UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Dipartimento di Economia e Finanza

Southern Europe Research in Economic Studies

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

Paolo Brunori & Giuliano Resce

SERIES Working Papers n. 04/2020

SERIES sono pubblicati a cura del Dipartimento di Scienze economiche e metodi matematici dell'Università degli Studi di Bari "Aldo Moro". I lavori riflettono esclusivamente le opinioni degli autori e non impegnano la responsabilità del Dipartimento. SERIES vogliono promuovere la circolazione di studi ancora preliminari e incompleti, per suscitare commenti critici e suggerimenti. Si richiede di tener conto della natura provvisoria dei lavori per eventuali citazioni o per ogni altro uso.

SERIES are published under the auspices of the Department of Economics of the University of Bari. Any opinions expressed here are those of the authors and not those of the Department. Often SERIES divulge preliminary or incomplete work, circulated to favor discussion and comment. Citation and use of these paper should consider their provisional character.

Searching for the peak
Google Trends and the Covid-19 outbreak in Italy[*]

Paolo Brunori[†], Giuliano Resce[‡]

April 4, 2020

Abstract

One of the difficulties faced by policy makers during the Covid-19 outbreak in Italy was the monitoring of the virus diffusion. Due to changing criteria and insufficient resources to test all suspected cases, the number of 'confirmed infected' cases rapidly proved to be unreliably reported by official statistics. This limited the ability of epidemiologic models to predict the evolution of the infectious disease. This paper explores the possibility of using information obtained from Google Trends to supplement official statistics in order to predict when the number of deaths due to Covid-19 will peak in Italy. We estimate and regularize a panel model with regional and time fixed effects. Our preferred specification shows a positive and significant correlation between Google searches for commonly reported Covid-19 symptoms and deaths recorded. The analysis suggests that the social distancing measures implemented in early March in Italy were effective in slowing down the spread of the virus.

Keywords: Covid-19, Google Trends, Lasso.
JEL: I18, D83, C10.

[*] Although the authors alone are responsible for what is written in this paper, thanks are extended to David Clément, Bia Carneiro, Giovanni Cerulli, Michele Raitano, Lorenzo Busoni, Vincenzo Mariani for helpful comments.

[†] University of Florence, paolo.brunori@unifi.it.

[‡] Sose, giuliano.resce@uniroma3.it.

I BACKGROUND

Italy was the first European country to discover a serious outbreak of Covid-19, the infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). For this reason, policy makers around the globe have been looking carefully at the responses implemented in Italy and their effectiveness in slowing the spread of the disease. Italy struggled for weeks trying to respond to the Coronavirus pandemic and was the first Western country to implement a strict limitation of the freedom of movement of its citizens, on March 9 (Saglietto et al., 2020; Remuzzi and Remuzzi, 2020). Monitoring the virus diffusion was among the many difficulties faced by policy makers during the Covid-19 outbreak in Italy. The press conference by the Civil Protection - the office of the Italian Presidency of the Council of Ministers responsible for coordinating actions in response to the health emergency - held every evening, became a collective ritual. Families gathered in front of their televisions waiting for news and hoping to glimpse a slowdown of the epidemic.

However, days after days it became clear that the official ‘number of confirmed infected’ cases were hardly useful to monitor the Covid-19 widespread, mainly because the number of infected critically depends on the number people tested, and the latter number is to a large extent determined by criteria adopted by the health care system to recommend the test. In Italy, testing criteria changed during the epidemic. The first guidelines, issued February 27, recommended testing only individuals presenting Covid-19 symptoms and having a clear link with someone infected, or returning from China or from particular areas in Italy, mainly Lodi province (Ministero della Salute, 2020a). After the virus was declared a pandemic and Italy became the second country for number of infected after China, the Ministry approved slightly less stringent guidelines. The new guidelines, published March 9, allowed testing anyone presenting serious symptoms compatible with Covid-19 (Ministero della Salute, 2020b). Even though a number of epidemiologists and the World Health Organization suggested increasing testing efforts, the number of tests performed in Italy remained low compared with other countries, such as South Korea (WHO, 2020b; Ministero della Salute, 2020c).

