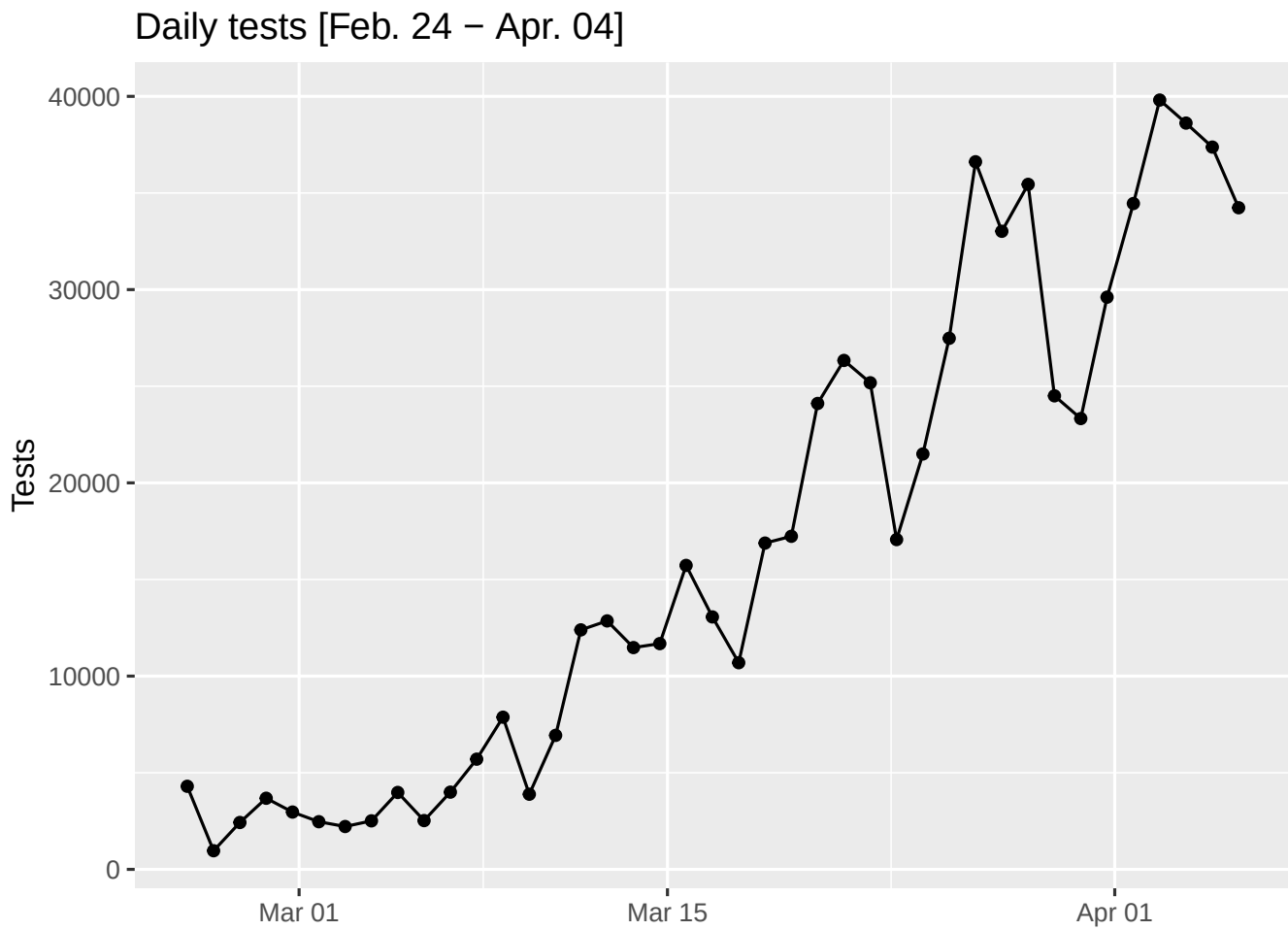
Searching for the peak
Google Trends and the monitoring of the Covid-19 outbreak in Italy

09 July 2020

Paolo Brunori
University of Florence & University of Bari

Giuliano Resce
SOSE - Soluzioni per il Sistema Economico S.p.A.

COVID-19 outbreak official data



Source: Dipartimento Protezione Civile

COVID-19 outbreak official data

Approfondimento Coronavirus

Borrelli: "I numeri sono altri. L'epidemia va più veloce della nostra burocrazia"

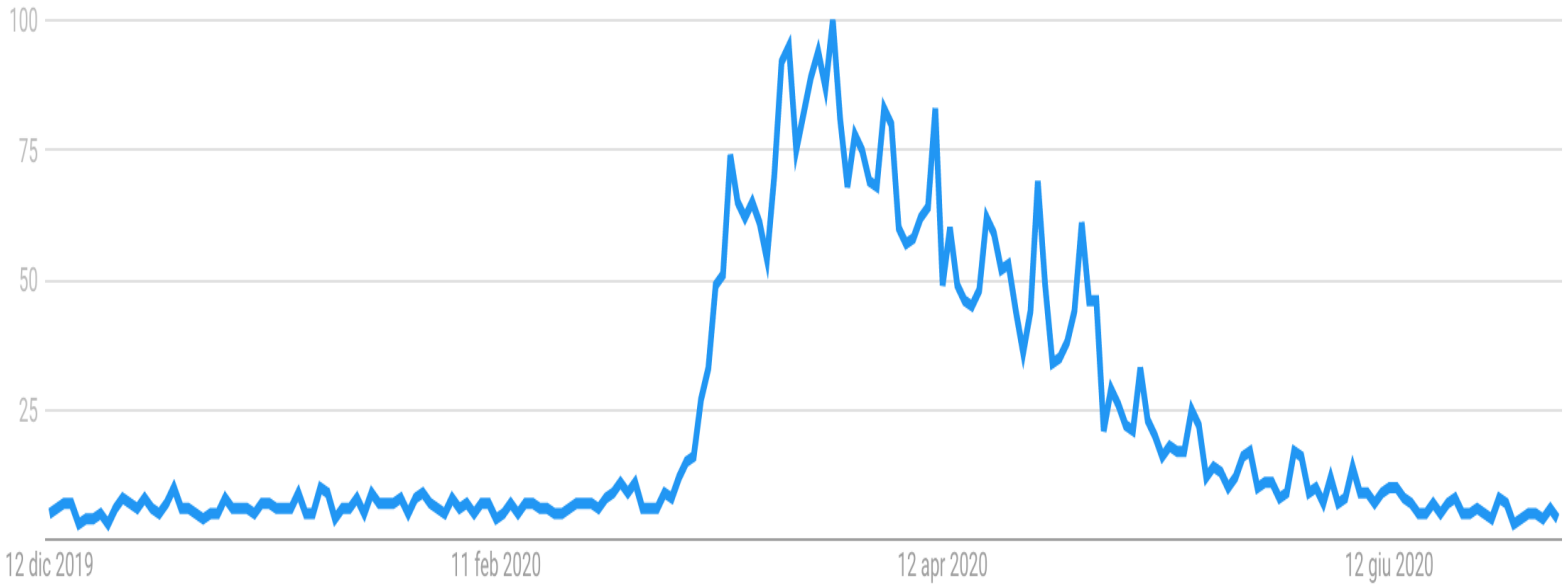
23 MARZO 2020

Intervista al capo della Protezione civile: "I contagiati ufficiali a ieri sera erano 63 mila. Ma il rapporto di un malato certificato ogni dieci non censiti è credibile"

Big data Nowcasting

- Big data can improve our understanding of human-related phenomena;
- Processed in real time;
- Large samples (often the entire population is available);
- In a few cases freely available (Google Trends).

