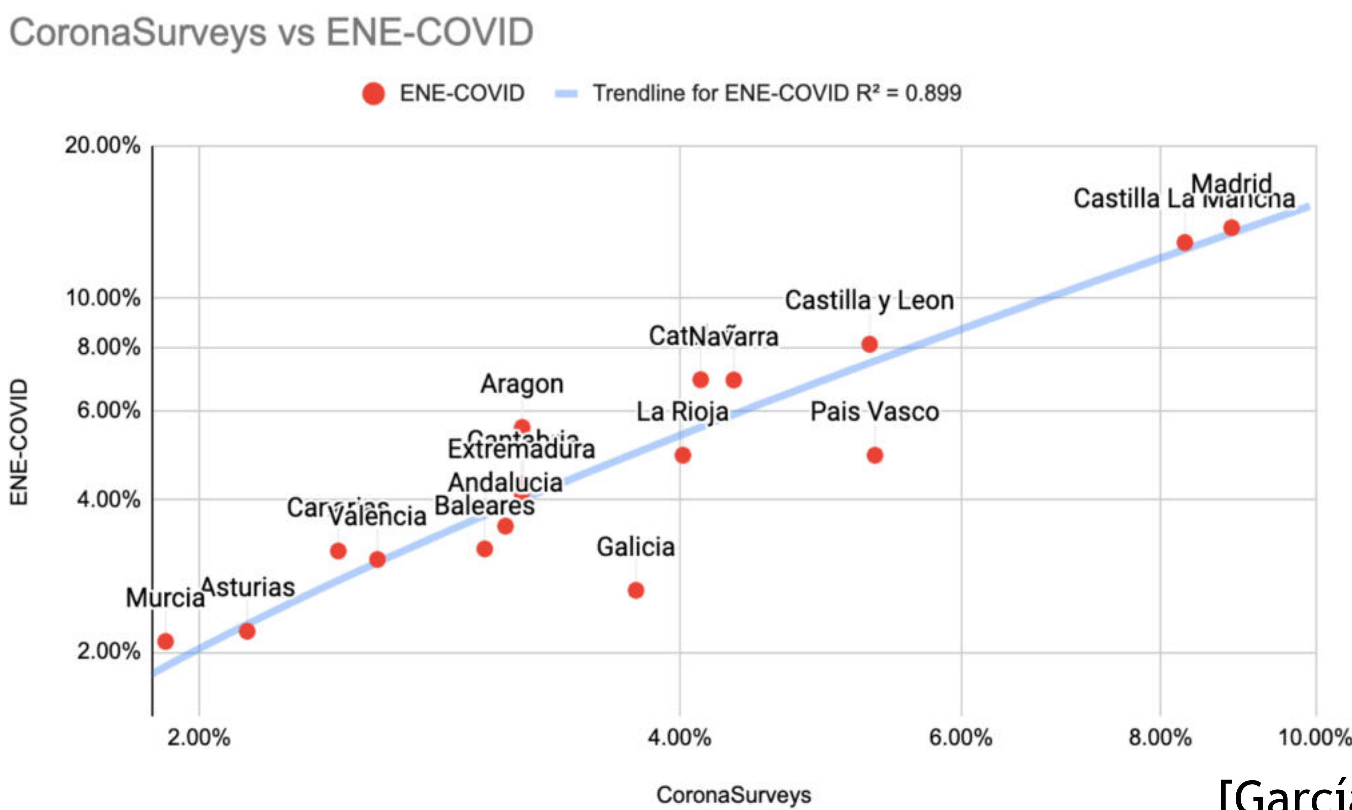
ARD collected in one-shot surveys to estimate:

- Casualties in an earthquake [Bernard et al. 1989]
- Female sex workers conditions [Jing et al. 2018]
- Prevalence of drug use [Salganik et al. 2010]
- Prevalence of HIV [Teo et al. 2019]
- Prevalence of COVID-19 [Garcia-Agundez et al. 2021]

Opportunity of online indirect surveys for continuous ARD collection and trend nowcasting

Validation: ENE-COVID

Serology (IgG) study in Spain with ~60,000 people on April 27th to May 11th, 2020: ENE-COVID