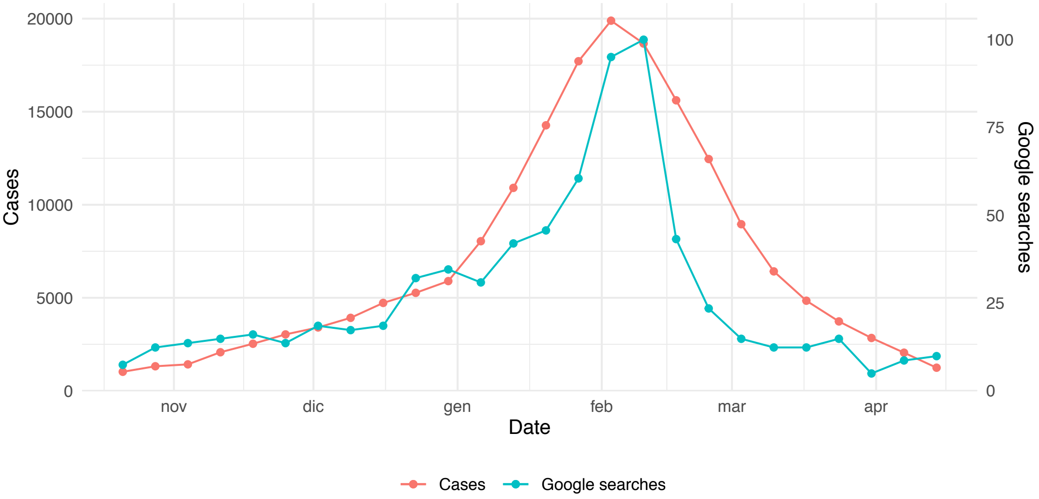
The stringency of Italian criteria led many experts to warn against the possibility that the official number of infections could be severely downwards biased. Moreover, later in March, the worsening of the health crisis put into question the viability of performing tests in many areas of the country. This is likely to have sharpened the underestimation of the diffusion, making official statistics less and less reliable over time.

This explains why the attention of media and experts has increasingly focused on the number of hospital admissions, the number of occupied beds in intensive care units, and, eventually, the number of deaths. However, even if those variables are easier to verify, they have been questioned. With the Italian health care system increasingly operating beyond its limits, the number of infected that can feasibly be admitted to hospitals is constrained. Eventually, when people die at home before a diagnosis, even the number of deaths due to Covid-19 can become unreliable. Such an awful scenario is unfortunately not implausible in areas like Lombardy, where hospitals have been overloaded for weeks.

In cases for which official statistics are not readily available, the use of big data can improve our ability to understand and predict the evolution of complex phenomena. The use of Googles queries to predict the outbreak of infections was first proposed by Ginsberg et. al (2008). Similarly, other authors have suggested the use of information retrieved from social media as a source of real-time influenza surveillance (Broniatowski et al., 2013). The idea is surprisingly simple: users suspecting an illness tend to search information about the symptoms and complications. Based on this premise, Google launched the tool Google Flu Trends in 2008, which operated until 2015 in predicting, almost in real-time, how influenza and dengue fever were spreading based on peoples queries. Although the algorithm was updated and calibrated yearly to minimize its prediction error, Google Flu Trends was criticized for having overestimated flu prevalence for more than one season (Lazer et al., 2014) and for having underestimated N1H1 influenza activity in 2009 (Cook et al., 2011). Nevertheless, a strong correlation between queries and integrated

flu surveillance data, such as the U.S. Outpatient Influenza-like Illness Surveillance Network, is found in all contributions (see also Yang et al., 2015). This type of correlation is also found in the Italian data about seasonal flu.
Figure 1 presents the strong correlation between the number of seasonal flu infections estimated by the National health institute for the 2018-19 season (Istituto Superiore di Sanità, 2020) and, on the right axis, the volume of Google queries for 'flu symptoms' in the same weeks. The National health institute uses integrated surveillance data coming from a variety of reliable sources including outpatient visits to health care providers, web surveys and information published by clinical laboratories. Contrary to what happened for the Covid-19 outbreak, in the case of seasonal flu, the number of infected is very unlikely to be incorrectly recorded by the official statistics.

Figure 1: Number of infections for seasonal flu 2018-2019 and Google searches



Data: Google Trends and Istituto Superiore di Sanità.
Note: Google Trends normalizes search volumes by setting the maximum recorded in the period considered to 100.