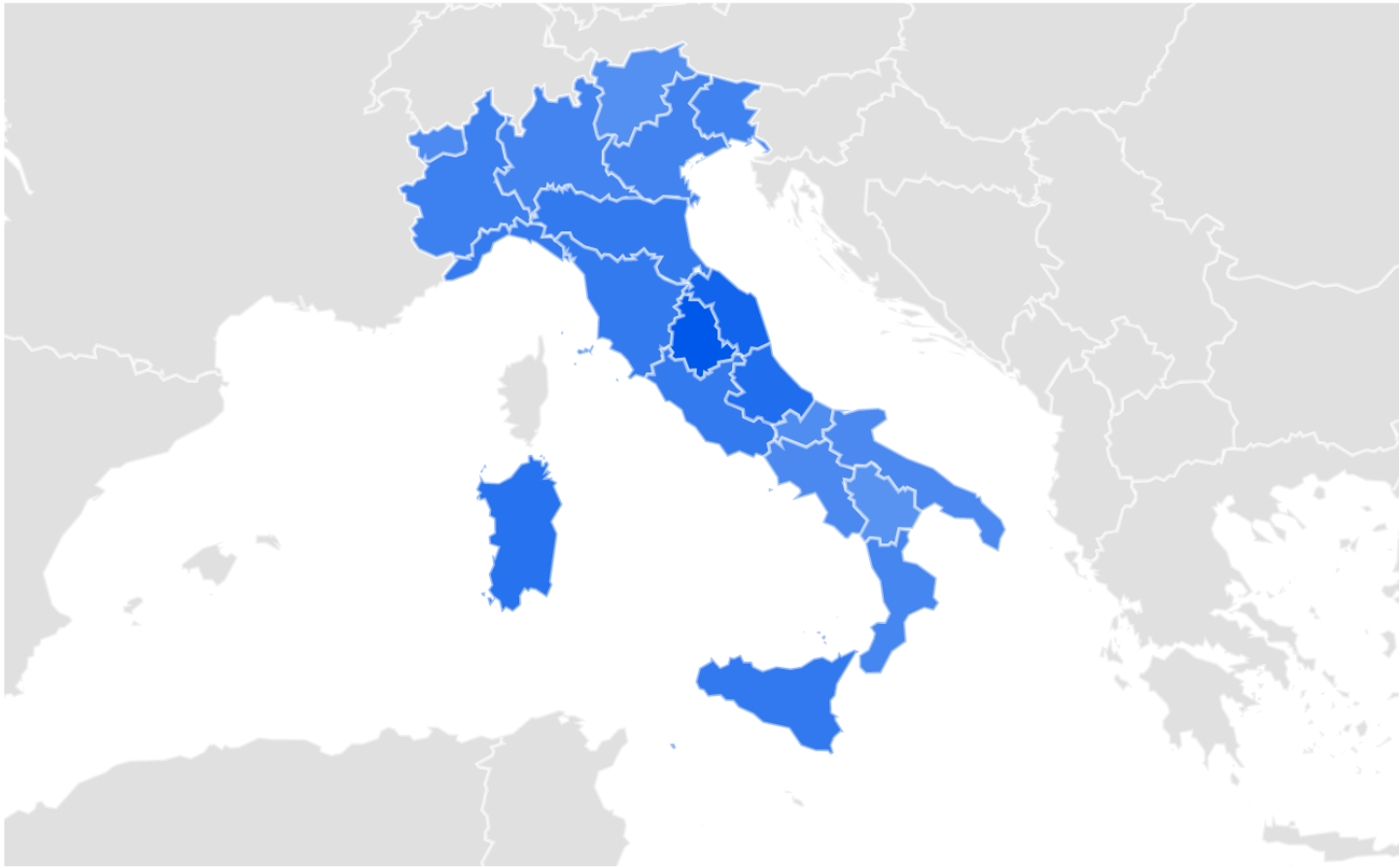
Google Trends: time series



Query: 'ricetta del pane'

Source: Google Trends [Dec. 8, 2019 - July 7, 2020]

Google query: regional disaggregation



Query: 'ricetta del pane'

Source: Google Trends [Dec. 8, 2019 - July 7, 2020]

Google Trends for macroeconomic nowcasting

- unemployment rate (Choi and Varian, 2009; Askitas and Zimmermann, 2009; Scott and Varian, 2014; D'Amuri and Marcucci, 2017; Naccarato et al., 2018);
- GDP growth (Narita and Yin, 2018; Ferrara and Simone, 2019);
- Inflation (Koop and Strathclyde, 2019) and housing prices (Wu and Brynjolfsson, 2013).

Epidemiology: Google Flu Trends

- Based on Ginsberg et al. (2008) Google developed Google Flu Trends (GFT):
- GFT operated in 25 countries predicting flu-like illness spread between 2008 and 2015;
- GFT's predictors were 45 queries trained on flu diffusion official statistics;
- Strong correlation with flu diffusion in all flu seasons 2009-2014 (Yang et al., 2015).

Why Google Flu Trends fails?

- February 2013 failure;
- Lazer et al. (2014): prediction errors were following a seasonal pattern ('big data hubris');
- Google algorithm was modified;
- users' behaviour may change (e.g. 2009-2010 'swine' flu pandemic).

