Joint Discussion Paper
Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 17-2020

Mohammad Reza Farzanegan, Mehdi Feizi and Saeed Malek Sadati

Google It Up! A Google Trend-based Analysis of COVID-19 Outbreak in Iran

Google It Up! A Google Trends-based analysis of COVID-19 outbreak in Iran

Mohammad Reza Farzanegan[a]; Mehdi Feizi[b], Saeed Malek Sadati[b]

[a] Philipps-Universität Marburg, Center for Near and Middle Eastern Studies (CNMS), Economics of the Middle East Research Group, Marburg, Germany & CESifo (Munich), ERF (Cairo) (farzanegan@uni-marburg.de).

[b] Ferdowsi University of Mashhad, Faculty of Economics and Administrative Sciences, Mashhad, Iran.

Abstract

Soon after the first identified COVID-19 cases in Iran, the spread of the new Coronavirus has affected almost all its provinces. In the absence of credible data on people's unfiltered concerns and needs, especially in developing countries, Google search data is a reliable source that truthfully captures the public sentiment. This study examines the within province changes of confirmed cases of Corona across Iranian provinces from 19 Feb. 2020 to 9 March 2020. Using real-time Google Trends data, panel fixed effects, and GMM regression estimations, we show a robust negative association between the intensity of search for disinfection methods and materials in the past and current confirmed cases of the COVID-19 virus. In addition, we find a positive and robust association between the intensity of the searches for symptoms of Corona and the number of confirmed cases within the Iranian provinces. These findings are robust to control for province and period fixed effects, province-specific time trends, and lag of confirmed cases. Our results show how not only prevention could hinder affection in an epidemic disease but also prophecies, shaped by individual concerns and reflected in Google search queries, might not be self-fulfilling.

Keyword: Google Trends, COVID-19, Iran, epidemic disease

1. Introduction

Following the Coronavirus outbreak in China, the first known cases of COVID-19 in Iran were officially confirmed on February 19, 2020. To lessen the speed of the outbreak, many public events (such as Friday prayers, concerts, and sports competitions) and public places (including cinemas, schools, and universities) were closed, and government office hours were cut in several provinces. Nevertheless, in just about two weeks since the first identified cases in Qom city, the spread of the new Coronavirus has affected almost all provinces of Iran.

Having the opportunity to diagnose illness early and respond rapidly could lessen the influence of pandemic disease and increase the health safety of people. However, worldwide, there is a lack of real-time and reliable data on human health behavior to predict the outbreak of epidemic diseases. Keyword-driven internet searches appears to be a worthy information base to construct proxies of social indicators in advance compared to official statistics (Di Bella et al., 2018). In the absence of credible data on people's unfiltered concerns, wants and needs, especially in developing countries, the Google search data is a reliable and unintentionally created source that candidly encapsulates the public sentiment. It shapes a pattern of information that otherwise would have not been visible and provides an insightful and intuitive picture of the extent to which individuals are involved with everyday issues, such as which movie to watch, which stock to invest, and how to prevent epidemic diseases.

Google Trends shows great potential as a method for vigorous and delicate surveillance of epidemics and diseases with high prevalence (Carneiro & Mylonakis, 2009). It has been used widely to detected influenza epidemics. Google Flu Trends (GFT) was aggregating Google Search query data, from 2008 to 2015, for more than 25 countries to predict outbreaks of flu. The concept behind this was that huge numbers of Google search queries can be examined to disclose if flu-like illness exists in a region. Estimations indicated that there was a high correlation between the number of Google queries and official influenza surveillance data.

The purpose of this study is to examine the recent development of confirmed cases of Corona across and within provinces of Iran. We extract the information which the Iranians have searched for regarding different aspects of Corona virus in each province over time. The extracted information from Google Trends show that in areas where the people are actively looking for sanitation, disinfection methods, and materials, we also observe lower records of confirmed cases of Corona. The higher frequency of searches for disinfection methods may capture the higher levels of awareness of households in dealing with Coronavirus.