CoronaSurveys.org

April 20th, 2020

|S|= 999 responses

(9-100 resp/region)

Ratio of Sums (RoS)

R^2=0.8994

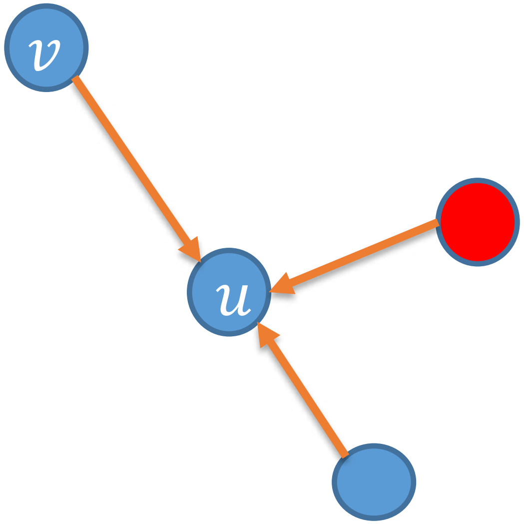
[Pollán et al., 2020]
[García-Agundez et al., 2021]

Contributions

- Latent dynamic graph formulation to prove that the estimated prevalence is proportional to the real prevalence
- ARD provides better prevalence estimate than a direct survey (w/ assumptions on degree variance of latent graph)
- Weighted moving average provides better estimates than a series of individual estimates
- Validate claims via simulations and real COVID-19 data

Latent Dynamic Graph

- Population N
- At time t:
    - Infected population H_t, and prevalence $f_t = \frac{|H_t|}{|N|}$
    - Graph $G_t = (N, E_t)$, where $(v, u) \in E_t$ if u knows whether v is infected

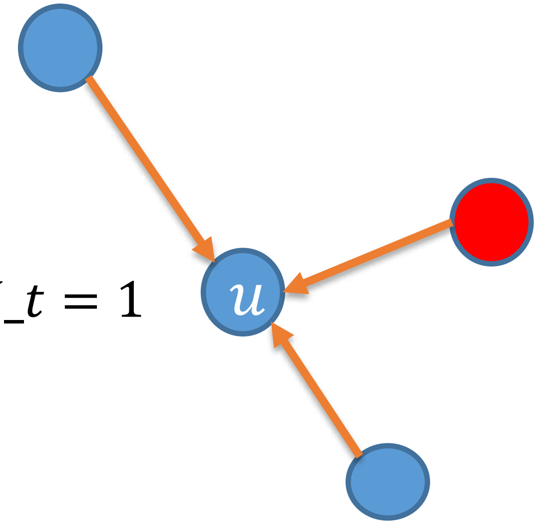


- Assumptions:
    - In-degree distribution of all G_t have the same mean μ and variance σ^2(supported empirically [Dunbar 2010])
    - If $(v, u) \in E_t$, $\Pr(v \in H_t)$ does not depend on the in-degree of u in G_t

Sampling

- Select a node u at time t from G_t uniformly at random
- Let X_t be the (random variable of) number of infected in-neighbors of u

In this case, $X\_t = 1$



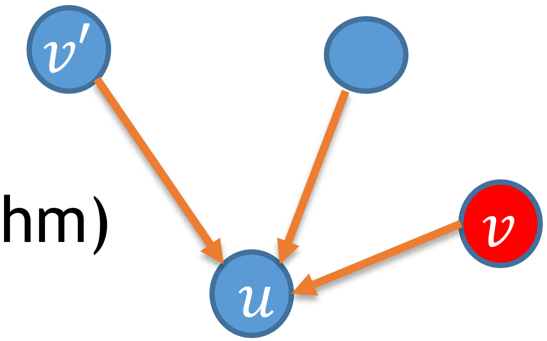
Prevalence Trend Estimation

Theorem: $E(X_t) = \mu \cdot f_t$

- I.e., the expectation of the indirect response X_t is proportional to the prevalence f_t we wish to estimate
- The time series $E(X_t)$ is proportional to the time series of f_t, with μ as the constant of proportionality
- The trend of f_t can be estimated without knowing μ
- If precise f_t values are needed, μ can be estimated (once) from reliable data