II SEARCHING FOR THE PEAK

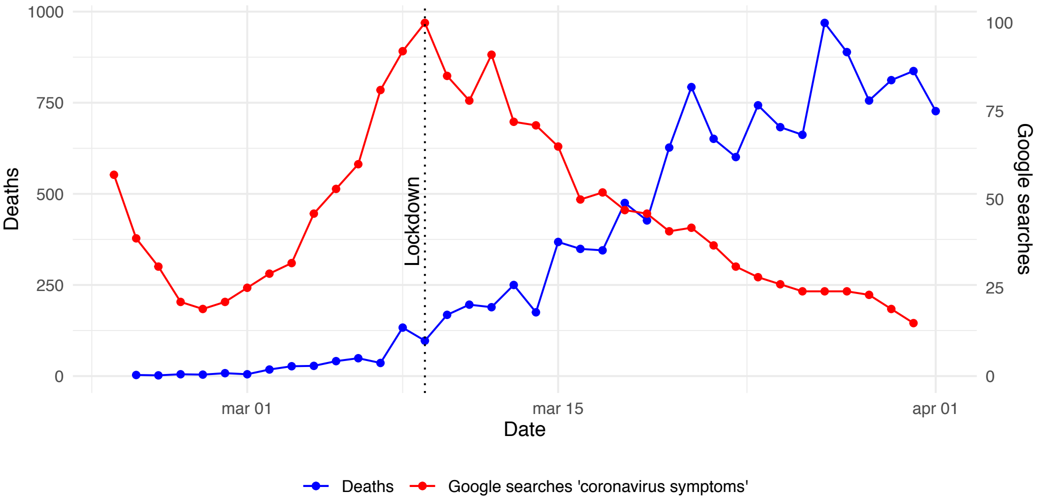
We adopt the same simple approach to predict the peak of the Covid-19 epidemic in Italy. Instead of considering the covariance of Google queries and the number of recorded infected we opt for the number of deaths. We consider such number far more reliable given the current situation.

The key weakness of adopting this approach, in the case of Coronavirus, is the massive media coverage received by the outbreak, particularly on the day Prime Minister Giuseppe Conte announced the lockdown for the entire country (March 9). Media emphasis has certainly influenced Google users queries. Many are likely to have searched information about the virus, including symptoms, without really suspecting to be infected. Moreover, once the symptomatology of a new illness is learnt some users may not need to search again when they start feeling sick.

A similar situation occurred in 2008-2009, when during the outbreak of the N1H1 flu the correlation between the number of infections and Google queries worsened. In this case, due to a much lower level of social alarm about the new flu, Google Flu Trends was found to severely underestimate the virus widespread. However, as noted by Cook et al. (2011), the correlation substantially improved in the second wave of the swine flu and, more importantly for our purpose, the model estimated on the Google users behaviour did correctly predict the peak of the epidemic also in 2009.

Figure 2 shows, on the left axis, the number of deaths per day due to Covid-19 in Italy and, on the right axis, the volume of searches for the term 'Coronavirus symptoms' in Italy. Considering that the median time from first symptoms to death is estimated in eight days in Italy (Istituto Superiore di Sanità, 2020) but is estimated in 2 weeks or longer by other sources sources (Wang et al., 2020, World Health Organization, 2020a), one could expect that at a peak in Google searches about symptoms could correspond a peak in the number of deaths after 8-14 days . However, as shown in Figure 2 the peak for Google trends is exactly March the 9, the day of the Italian lockdown, and after that the number of deaths continued to rise, suggesting that queries for 'Coronavirus symptoms' may be strongly correlated with media coverage rather than a good proxy for the number of individuals suspecting to have contracted Covid-19.

Figure 2: Number of deaths per day and Google searches for 'Coronavirus symptoms'



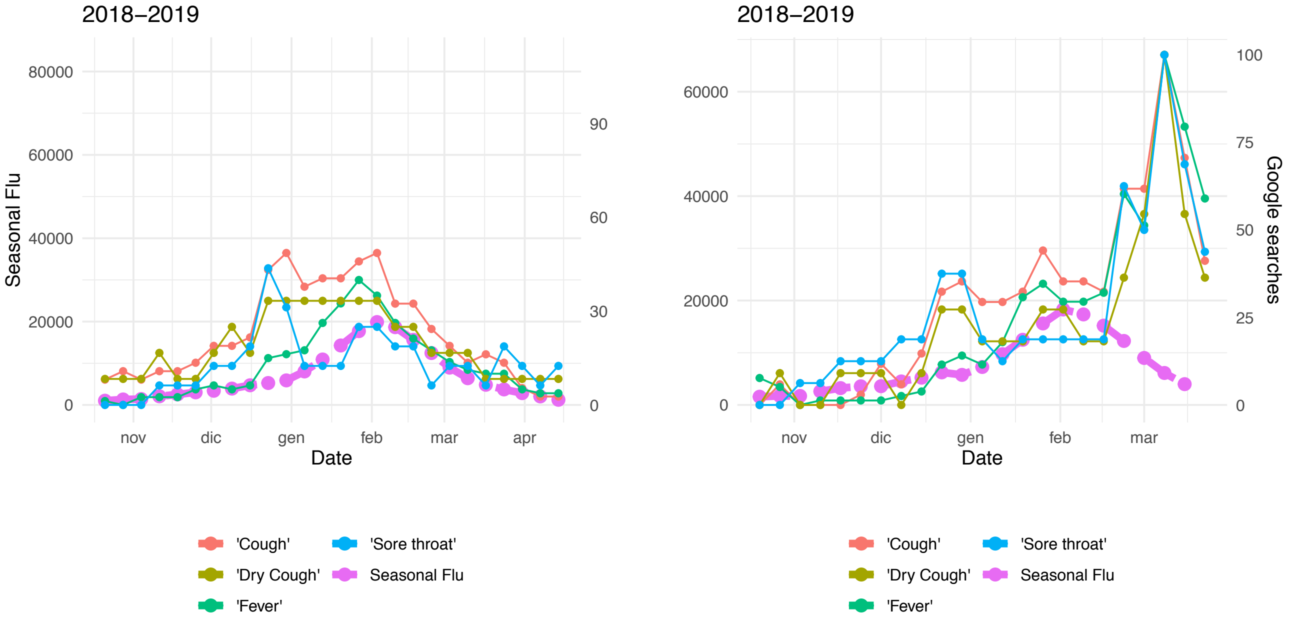
Data: Google Trends and Istituto Superiore di Sanità (Downloaded from https://github.com/pcm-dpc/COVID-19, last update April 4 2020).
Note: Google Trends normalizes search volumes by setting the maximum recorded in the period considered to 100.