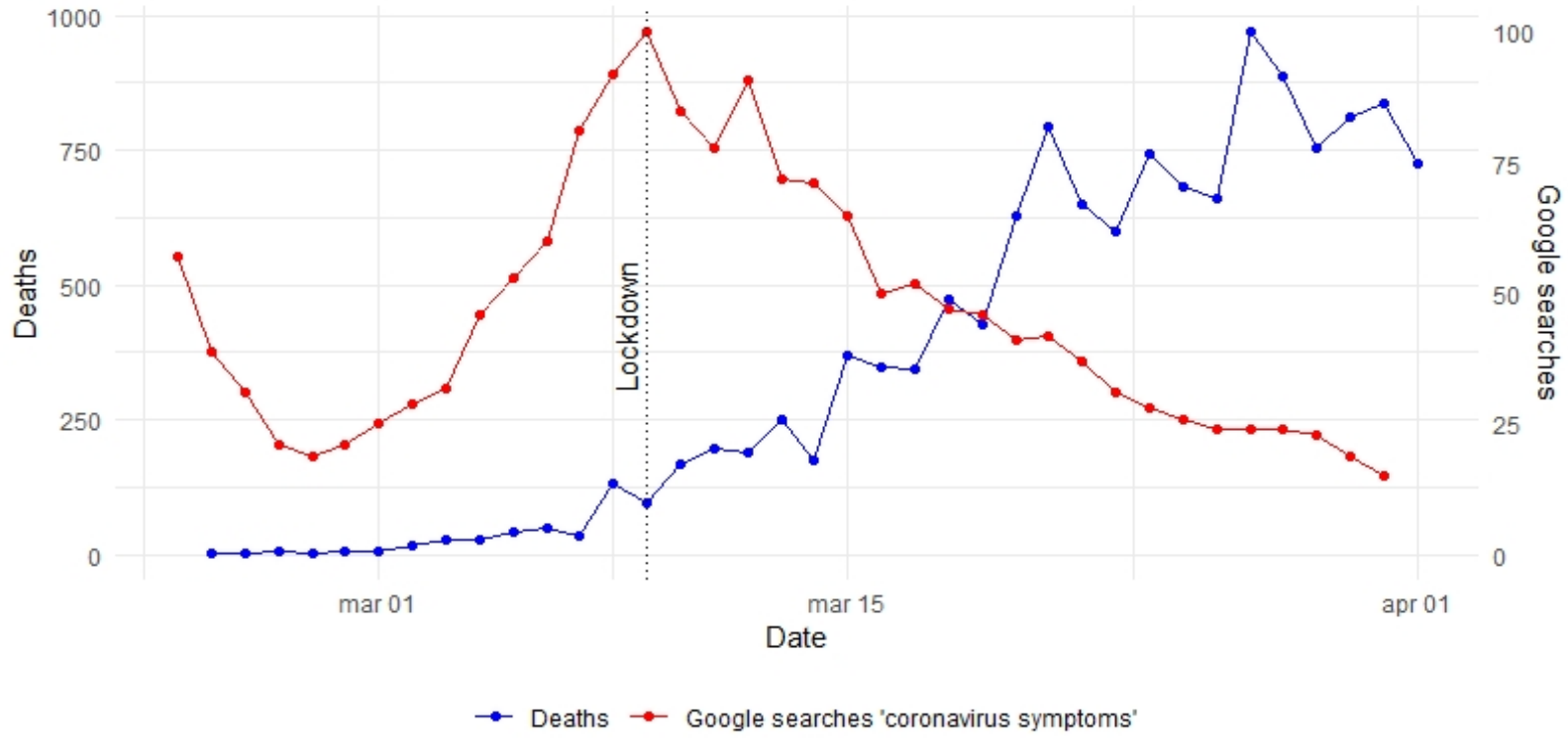
Google Trends and COVID-19 outbreak

- key question in March 2020: did we pass the peak?
- Cook et al. (2011): FGT produced biased estimates but did identify peaks;
- Google queries for 'Coronavirus symptoms' were steeply declining;
- Was this a sign of a diffusion slow down?

Methodological choices

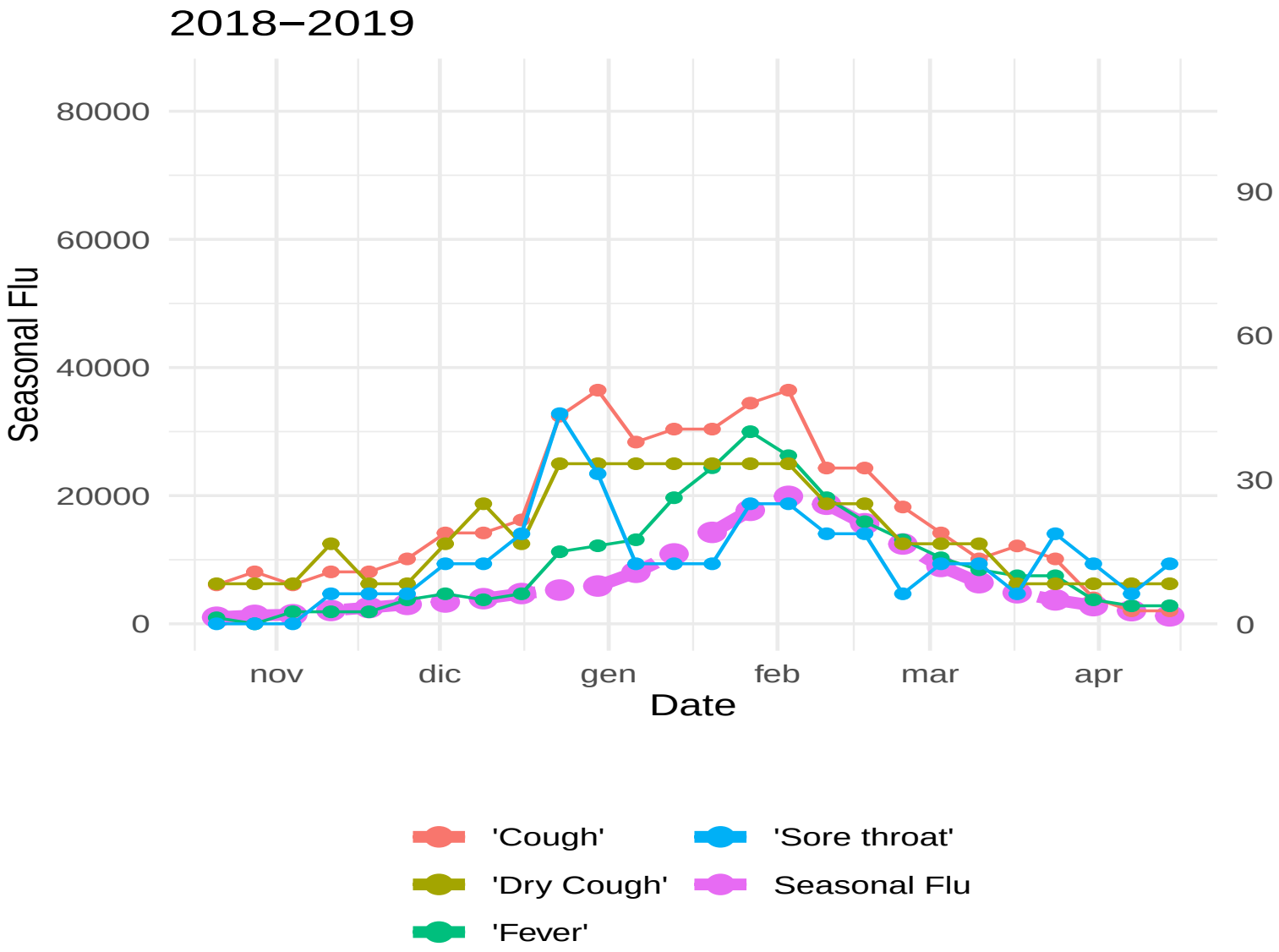
1. Focus on ‘number of deaths’ and ‘ICU patients’ rather than ‘number of positive’;
2. Exploit regional heterogeneity in deaths/ICU patients and Google queries;
3. Ignore queries that contained the words ‘Coronavirs’ and ‘COVID-19’.