Indirect versus Direct Reporting

Y_t : random variable of whether node u is infected
$\bar{Y}_t$: sampled mean of Y_t
$\bar{X}_t$: sampled mean of X_t
ϕ_t: probability of co-infection
$|S| = n$: large sample size (Central Limit Thm)



Theorem: *For any* $\lambda > 0$, if

$$\sigma^2 \le \frac{\mu(\mu - 1)(1 - \phi_t)}{\phi_t}$$

then $\Pr(|\bar{X}_t/\mu - f_t| > \lambda) \le \Pr(|\bar{Y}_t - f_t| > \lambda)$

I.e., indirect surveys are better than direct surveys for same sample size n

Advantage of Smoothing

Assumption:

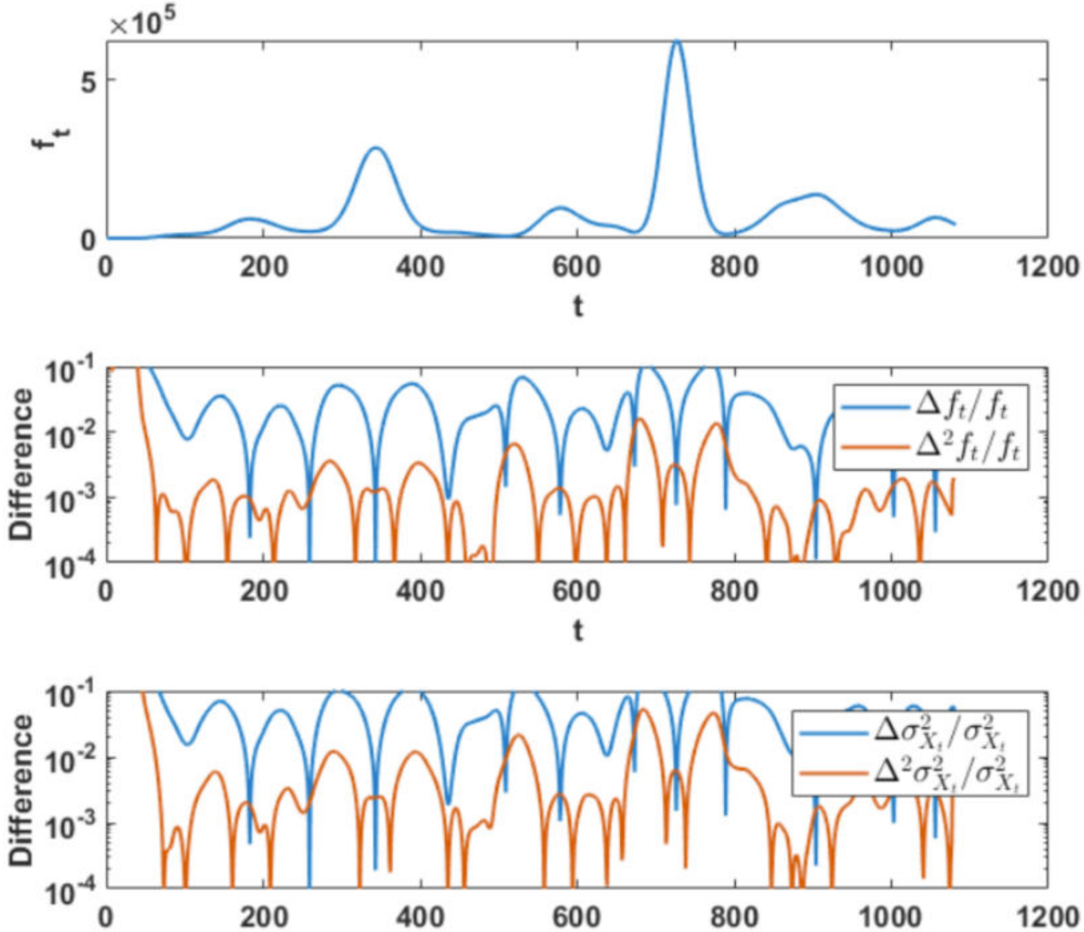
$$|\Delta f_t| \leq \varepsilon_f \, f_t$$

$$\left|\Delta\sigma^2_{X_t}\right| \leq \varepsilon_\sigma \, \sigma^2_{X_t}$$

for small $\varepsilon_f, \varepsilon_\sigma \geq 0$

(f_t and $\sigma^2_{X_t}$ change slowly over time)