A more interesting picture emerges when considering queries describing separately each symptom, without including the term 'Coronavirus'. Figure 3, left panel, shows the number of searches for the terms 'fever', 'couch', 'dry cough', 'sore throat' and the number of seasonal influenza infected recorded by the National health institute. The comparison is interesting because these symptoms are common to Covid-19 and seasonal flu. In 2018-19, left panel, symptom queries peaked between January and February and then declined. The right panel containing data for the season 2019-20 queries displays the same trend until the outbreak of Covid-19 in late February, when the searches for symptoms skyrocketed.

Focusing on the first three months of 2020 and considering the daily volume of searches, a first peak is noticeable at the end of January (with the number of searches almost identical to the same period in 2019). The trend started to decline in February and peaked again, at much higher volume of searches, around March 15, that is, one week after the national lockdown was enforced. After mid-March, queries for all symptoms have been monotonically declining. Figure 4 shows this trend together with the number of deaths for Covid-19 recorded in Italy. Note that here we added two symptoms recently reported to be strongly associated with Coronavirus: loss of smell and taste (Steves and Spector, 2020).

The timing of the search peak suggests that the unprecedented measures imposed in Italy to contain the virus have been successful in slowing the widespread of Covid-19. Moreover, if information about symptoms are searched at the earlier stage of the infection, and given an

Figure 3: weekly queries for specific symptoms of Covid-19 and seasonal flu number of infected: season 2018-2019 (left) and season 2019-2020 (right).



Data: Google Trends and Istituto Superiore di Sanità.
Note: 'Seasonal flu' refers to the number of total active cases recoded in the week. Google Trends normalizes search volumes by setting the maximum recorded in the period considered to 100. Number of searches for each symptom are normalized to have the same mean.

expected time from symptoms to death, the number of deaths in Italy could have reached its maximum at the end of March.

III A REGIONAL ANALYSIS

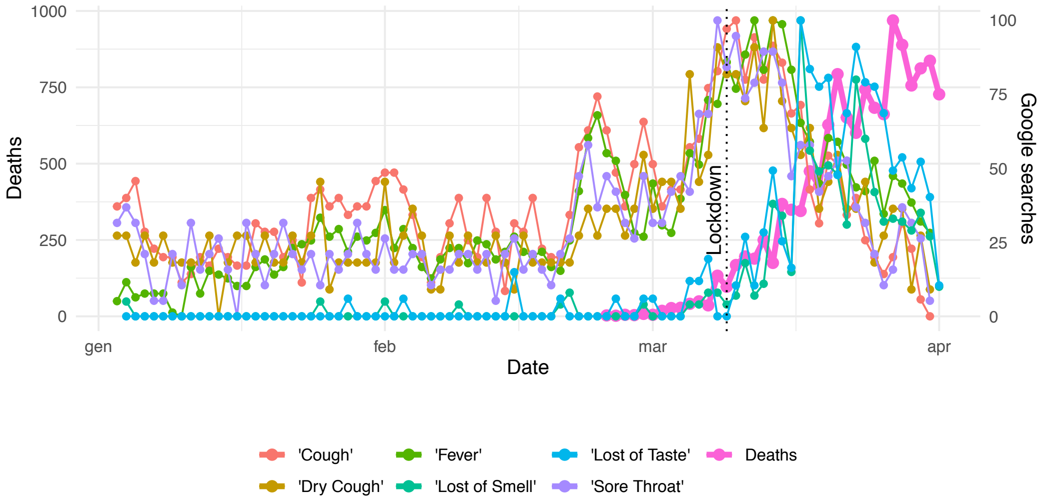
As the spread of Coronavirus in Italy started in Lombardy and Veneto and then slowly moved towards Southern regions, a certain degree of geographical heterogeneity can be exploited to show a more robust correlation between Google Trends data and Covid-19 deaths. Italy is divided in 20 regions and, after one of the most important federal reforms (Legislative Decree 56/2000), each region is responsible for the organization of its health system, following the guidelines defined by the central government. We were therefore able to test the ability of Google Trends to predict deaths using a two-way fixed effect model at the regional level.

