Dai Big-Data un motivo di fiducia in più

Categories : Articoli, Istituzioni e regole

Tagged as : Giuliano Resce, Menabò n. 121/2020, Paolo Brunori

Date : 27 Marzo 2020

Ogni giorno alle 18 la conferenza stampa del Dipartimento della Protezione Civile è diventata un rito propiziatorio collettivo, nell'attesa che l'andamento dell'epidemia inizi a dare segnali di rallentamento. Purtroppo però, con il passare del tempo, è apparso chiaro che i dati del contagio sono difficilmente utilizzabili per farsi un'idea precisa riguardo l'andamento della pandemia. Il motivo fondamentale è che il numero di contagi dipende dal numero di tamponi effettuati e questo dipende in primo luogo dai criteri utilizzati per decidere se eseguire il test o meno. Criteri che sono cambiati nel corso dell'epidemia, e che, malgrado si parli da tempo di aumentare fortemente il numero di tamponi, rimangono ad oggi molto restrittivi. Se soltanto i soggetti che rispettano i criteri stringenti per il tampone entrano effettivamente nelle statistiche, è possibile che il numero degli infetti sia sottostimato. A questo si deve aggiungere il fatto che, con l'aggravarsi della situazione, le risorse disponibili per effettuare test sono diventate insufficienti. Questo ha verosimilmente accentuato il problema di sottostima rendendo il numero ufficiale di nuovi infetti sempre meno affidabile nel tempo.

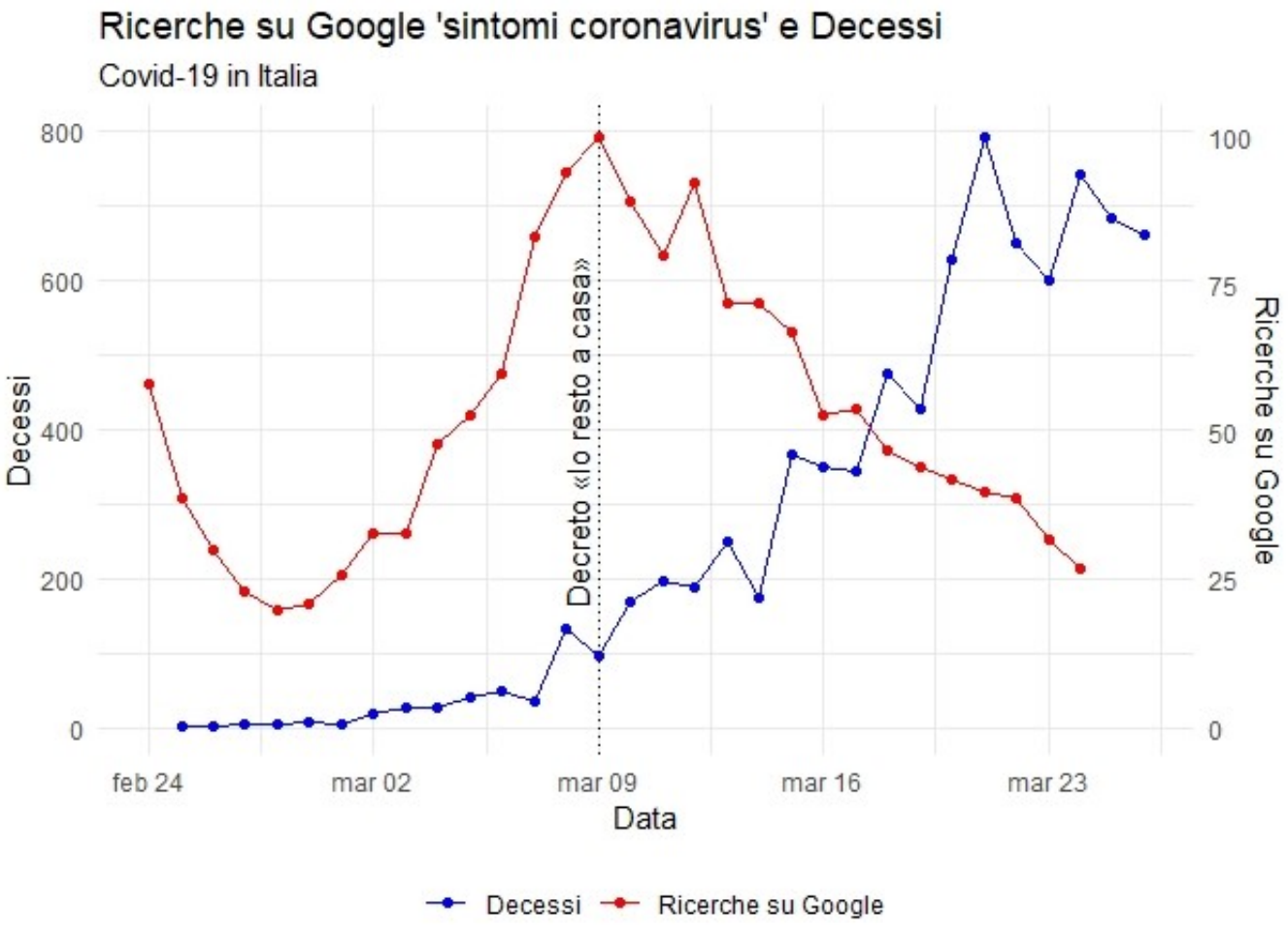
Alcuni analisti hanno quindi preferito concentrarsi sul numero di ricoveri, sul numero di ricoveri in terapia intensiva e sul numero dei decessi a causa del Covid-19. Si tratta certamente di numeri più affidabili, ma che rischiano di diventare difficilmente interpretabili nel momento in cui il sistema sanitario, ormai al limite della sua capacità, non riesce ad operare normalmente. Se non ci sono posti disponibili in corsia la crescita del numero dei ricoverati tenderà a rallentare anche se l'epidemia sta accelerando. Questo scenario è particolarmente verosimile in Lombardia, dove gli ospedali operano da tempo in emergenza.