Ex: COVID-19 cases in California



Advantage of Smoothing

$\bar{X}_{t,w}$: mean of X_t on $[t-w, t+w]$
n_t : number of samples at time t
n_w : sum of n_t on $[t-w, t+w]$

Theorem: *If*

$$\lambda \geq \frac{w\,\varepsilon_f}{1-\left(\frac{1}{1-w\,\varepsilon_\sigma}\right)\sqrt{\frac{n_t}{n_w}}}$$

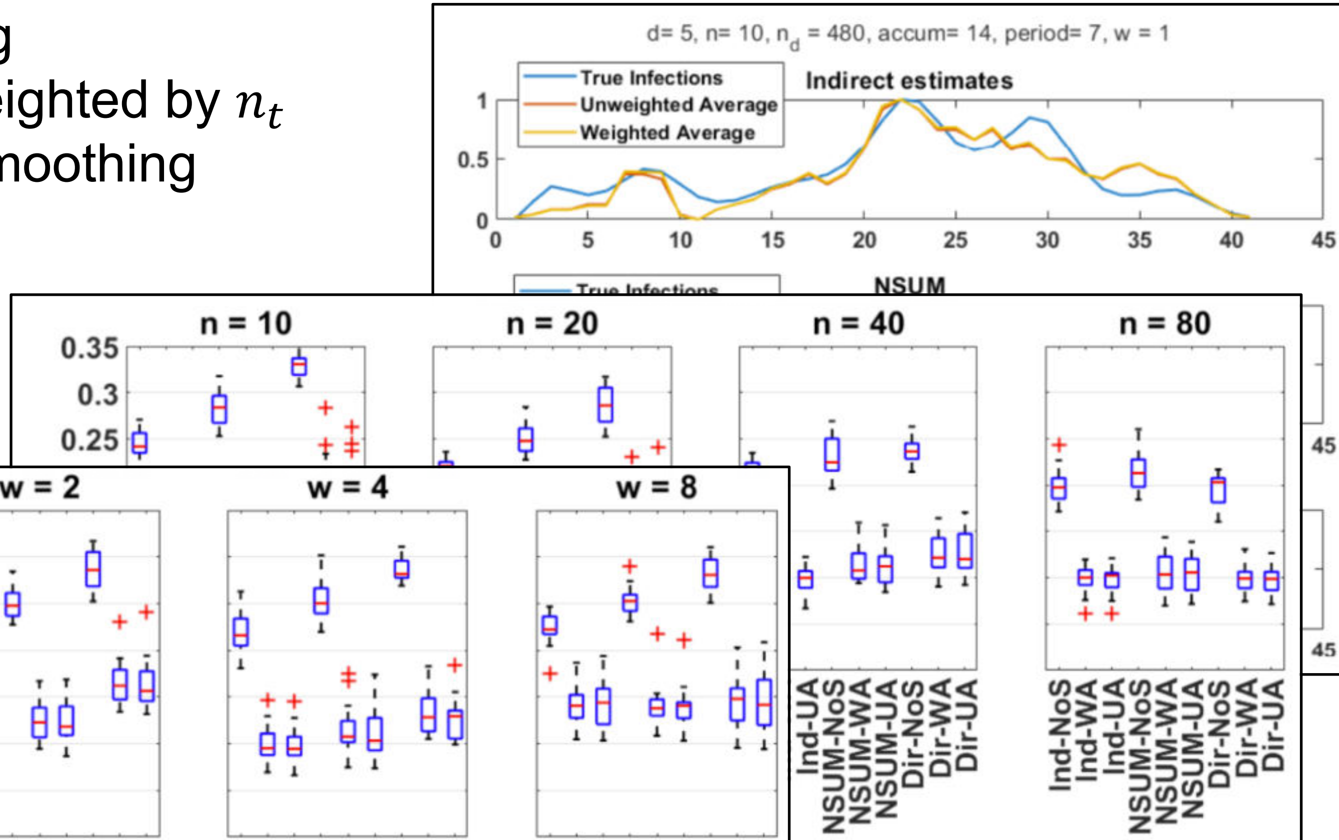
then

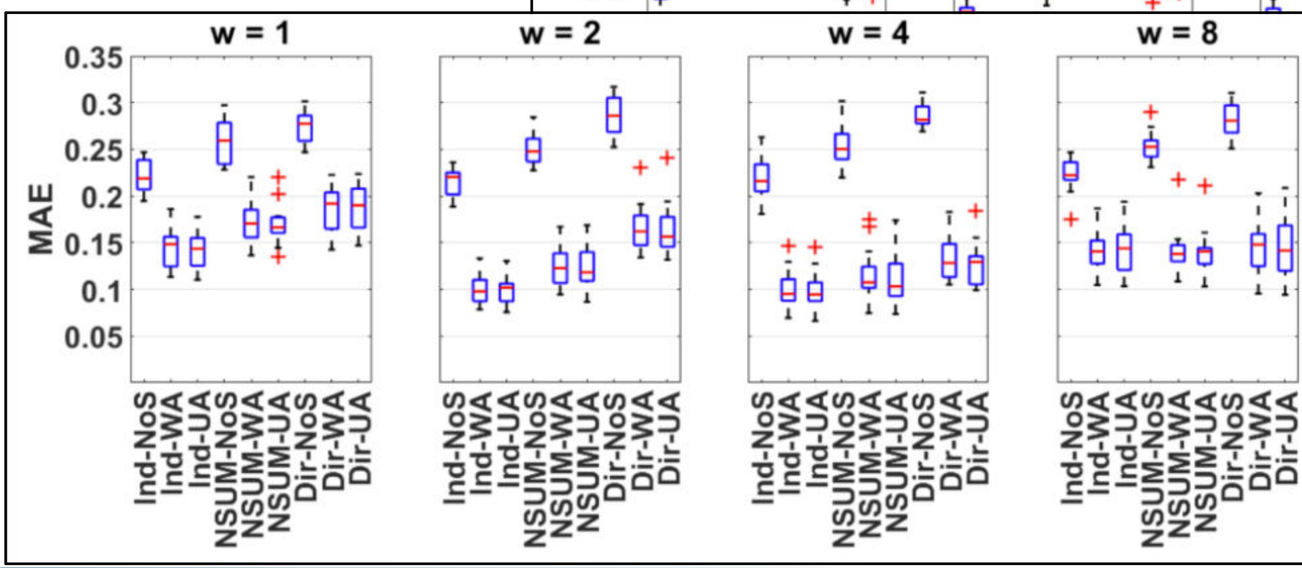
$$\Pr(|\bar{X}_{t,w}/\mu - f_t| \geq \lambda\, f_t) \leq \Pr(|\bar{X}_t/\mu - f_t| \geq \lambda\, f_t)$$

I.e., the smoothed estimate $\bar{X}_{t,w}/\mu$ is less likely to deviate by λ from the true value than the instantaneous value $\bar{X}_t/\mu$