Formally, we consider a population of 14 out of the 20 Italian regions observed daily from 2020-02-24 to 2020-03-28. We exclude from this analysis the five smallest regions (Friuli Venezia Giulia, Trentino-Alto Adige, Umbria, Basilicata, Molise, and Valle d'Aosta) because of lack of robust Google Trends available for the items selected in the considered period (daily search in smallest regions are mainly zeros and 100 - i.e., the maximum normalised). This is particularly problematic in our analysis because the analysis implicitly give the same weight to all regions. However, the sum of population in the excluded regions is 4.163 millions out of the 60.359 Italian inhabitants, so the following analysis is representative of about 93% of Italian population.

The estimated model is as follows:

$$y_{it} = \theta_i + \phi_t + \sum_{k=0}^{n} \beta_{it-k} Trends_{it-k} + \epsilon_{it} \tag{1}$$

Figure 4: Number of deaths per day and Google searches for commonly reported symptoms of Covid-19



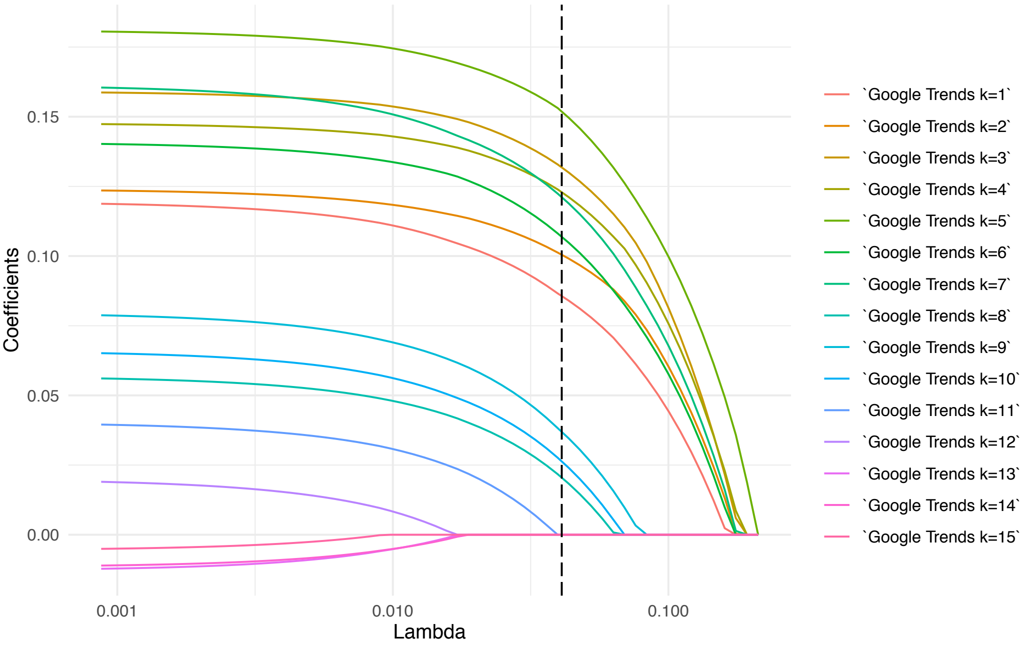
Data: Google Trends and Istituto Superiore di Sanità (Downloaded from https://github.com/pcm-dpc/COVID-19, last update April 3 2020).
Note: Google Trends normalizes search volumes by setting the maximum recorded in the period considered to 100.

Where y_{it} is the number of deaths for Covid-19 in region i the day t, θ_i is a regional fixed effect, ϕ_t is a daily fixed effect, $Trends_{it-k}$ is the volume of Google queries from region i the day t, k is the lead we test, and ϵ_{it} is the error term. We stop the model at 15 days.