Google queries and deaths



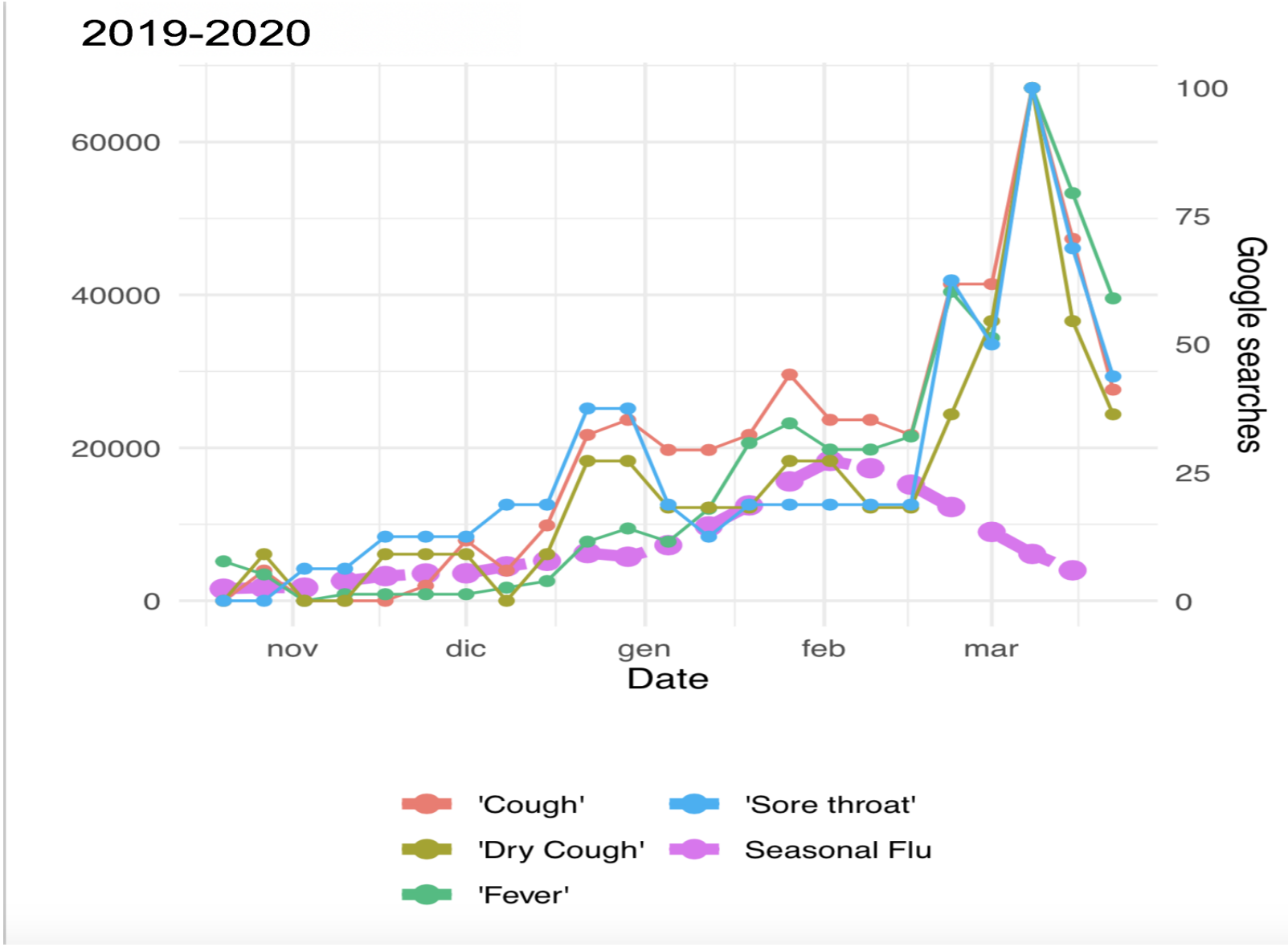
Source: Google Trends and Dipartimento Protezione civile