Nei casi in cui le statistiche ufficiali non possono garantire il consueto standard di affidabilità, l'utilizzo dei *big data* per migliorare la nostra capacità di comprendere l'evoluzione di un fenomeno rappresenta un'opportunità preziosa. La possibilità di utilizzare le ricerche degli utenti sulla rete per avere una quantificazione dell'evoluzione di una malattia infettiva è stata per la prima volta mostrata da Jeremy Ginsberg e i suoi colleghi ricercatori di Google in una pubblicazione su Nature (Ginsberg, Mohebbi, Patel, Brammer, Smolinski, "Detecting influenza epidemics using search engine query data", *Nature,* 2008). Il meccanismo è semplice: le persone sospettano di avere sintomi tipici di una malattia e cercano conferma sulla rete utilizzando i motori di ricerca.

Questo comportamento è rispecchiato anche in Italia, ogni anno, durante il periodo dell'influenza stagionale. L'andamento delle ricerche delle parole "sintomi dell'influenza" su Google si muove di pari passo con il numero di infetti registrati dall'Istituto Superiore di Sanita?, che, nel caso dell'influenza stagionale, sono ottenuti integrando una serie di fonti differenti di dati e possono essere considerati affidabili. Il grafico sotto riporta l'andamento della variazione settimanale del numero di contagiati della sindrome influenzale durante la stagione invernale 2017-2018 sovrapposto al numero di ricerche su Google riguardo ai sintomi.