The variable has been constructed as the sum of Google queries for words related to the most common symptoms of Covid-19: i.e. 'fever', 'dry cough', 'cough', 'sore throat', 'loss of sense of smell', and 'loss of sense of taste'. Because both the frequency of symptoms reported and the volume of the queries show a certain degree of heterogeneity, we normalised all the queries to the maximum so that all have the maximum at 100 as the data downloaded by Google. We also test as robustness a weighted measure of Covid-19 related queries summing commonly reported clinical symptoms weighted by the probability to be observed in positive patients, as mentioned by the European Centre for Disease Prevention and Control (ECDC, 2020): fever (47%), dry or productive cough (25%), and sore throat (16%). Results are the same. The parameters of interest are $\beta_{it-k}, k = 1, ..., n$, which capture the effect of the volume of Google queries the day $t - k$ to the number of deaths for Covid-19 the day t.
The model is regularized to minimize prediction error using the least absolute shrinkage and selection operator (Lasso). Regression coefficients are obtained searching for values that minimize the sum of squared prediction errors penalized by the sum of coefficients absolute value weighted by a penalization term λ. The larger λ the more coefficients are shrunk toward zero. The optimal λ is selected in order to minimize the mean squared prediction error out-of-sample (MSE). In practice, a large number of models are estimated letting λ to vary, for each model the MSE is estimated five times by 10-fold cross validation and stored. The value λ^* that produces the lowest average MSE is selected (Tibshirani, 1996).

Figure 5 shows the value of the coefficients estimated for lags when λ varies between 0 and 0.15. When λ is sufficiently close to zero, the vector of obtained are exactly the parameter one would obtain running an ordinary least squares regression. When λ is large enough, all coefficients are set to zero and the model estimated the number of deaths with a constant. Figure 5 shows the magnitude of the coefficients as a function of λ. The coefficients for Google searches for $k < 13$ are positive while coefficients associated to $k \geqslant 13$ are negative, all the coefficient are

Figure 5: Lasso coefficients when λ varies between 0 and 0.1



Data: Google Trends and Istituto Superiore di Sanità (Downloaded from https://github.com/pcm-dpc/COVID-19, last update April 3 2020).
Note: Google Trends normalizes search volumes by setting the maximum recorded in the period considered to 100.

monotonically decreasing in magnitude with λ. Searches in previous days (1-7) have the largest coefficients and their ranking remains quite constant with the increase of the penalisation. The vertical dashed line indicates the model that produces the lowest MSE. Coefficients for searches with a delay of more than 10 days are set to zero by the algorithm.

Table 1 reports the results of the two ways fixed effect model, in which the dependent variable is the number of deaths officially recorded for Covid-19, and the independent variable is the weighted sum of Google queries for words related to symptoms: i.e. 'fever', 'dry cough', 'cough', 'sore throat','loss of sense of smell', and 'loss of sense of taste'. Coefficients obtained with the Lasso are reported in the first column, coefficients obtained running a standard panel analysis that minimize sum of squared residuals in sample are shown in the second column. Standard errors, t-value and p-value refer to the latter model.

Regarding the fixed effect model, significant queries in explaining regional deaths for Covid-19 at time are Google searches for symptoms made on $t-k, k=1,...,7$. The queries made on $t-k, k=8,...,12$ are positively associated with deaths, while the queries made on $t-k, k=13,...,15$ are negatively associated with deaths. The association between queries and deaths for Covid-19 is not significant for $k>7$, which means that older queries ability to explaining the Covid-19 deaths is not statistically robust. In terms of Root Means Square Error out-of-sample, the best Lasso specification outperforms the fixed effect model (MSE = 0.901 with fixed effect model; MSE = 0.889 after regularization). The fact that regularizing the model we obtain substantial shrinkage of all the parameters and a reduction in the prediction error of about 1.3% suggest that the standard regression model would result in a slightly over-fitted model.

While the main rationale for regularizing our model with the Lasso is finding the model specification that minimizes prediction error, the regularization also provides an immediate method to retrieve importance scores for each attribute. Importance provides a score that indicates how useful or valuable each feature was in the construction of the final model. It is calculated using