Google queries and flu cases in Italy



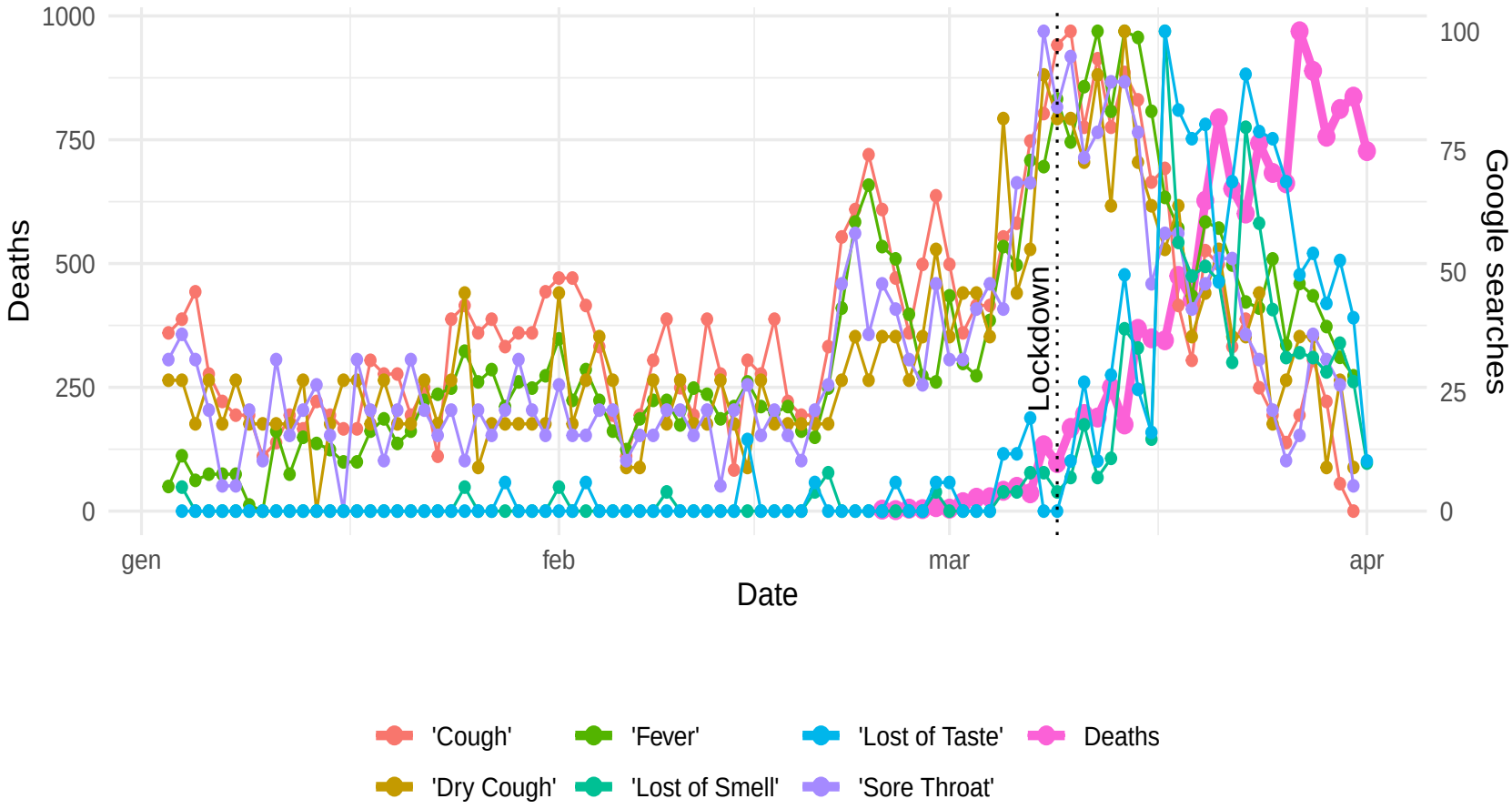
Source: Google Trends and Ministero della Salute

Google queries and flu cases in Italy



Source: Google Trends and Ministero della Salute

Google queries and COVID-19 deaths



Source: Google Trends and Protezione Civile.

A regional analysis

- Period considered: 2020-02-24 to 2020-03-28;
- Regions: 14 out of the 20 Italian regions (93% of the national population);
- Queries: most frequently reported symptoms as listed by the European Centre for Disease Prevention and Control;

A regional analysis, cnt

$$y_{it} = \theta_i + \phi_t + \beta_{t-1} \sum_{j=0}^{p} S^j_{it-1} + ... + \beta_{it-15} \sum_{j=0}^{p} S^j_{it-15} + \epsilon_{it}$$

y_{it} = deaths/ICU patients in region i at time t;

θ_i = regional fixed effect;

ϕ_t = time fixed effect;

S^j_{it-k} = query j in region i.

Estimation

- The model is estimated using the least absolute shrinkage and selection operator (LASSO);
- Similar to an OLS regression LASSO assumes an additive and linear DGP;
- It searches for parameters that minimize:

$$\sum_{i=1}^{n} (\text{Residual Sum of Squares}) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Where λ is the shrinkage parameter.

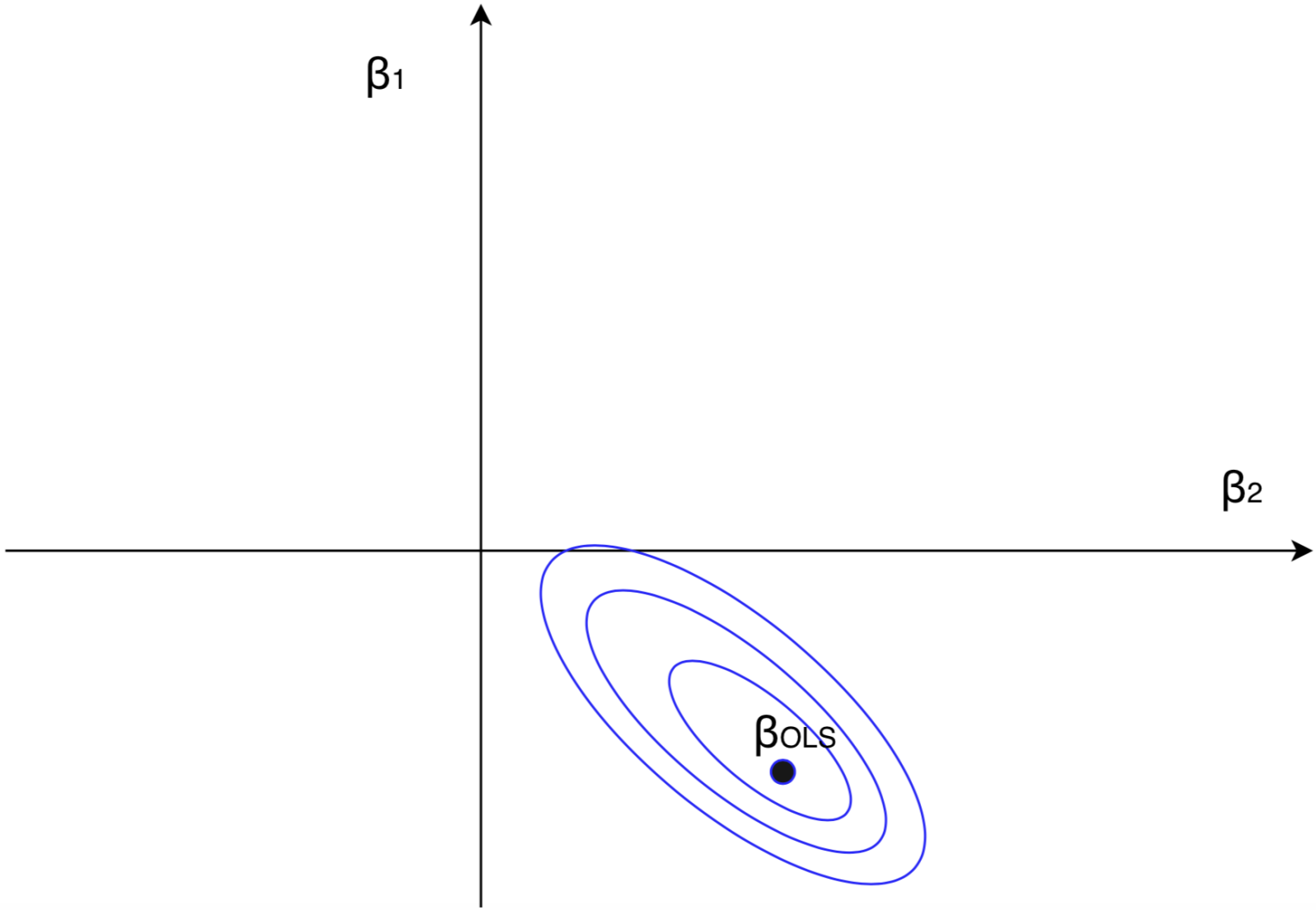
LASSO

This is equivalent to:

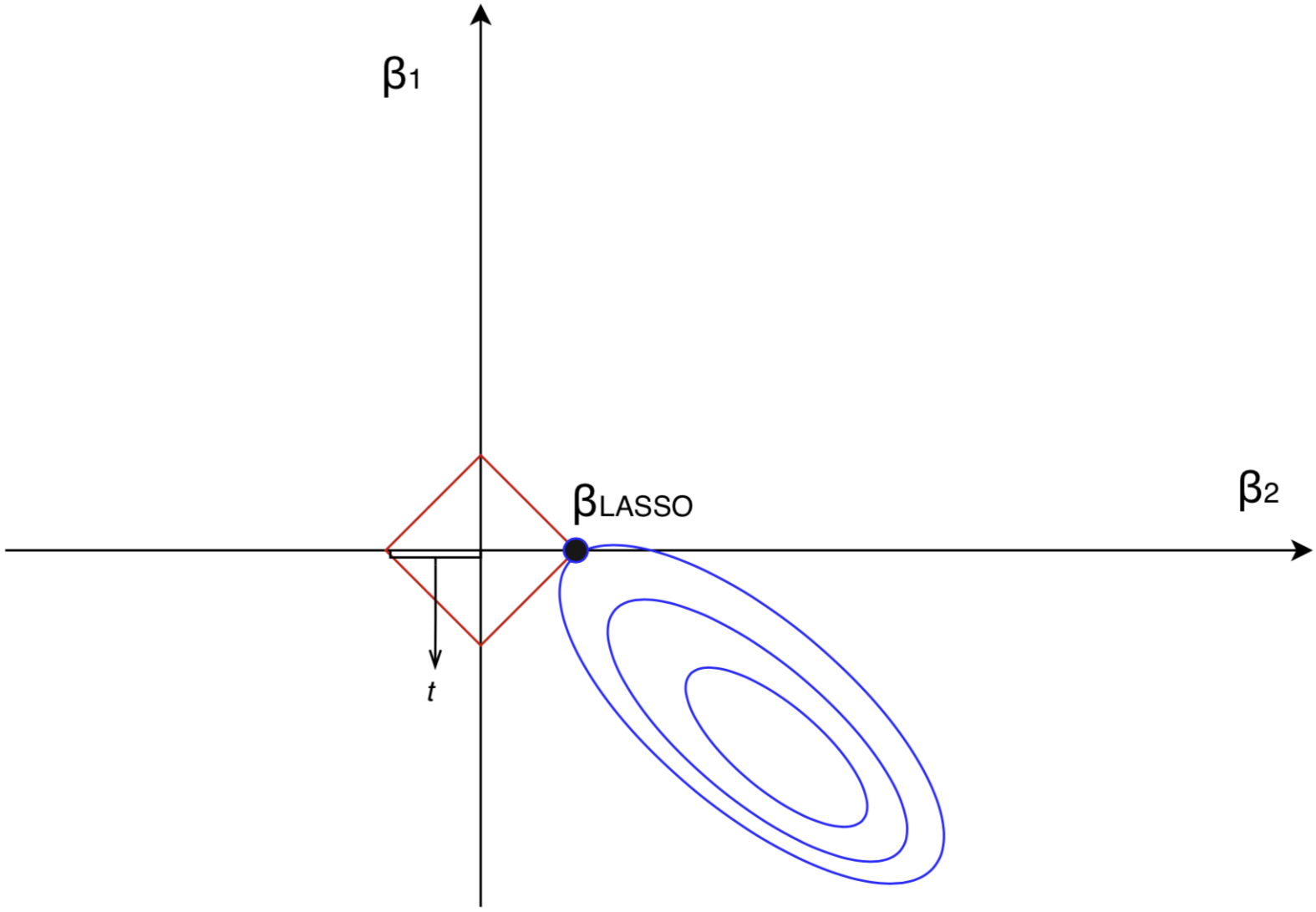
$$\hat{\beta}_{LASSO} = \underset{\beta}{\text{argmin}} \, (\text{Residual Sum of Squares})$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