We also find that the provinces with a higher frequency of searches for symptoms of Coronavirus are recording higher levels of confirmed cases. The higher frequency of searches for symptoms may capture the higher levels of concerns due to increased levels of Coronavirus in affected regions. To control for possible omitted variables, we are controlling for province and period fixed effects. In addition, our results are robust to control for province specific time trend which is controlling for those factors which are different across provinces and are changing over time but, due to lack of data with daily frequency, we cannot control them explicitly in the model.

Recently studies have utilized Google Trends to predict the novel Coronavirus pandemic (Hu et al., 2020; Husnayain et al., 2020; Li et al. 2020) and analyze the effects of fears about it on current economic sentiment (Fetzer et al., 2020). Our work is different from other similar works in some respects. To the best of our knowledge, it is the first study which assesses the outbreak of the COVID-19 pandemic in a developing country using search queries in a local language, rather than English. Moreover, we use different possible keywords that people are expected to search to prevent the illness. The rest of the paper is organized as follows. Section 2 provides a brief review of literature. Data and method are presented in Section 3. Results are shown and discussed in Section 4. Section 5 concludes.

2. Review of literature

In the last decade, Google Trends has been used in an extensive range of academic research such as health, economics, finance, tourism and more[1]. However, the common trait of most papers citing Google Trends is the need for real-time data for timely analysis of the variables' patterns. The seminal work of Ginsberg et al. (2009),

[1] For instance, Google Trends has been also used to predict near-term values of economic indicators, such as automobile purchases (Choi & Varian, 2012), unemployment claims, travel planning and consumer confidence (Choi & Varian, 2012), cinema admissions (Hand & Judge, 2012), fashion consumer behavior (Silva et al., 2019), tourists' arrivals (Dergiades et al., 2018), oil consumption (Yu et al., 2019), and social interest in burnout (Aguilera et al., 2019). Moreover, it has applied to estimate suicide occurrence (Kristoufek, Moat, & Preis, 2016) and quantify trading behavior in financial markets (Preis, Moat, & Stanley, 2013).

published in Nature, is the first study to utilize Google Trends with such an approach in health. They found a high correlation between the frequency of certain web search queries and the percentage of patients with influenza-like symptoms, additionally confirming that Google Trends can detect influenza diffusion one week or two weeks earlier than the Centers for Disease Control and Prevention (CDC).

Ginsberg et al. (2009) inspired subsequent scholars to apply GFT to predict the behavior of epidemic infections. For example, Cook et al. (2011) evaluated the accuracy of original and updated GFT models during a non-seasonal influenza outbreak, Dugas et al. (2012) assessed the City-level GFT as a surveillance tool and Olson et al. (2013) reassessed the ability of GFT models in the detection of seasonal and pandemic influenza. Among scholars outside the US, Kang et al. (2013) and Verma et al. (2018) noticed that Google Trends can be utilized as a source of data for influenza surveillance respectively in south China and India. However, despite a variety of local languages used as primary languages by devices like mobile or PCs in these two countries, they employed English to retrieve the search results.

In other epidemic diseases and with a similar methodology, Chan et al. (2011), Althouse et al. (2011), and Husnayain et al. (2019) concentrated on the prediction of Dengue epidemics based on the web search query data; Lu et al. (2019) predicted epidemic Avian Influenza and Teng et al. (2017) developed a dynamic model for surveillance Zika virus.

3. Data and model specification

Because the Coronavirus stays on different surfaces for a long time, the best way to prevent COVID-19 is to disinfect surfaces, regularly wash hands, and avoid contact between the hands, nose, mouth and eyes. Although the use of a mask does not prevent the virus, it is recommended to use in crowded areas. Like many countries involved with the virus, concerns about its spread in Iran have also led to increased demand for sanitary equipment such as surface cleaners, hand sanitizers and face masks. As a result, these products became scarce and their prices rose sharply in Iran.

Our key explanatory variable is based on Google Trends for a selection of relevant keywords which are searched for by individuals with reference to Corona, namely "Corona Symptoms", "Masks", "Disinfection", and "Corona". These are represented by an index of 0-100, where 100 stand for the day with the highest numbers of search queries for mentioned terms across 31 provinces over 19 days.