Table 1: Lasso and ordinary least square coefficients

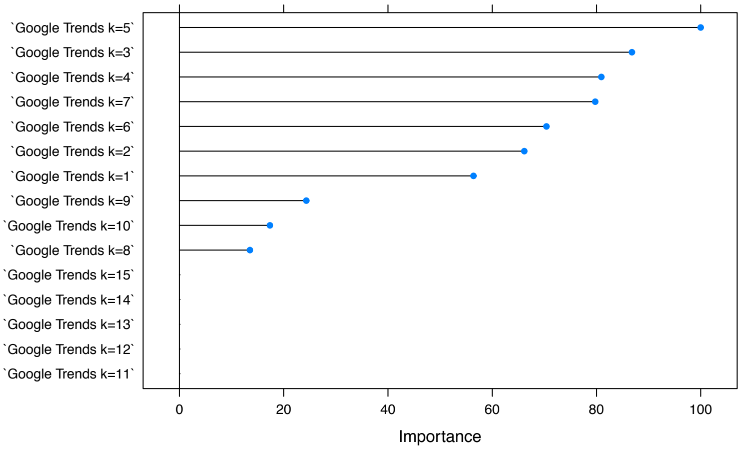
	Lasso estimate	Estimate	Std. Error	t-value	p-value	
Google Trends k=1	0.086	0.037	0.020	1.824	0.070	.
Google Trends k=2	0.100	0.039	0.020	1.921	0.056	.
Google Trends k=3	0.132	0.048	0.019	2.472	0.014	*
Google Trends k=4	0.123	0.044	0.019	2.314	0.022	*
Google Trends k=5	0.152	0.054	0.018	2.922	0.004	**
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 (Downloaded from https://github.com/pcm-dpc/COVID-19, last update April 4 2020).
Note: Note: estimates are based on 'glmnet' package (Friedman et al., 2009). 280 samples, 15 predictors. Resampling: Cross-Validated (10 fold, repeated 5 times).
Balanced Panel: n = 14, T = 19, N = 266.
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
R-Squared: 0.21725
Adj. R-Squared: 0.052843
F-statistic: 4.0523 on 15 and 219 DF, p-value: 1.6571e-06
MSE Panel Linear Model = 0.901
MSE Lasso Best Model = 0.889

the absolute value of the t-statistic for each model parameter. Overall, Figure 6 shows that in the final model the volume of queries on Google at $t-5$ are the most important in predicting the number of deaths at t. In terms of importance, queries on Google at $t-5$ are followed by queries on Google at $t-3$, $t-4$ and $t-7$ in predicting deaths at t. It is worth noting that βs associated to queries on Google with $k > 10$ are set to zero in the final model, meaning that older symptom searches on Google (more than 10 days before) are not predictive of daily number of deaths. These results are consistent with the Fixed effect model: the most important features shown in Figure 6 are also the significant regressors in Table 1. As a robustness check we test the same specification using number of serious/critical cases of Covid-19 as dependent variable and we obtained similar results. Table A1 in the Appendix reports an even stronger ability of Google queries to predict Covid-19 critical cases. Queries made on $t-k, k = 8, ..., 12$ have a t-value larger than 3.5, moreover, the MSE of the regularized model is reduced to 0.844.

Overall, the regional analysis suggests that the time between searches and death is between 3 and 10 days. This is shorter than what appears in the descriptive analysis based on national data, but still in the range of time depicted by literature. Possible explanations for the differences between the national and the regional analysis could be due to variations in the quality of the regional health care systems (Lagravinese et al. 2019), regional differentiations in access to the Internet and other related services (Greco et al. 2019), cultural factors regarding individual attitudes towards the use of search engines for health self-assessment searches, and regional differentiations in timing and intensity of the Covid-19 spread. In particular, it should be noted that while some northern regions like Lombardy already had cases and deaths on February 24 (specifically, 166 cases and six deaths in Lombardy), the majority of southern regions remained

Figure 6: Feature Importance to predict number of deaths



Data: Google Trends and Istituto Superiore di Sanità (Downloaded from https://github.com/pcm-dpc/COVID-19, last update April 4 2020).
Note: Estimates based on 'glmnet' package (Friedman et al. 2009). 280 samples, 15 predictors. Resampling: Cross-Validated (10-fold, repeated 5 times) The absolute value of the t-statistic for each model parameter is used to estimate importance.

with zero cases until the beginning of March and at the last observation (April 2) they remain with low infection rates. This last factor, together with the huge media coverage of the phenomenon, may have caused a disconnection between searches and cases of Covid-19 in that part of the country.

IV CONCLUSIONS

Among the problematic aspects of the early Covid-19 outbreak in Italy, the difficulties of institutions to provide real-time and reliable information about the spread of the virus stands out. Lack of precise information represented a major issue in a moment of crisis in which effective decisions to respond to the pandemic had to be made immediately. This paper explored the possibility to use Google Trends data to predict the peak of the number of deaths and critical cases due to Coronavirus in Italy.

Although searches for the term 'Coronavirus symptoms' seem to be to a large extent determined by media coverage rather than by virus diffusion, the analysis of more specific queries about commonly reported Covid-19 symptoms appears predictive of the number of deaths attributed to Covid-19 later in time. We estimated a prediction model based on Google queries controlling for regional fixed characteristics and time fixed effects. The model was tuned by estimating a least absolute shrinkage and selection operator that shows a systematic positive relationship between number of searches for symptoms and number of deaths. The same model showed an even stronger predictive ability when used to explain critical cases recorded in the Italian hospitals. Because the trends in queries for symptoms have all been monotonically decreasing since mid-March, the analysis suggests that the number of deaths in Italy should peak and start to decline with the end of the same month.