OLS



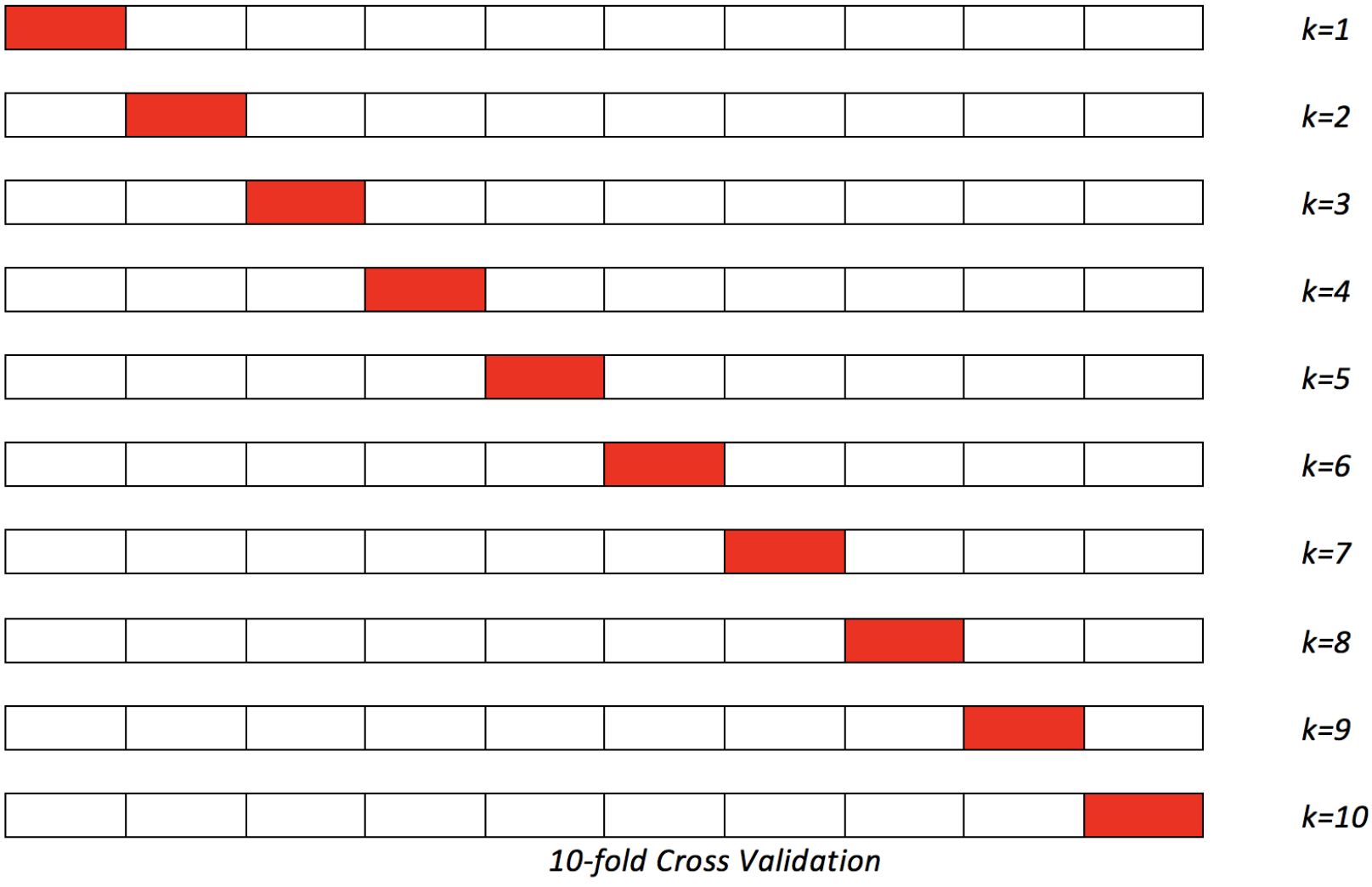
LASSO, cnt



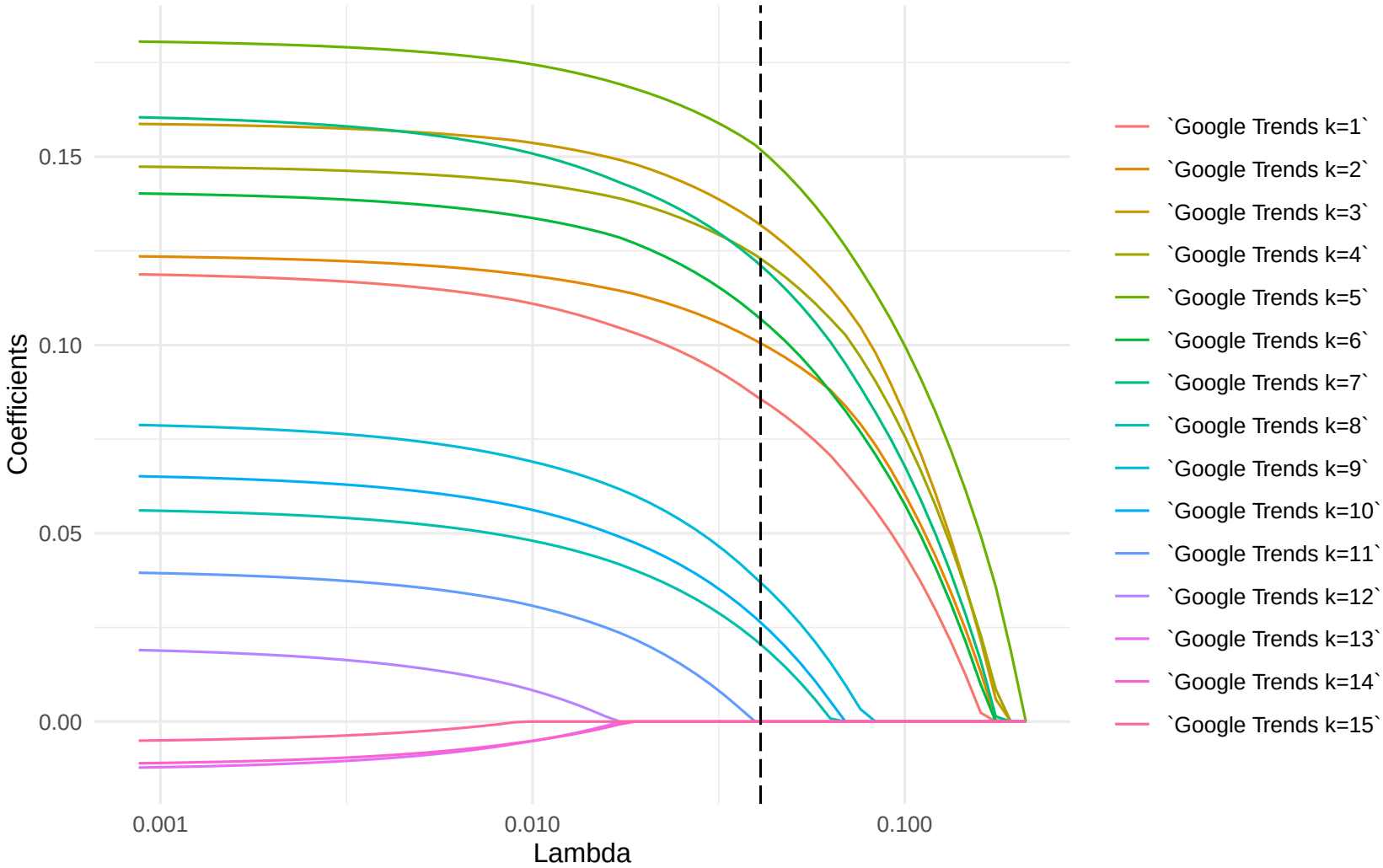
Tuning

- λ is determined to minimize the out-of-sample Mean Squared Error (MSE);
- MSE is estimated by 10-fold cross-validation.