Google is indeed the most used search engine in Iran. According to StatCounter Global Stats (2020), the search engine market share of Google in Iran in 2019 was more than 99%. To explain the within province pattern of confirmed cases of COVID-19, we extract information about residents of each provinces search activities in Google with reference to Corona.

The dependent variable is confirmed COVID-19 cases in each province over time. Ministry of Health and Medical Education in Iran officially releases daily statistics on the number of COVID-19 cases, in the province level, and deaths, in total.

The World Health Organization (WHO) has approved these statistics[2] and announces them daily alongside statistics from other countries[3].

We expect to see lower levels of confirmed cases in provinces where there is greater interest in disinfection materials and methods to combat the spread of Corona virus. All variables are normalized by the size of the population of a province and are in logarithmic transformation. To control for province specific characteristics which may be relevant for extension of Corona cases, such as geographic location, cultural differences especially in social interactions, religiosity and perception of risk, we include province fixed effects.

In addition, time fixed effects are taken in to account. To control any other possible factors related to provinces which may change over time, we have also controlled for province specific time trend. A one period lag of search outcomes in Google Trend is used in estimations. Furthermore, a one period lag of the dependent variable, which is number of confirmed cases of Corona, is included in the right hand side of model. Inclusion of the lag of dependent variable in fixed effects regressions may call for application of Generalized Methods of Moments (GMM).

We apply both (one and two steps) first difference and system GMM methods[4] (see Arellano and Bover, 1995 and Blundell and Bond, 1998). The GMM method

[2] https://en.irna.ir/news/83700376/

[3] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[4] Baltagi (2008) shows that the system GMM produces more efficient and precise estimates compared to differenced GMM by enhancing precision and decreasing the finite sample bias.

differences the model to get rid of any time-invariant country specific factors and may eliminate any endogeneity due to the correlation of these time invariant country characteristics and the right hand side regressors (Baltagi, Demetriades, & Law, 2009).

We use robust standard errors which are consistent with panel-specific autocorrelation and heteroskedasticity. We treat the lag of the dependent variable as potentially endogenous variable and use two lags as internal instruments. The model specification is as following:

$$\log of\ CASES_{it} = \delta i + \theta t + \beta_1 \cdot \log of\ \text{Corona Symptoms}_{it-1} + \beta_2 \cdot \log of\ \text{Corona Masks}_{it-1} + \beta_3 \cdot \log of\ \text{Disinfection for Corona}\ _{it-1} + \beta_4 \cdot \log of\ \text{Corona}\ _{it-1} + \beta_5 \cdot \log of\ CASES\ _{it-1} + \varepsilon_{it}$$

We expect to observe $\beta_3 < 0$ and $\beta_1 > 0$.

4. Empirical results

Table 1 shows the results of panel regression with province and period fixed effects. In models 1 to 8 we check the association between the intensity of searches for each of the keywords related to Corona in the past period and the current level of confirmed cases of Corona within provinces. In Models 1 to 4, we control for province fixed effects while Models 5 to 8 also include the time fixed effects. In Model 9, we control for intensity of all searched key words related to Corona in one specification. Model 10 includes the lag of confirmed cases to check the possible persistence in identified cases of Corona in province. In Model 11, in order to control for other possible time varying factors related to each province, we control for province specific time trend.

We can observe two robust and consistent findings: The first finding is the negative association between earlier searches for disinfection methods, and materials related to Corona by people in province, and current records of confirmed cases. Across different specifications, a one percent increase in the intensity of searches for disinfection methods and materials by individuals is associated with a reduction of 0.23% to 1.17% in confirmed cases within provinces, controlling for province and time fixed effects, lag of confirmed cases and province specific time trend. This shows that the higher levels of awareness of individuals to protective measures, the lower the risk is of infection by Corona virus. Free flow of information and public education is an effective tool to dampen the number of confirmed cases.