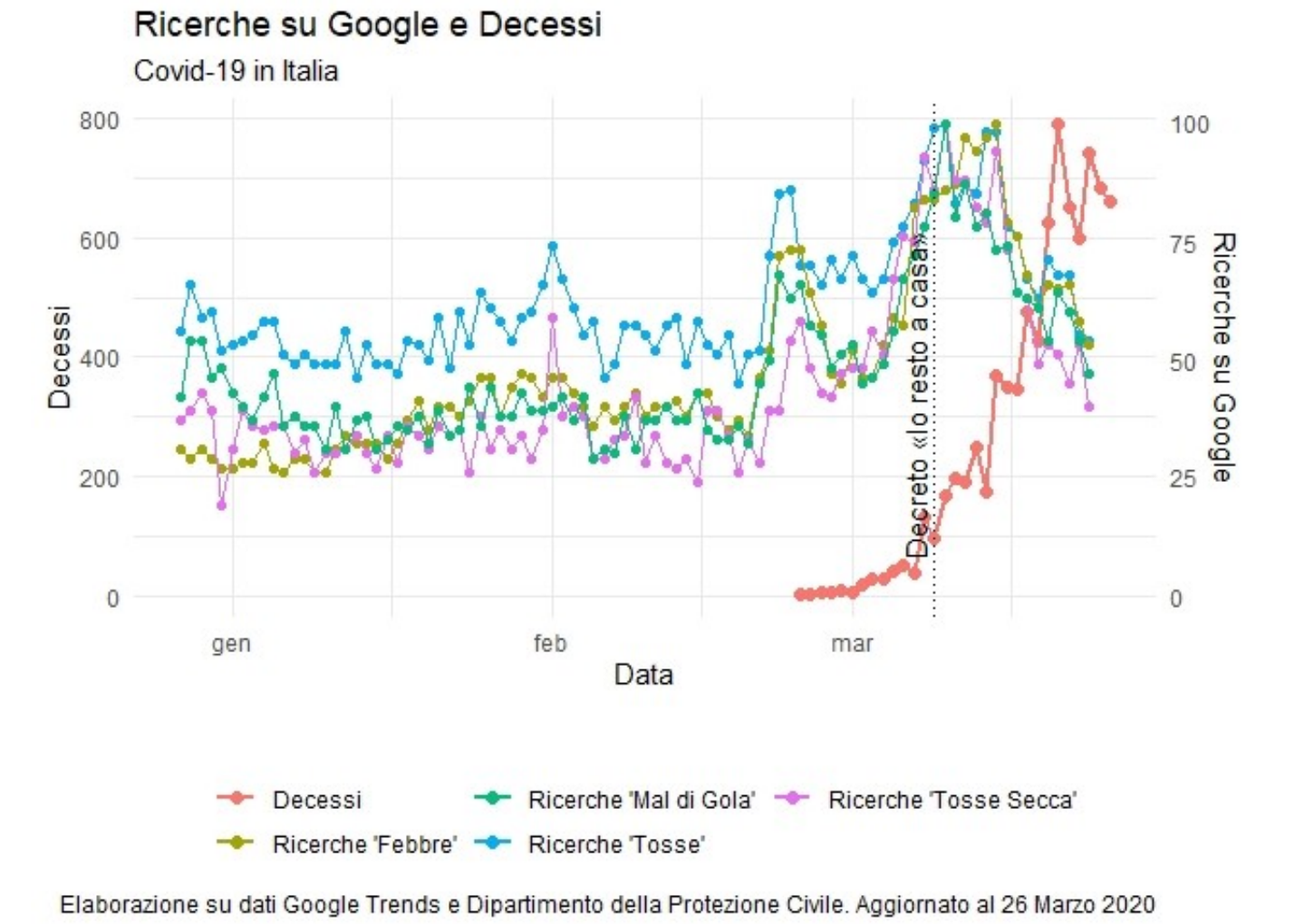
Utilizzare lo stesso approccio per l'epidemia da Covid-19 è possibile. Certo è importante tener conto che l'esposizione mediatica che ha avuto l'epidemia durante queste settimane può aver indotto molti a ricercare una descrizione online dei sintomi anche senza sospettare di essere malati, e questo in qualche misura limita l'affidabilità dell'esercizio predittivo. Una modifica della correlazione fra ricerche internet riguardo l'influenza e numero di infetti è stata ad esempio dimostrata per gli Stati Uniti in uno studio relativo alla stagione influenzale 2009, nel momento in cui si è iniziata a diffondere l'influenza N1H1, la così detta influenza suina (Cook, Conrad, Fowlkes, Mohebbi, "Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic," *PLoS ONE*, 2011). È chiaro che le persone, anche in assenza di uno stato di malessere, sentendo parlare di una nuova malattia, si interessino ai sintomi e effettuino ricerche su internet, tanto più se sono costrette a passare molto tempo in casa. Allo stesso tempo, anche nel caso dell'insorgere dell'influenza suina, Cook e colleghi sottolineano come si sia continuata a registrare una forte correlazione fra ricerche attraverso Google e diffusione influenzale. Quindi, con un certo grado di cautela, è possibile pensare che un aumento delle ricerche online riguardo termini collegati al virus Covid-19 possa essere fortemente collegato al numero di persone che sospettano di essersi ammalate.