K-fold cross-validation



Results


0.15
0.10
0.05
0.00
Coefficients
0.001
0.010
0.100
Lambda
`Google Trends k=1`
`Google Trends k=2`
`Google Trends k=3`
`Google Trends k=4`
`Google Trends k=5`
`Google Trends k=6`
`Google Trends k=7`
`Google Trends k=8`
`Google Trends k=9`
`Google Trends k=10`
`Google Trends k=11`
`Google Trends k=12`
`Google Trends k=13`
`Google Trends k=14`
`Google Trends k=15`

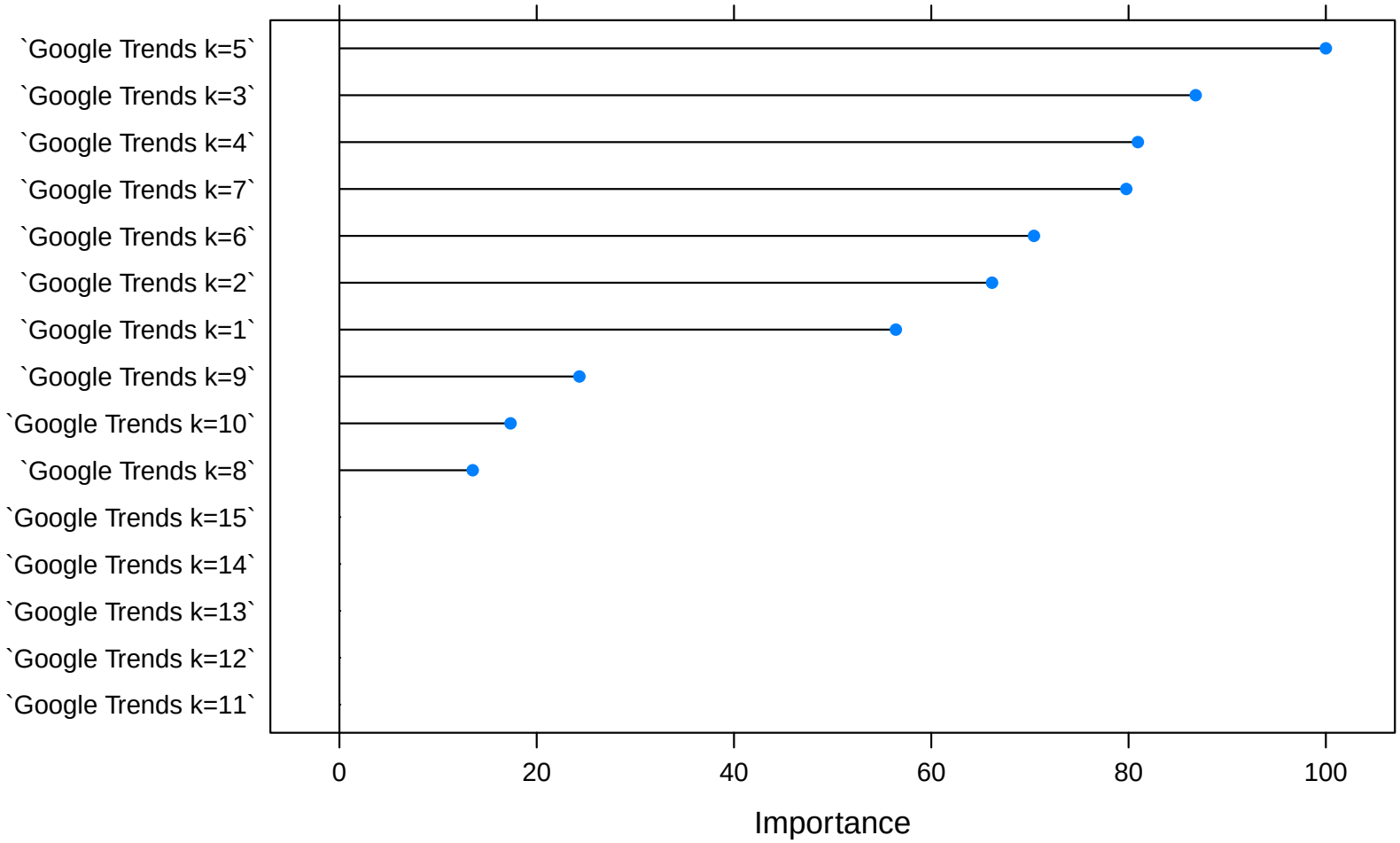
Results, cnt

Table: LASSO and ordinary least square coefficients

	LASSO estimate	OLS Estimate	Std. Error	t-value	p-value	
Google Trends k=1	0.086	0.037	0.020	1.824	0.070	.
Google Trends k=2	0.100	0.039	0.020	1.921	0.056	.
Google Trends k=3	0.132	0.048	0.019	2.472	0.014	*
Google Trends k=4	0.123	0.044	0.019	2.314	0.022	*
Google Trends k=5	0.152	0.054	0.018	2.922	0.004	**
Google Trends k=6	0.107	0.040	0.018	2.239	0.026	*
Google Trends k=7	0.121	0.045	0.017	2.615	0.010	**
Google Trends k=8	0.021	0.015	0.017	0.887	0.376	
Google Trends k=9	0.037	0.020	0.017	1.213	0.226	
Google Trends k=10	0.026	0.015	0.016	0.935	0.351	
Google Trends k=11	.	0.008	0.016	0.510	0.611	
Google Trends k=12	.	0.003	0.016	0.202	0.840	
Google Trends k=13	.	-0.005	0.016	-0.335	0.738	
Google Trends k=14	.	-0.006	0.016	-0.351	0.726	
Google Trends k=15	.	-0.004	0.015	-0.278	0.781	

Data: Google Trends and Iastituto Superiore di Sanità.
Note: Balanced Panel: n = 14, T = 19, N = 266.
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
MSE Panel Linear Model = 0.901
MSE Lasso Best Model = 0.889

Relative importance



Results: number of ICU patients

Table: LASSO and ordinary least square coefficients

	Lasso estimate	Estimate	Std. Error	t-value	p-value	
Google Trends k=1	0.121	0.269	0.141	1.907	0.058	.
Google Trends k=2	0.102	0.223	0.128	1.739	0.083	.
Google Trends k=3	0.175	0.423	0.114	3.714	0.000	***
Google Trends k=4	0.188	0.551	0.107	5.167	0.000	***
Google Trends k=5	0.179	0.482	0.100	4.826	0.000	***
Google Trends k=6	0.158	0.362	0.091	3.974	0.000	***
Google Trends k=7	0.108	0.273	0.086	3.182	0.002	**
Google Trends k=8	0.046	0.215	0.084	2.559	0.011	*
Google Trends k=9	0.075	0.162	0.083	1.959	0.051	.
Google Trends k=10	0.026	0.073	0.077	0.939	0.349	
Google Trends k=11	0.013	0.071	0.076	0.929	0.354	
Google Trends k=12	.	0.029	0.075	0.388	0.698	
Google Trends k=13	.	-0.023	0.074	-0.317	0.751	
Google Trends k=14	.	0.023	0.069	0.336	0.737	
Google Trends k=15	.	0.021	0.067	0.319	0.750	

Data: Google Trends and Iastituto Superiore di Sanità.
Note: Balanced Panel: n = 14, T = 19, N = 266.
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
MSE Panel Linear Model = 0.901
MSE Lasso Best Model = 0.889

Conclusion

- Google Trends would have correctly predict COVID-19 peak in Italy in late March;
- Even in difficult times, Big data cannot substitute official statistics;
- But they can complement standard data sources, especially in difficult times.