The possibility to predict outbreaks based on web searches of Google users to supplement epidemiological models has been severely criticized in the last decade. Nevertheless, during crisis situations where institutions struggle to operate normally and the reliability of official statistics is questioned, supplementing official data sources with data obtained from Google Trends appear a promising option.

REFERENCES

- Broniatowski D.A., Paul M.J., Dredze M. (2013). National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. PLoS ONE 8(12): e83672.
- Cook S., Conrad C., Fowlkes A.L., Mohebbi M.H. (2011). Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. PLoS ONE 6(8): e23610.
- European Centre for Disease Prevention and Control - ECDC (2020). Disease background of COVID-19. Link https://www.ecdc.europa.eu/en/2019-ncov-background-disease, last access: March, 28th 2020.
- Friedman, J., Hastie, T., Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. R package version, 1(4).
- Ginsberg J., Mohebbi M.H., Patel R.S., Brammer L., Smolinski M.S.. (2008). Detecting influenza epidemics using search engine query data. Nature 457: 1012-10155.
- Greco, S., Ishizaka, A., Matarazzo, B., Torrisi, G. (2018). Stochastic multi-attribute acceptability analysis (SMAA): an application to the ranking of Italian regions. Regional Studies, 52(4), 585-600.
- Istituto Superiore di Sanità (2020). Rapporto della sorveglianza integrata dellÕInfluenza 2017-2018, Link: https://www.epicentro.iss.it/influenza/pdf/FluNewsMetodi.pdf
- Istituto Superiore di Sanità (2020). Report sulle caratteristiche dei pazienti deceduti positivi a COVID-19 in Italia - 17 Marzo 2020. Link: https://www.epicentro.iss.it/coronavirus/bollettino/Report-COVID-2019_17_marzo-v2.pdf
- Lagravinese, R., Liberati, P., Resce, G. (2019). Exploring health outcomes by stochastic multicriteria acceptability analysis: An application to Italian regions. European Journal of Operational Research, 274(3), 1168-1179.
- Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science, 343 (6176): 1203 -1205.
- Ministero della Salute. (2020a). Documento relativo ai criteri per sottoporre soggetti clinicamente asintomatici alla ricerca dÕinfezione da SARS-CoV-2attraverso tampone rinofaringeo e test diagnostico. 2020 February 27. Link http://www.trovanorme.salute.gov.it
- Ministero della Salute. (2020c). COVID-19. Aggiornamentodella definizione di caso. 2020 March 9. Link https://www.fnopi.it/wp-content/uploads/2020/03/Circolare_9_marzo_2020.pdf
- Ministero della Salute. (2020c). Covid-19 - Situazione in Italia - Situazione italiana al 26 Marzo. Link http://www.salute.gov.it/imgs/C_17_pagineAree_5351_36_file.pdf
- Remuzzi A. and Remuzzi G. (2020). COVID-19 and Italy: what next? The Lancet, Published on line first:March 13, 2020.
- Saglietto A., DÕAscenzo , F., Biondi Zoccai G., De Ferrari M.G. (2020). COVID-19 in Europe: the Italian lesson. The Lancet. Published on line: March 24, 2020.
- Steves C. and Spector T. (2020). COVID Symptom Tracker Research Update. March 30, 2020 Webinar King's College London. Link https://covid.joinzoe.com/post/research-update-april
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.
- Wang W. Jianming T. Fangqiang W. (2020). Updated understanding of the outbreak of 2019 novel coronavirus (2019?nCoV) in Wuhan, China. Journal of Medical Virology, 92, (4): 441-447.

- World Health Organization. (2020a). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). 16-24 February 2020. Link https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf
- World Health Organization. (2020b). Laboratory testing strategy recommendations for COVID-19. Interim guidance. 22 March 2020 Link https://apps.who.int/iris/bitstream/handle/10665/331509/WHO-COVID-19-lab_testing-2020.1-eng.pdf
- Yang S., Santillana M., Kou, S. C. (2015). Accurate influenza epidemics estimation via ARGO. Proceedings of the National Academy of Sciences. 112 (47): 14473-14478

A APPENDIX

Table A1: Lasso and ordinary least square coefficients, dependent variable number of serious/critical cases

	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 (Downloaded from https://github.com/pcm-dpc/COVID-19, last update April 4 2020).

Note: Note: estimates are based on 'glmnet' package (Friedman et al., 2009). 280 samples, 15 predictors. Resampling: Cross-Validated (10 fold, repeated 5 times). Balanced Panel: n = 14, T = 20, N = 280.

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

R-Squared: 0.40715

Adj. R-Squared: 0.28704

F-statistic: 10.6218 on 15 and 232 DF, p-value: ¡ 2.22e-16

MSE Panel Linear Model = 0.855

MSE Lasso Best Model = 0.844