Il grafico sotto riporta l'andamento del numero di ricerche "sintomi coronavirus" e il numero di decessi da Covid-19 in Italia. Considerando che il tempo che trascorre fra i primi sintomi e il decesso per Coronavirus è fra gli 8 e i 14 giorni, a seconda degli studi (Istituto Superiore di Sanità, 2020; Wang, Jianming, Fangqiang, "Updated understanding of the outbreak of 2019 novel coronavirus (2019?nCoV) in Wuhan, China", *Journal of Medical Virology*, 2020), è possibile che a un picco di ricerche riguardo ai sintomi sui motori di ricerca corrisponda, a distanza di circa 10 giorni, un picco nei decessi. Questo è quello che effettivamente si osserva nei dati degli ultimi giorni.



Elaborazione su dati Google Trends e Dipartimento della Protezione Civile. Aggiornato al 26 Marzo 2020

Il picco delle ricerche è verosimilmente anche collegato all'esposizione mediatica della malattia. È difficilmente casuale il fatto che il massimo numero di ricerche corrisponda al giorno di annuncio del decreto "io resto a casa" la sera del 9 marzo. Per questo motivo abbiamo ripetuto l'analisi per le ricerche di termine descrittivi dei sintomi da Coronavirus evitando di inserire il nome del virus. Il grafico sotto riporta l'andamento dei termini: "febbre", "tosse", "tosse secca", "mal di gola". Tutti i valori del volume delle ricerche sono normalizzati in modo che abbiano valore massimo identico. Storicamente queste ricerche hanno un picco fra gennaio e febbraio quando la sindrome influenzale stagionale raggiunge il massimo della diffusione. Quest'anno il picco per tutte i quattro criteri di ricerca considerati si raggiunge tra il 10 (per "tosse", "tosse secca", "mal di gola") e il 15 marzo (per "febbre"). Dal 15 marzo in poi, il volume di ricerche per tutti i termini considerati è nettamente decrescente. Questo potrebbe suggerire che in realtà il picco dei sintomi è stato raggiunto un po' più tardi di quanto non suggerirebbe la ricerca "sintomi Coronavirus", ma che comunque è da oltre dieci giorni in decrescita.



La disaggregazione regionale dei dati mostra una forte eterogeneità. Considerando la Lombardia, che ad oggi è la regione più colpita, i due picchi delle ricerche su Google "sintomi coronavirus" si osservano il 9 e il 12 marzo, che sono gli stessi giorni in cui si osservano i picchi a livello nazionale. Allo stesso tempo il giorno con il maggior numero di decessi in Lombardia è il 21 marzo; la distanza tra il picco di ricerche e il picco di deceduti è quindi di circa 10 giorni. Il dato delle ricerche su Google ha una tendenza assimilabile al dato nazionale per la gran parte delle restanti regioni. Una eccezione interessante a questa tendenza generale si registra nelle Marche. Le Marche hanno i due picchi di ricerche su Google che ritardano di 1-2 giorni rispetto al dato nazionale e rispetto al dato delle altre regioni: 10 e 14 marzo. A questi due picchi ritardati corrisponde un ritardo anche nel picco del numero di decessi, che nelle Marche si registra il 25 marzo. In altre regioni invece il picco di ricerche su internet segue quello nazionale mentre il numero di decessi, in alcuni casi, ha continuato a salire.

Soltanto i prossimi giorni potranno confermare se, anche nel caso di una nuova malattia sconosciuta come il Covid-19, il comportamento degli utenti sulla rete rappresenta un possibile strumento di supporto al monitoraggio del contagio o meno. L'uso di dati non strutturati provenienti da fonti eterogenee e non accreditate, come i Google Trends, non potrà mai sostituirsi ai dati ufficiali prodotti dalle autorità competenti. Tuttavia, al netto delle problematiche di qualità del dato, problemi di privacy e di trasparenza di cui i big-data sono intrisi, questo esercizio mostra che l'esplorazione di fonti alternative può rappresentare una opportunità per aggiungere conoscenza in tempi eccezionalmente rapidi allo studio di fenomeni nuovi e dalla difficile interpretazione.