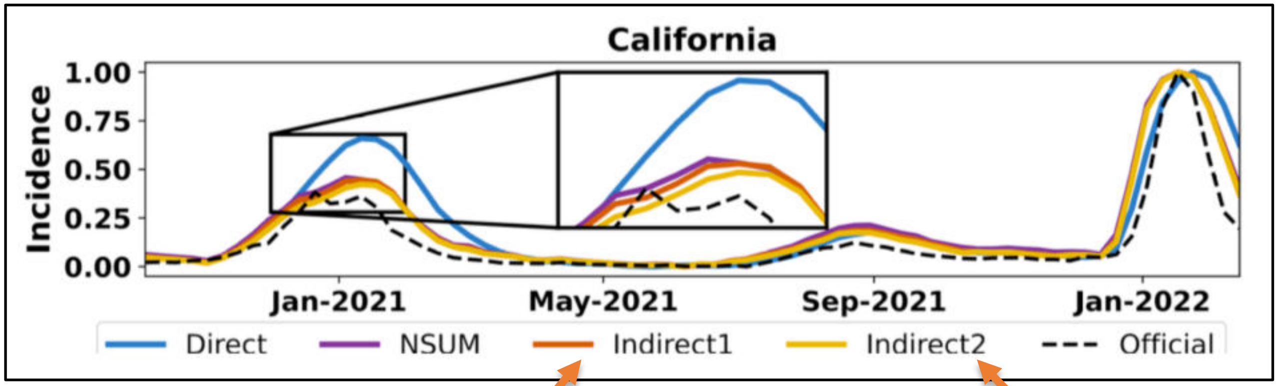
Simulations of an Epidemic Model

NoS: No smoothing
WA: Smoothing weighted by n_t
UA: Unweighted smoothing
Ind: Indirect
Dir: Direct
MAE of normalized curves





Real COVID-19 Data



CLI incidence reported in the household

CLI incidence reported within the local community

US COVID-19 Trends and Impact Survey (CTIS) [Salomon et al. 2021]

Real COVID-19 Data

accum	*w*	state	Direct	NSUM	Indirect1	Indirect2
7	1	CA	0.0703	0.0443	0.0341	**0.0292**
		TX	0.0661	0.0379	0.0289	**0.0270**
		NY	0.0785	0.0315	0.0301	**0.0299**
		PN	0.0572	0.0368	0.0300	**0.0263**
	3	CA	0.1148	0.0988	0.0881	**0.0811**
		TX	0.1236	0.0890	0.0813	**0.0782**
		NY	0.1210	0.0930	0.0910	**0.0886**
		PN	0.0956	0.0907	0.0816	**0.0691**
14	1	CA	0.0836	0.0624	0.0524	**0.0477**
		TX	0.0779	0.0385	0.0343	**0.0336**
		NY	0.0929	0.0520	0.0504	**0.0500**
		PN	0.0689	**0.0389**	0.0391	0.0429
	3	CA	0.1441	0.1217	0.1116	**0.1059**
		TX	0.1349	0.1058	0.1042	**0.1027**
		NY	0.1571	0.1165	0.1126	**0.1090**
		PN	0.1349	0.1182	0.1110	**0.1005**

CLI incidence reported in the household

CLI incidence reported within the local community

MAE of the normalized COVID-19 incidence curves

US COVID-19 Trends and Impact Survey (CTIS) [Salomon et al. 2021]

Conclusions

- Indirect surveys are a useful tool to monitor society
- Provide good estimates even with limited number of responses
- Can be easily used to monitor trends
- Limits and assumptions that make it applicable have to be explored
- Not widely exploited over time and space: opportunities for research in dynamic networks

Future Work

- Monitoring of social phenomena:
  - Epidemics (COVID-19, monkey pox, malaria)
  - Harassment and bullying incidence
  - Customer opinions and marketing
  - Vote intention
- Evolution over time of these phenomena
- We need to understand better the limitations of the method both for one-shot and continuous monitoring:
  - Worst cases
  - Average practical cases

Thank you!

Honorable Mention to Best Paper for the AI for Social Impact track at AAAI-2024

Ajitesh Srivastava, Juan Marcos Ramirez, Sergio Díaz [...], José Aguilar, Antonio Fernández Anta, Antonio Ortega, Rosa Elvira Lillo: **Nowcasting Temporal Trends Using Indirect Surveys.** AAAI 2024: 22359-22367
https://doi.org/10.48550/arXiv.2307.06643

coronasurveys.org

MINISTERIO DE CIENCIA E INNOVACIÓN
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia
AGENCIA ESTATAL DE INVESTIGACIÓN

This research was supported by SocialProbing project (TED2021-131264B-I00), and PID2019-104901RB-I00 funded by MCIN/AEI /10.13039/501100011033 and the European Union-NextGenerationEU/PRTR.