Second, we also observe a consistent and robust positive association between past intensity of searches for symptoms of Corona and current records of Corona. This may refer to the increasing concerns of individuals regarding Coronavirus and its consequences. The higher the levels of such concerns, the more people may look for symptoms of disease. In Model 11, approximately 88% of within province variation in the logarithm of confirmed cases (adjusted for the size of population) is explained by explanatory variables.

Table 1. Panel fixed effects regressions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
						Dependent variable: log (CASES)					
	with province fixed effects (fe)				with province & time fe				with province & time fe & all search cases	with province & time fe & all search cases + lag of DV	with province & time fe & all search cases + lag of DV + province specific time trend
L.log_search Corona	-1.354				-0.746				-1.273**	-0.846	0.099
	(-1.62)				(-1.30)				(-2.38)	(-1.12)	(0.07)
L.log_search symptom		0.610*				0.885***			1.187***	1.770***	2.189***
		(1.79)				(3.00)			(4.30)	(4.57)	(9.52)
L.log_search mask			0.460*				-0.136		-0.321	0.150	0.041
			(1.78)				(-0.89)		(-1.58)	(0.66)	(0.17)
L.log_search disinfection				-0.235**				-0.429*	-0.362*	-0.797**	-1.170***
				(-2.56)				(-1.99)	(-1.79)	(-2.68)	(-3.47)
L.log_CASES										0.083	-0.120
										(0.62)	(-0.64)
Obs.	108	108	108	102	108	108	108	102	102	65	65
R-sq	0.07	0.08	0.05	0.03	0.52	0.55	0.51	0.54	0.61	0.72	0.88

Robust t statistics are in ().L refer to one period lag. * Significantly different from zero at 90% confidence. ** Significantly different from zero at 95%confidence. *** Significantly different from zero at 99% confidence.

Table 2. Panel GMM regressions

	(1)	(2)	(3)	(4)
	Dependent variable: log (CASES)			
	diff GMM (one step)	diff GMM (two steps)	sys GMM (one step)	sys GMM (two steps)
L.log_search symptom	0.322	0.661	0.459	0.617
	(0.65)	(1.06)	(0.82)	(0.95)
L.log_search mask	0.548	0.166	0.357	0.192
	(1.09)	(0.26)	(0.70)	(0.31)
L.log_search disinfection	-0.865***	-1.038***	-0.561***	-0.511**
	(-4.05)	(-6.19)	(-4.09)	(-2.38)
L. log_CASES	0.482**	0.408	0.765***	0.733**
	(2.46)	(1.72)	(4.06)	(2.64)
Obs.	39	39	65	65
Hansen test of overid. (p-value)	0.963	0.963	0.789	0.789
AR(2) (p-value)	0.442	0.375	0.327	0.303

Robust t statistics are in ().L refer to one period lag. ** Significantly different from zero at 95%confidence. *** Significantly different from zero at 99% confidence.

In Table 2, we report the estimation results from both first difference and system GMM methods. We treat the lag of the dependent variable as endogenous and use two period lags as instruments. The AR (2) test and the Hansen J test indicate that there is no further serial correlation, and the overidentifying restrictions are not rejected in estimated models.

Similar to what was observed in the fixed effects regressions, we find robust negative association between the earlier search history of individuals pertaining to disinfection methods and confirmed cases of Corona. The other search outcomes in Google Trends with reference to Corona do not show a significant association. The GMM results reinforce our earlier discussion on the importance of earlier awareness and public education on the protective measures against the Corona virus. On average and depending on specification, a one percent increase in intensity of search for disinfection methods reduces the confirmed cases from 0.51% to approximately 1% in confirmed cases of Corona.

5. Conclusion

Our study of 31 Iranian provinces from 19 Feb. 2020 to 9 March 2020 aims to study the possible drivers of within province changes on confirmed cases of COVID-19 Coronavirus. We show that the Google Trend database can give us some interesting insights on possible factors in reducing the confirmed cases. Using province and time fixed effects and different GMM models, we show that earlier records of searches for disinfection methods and materials have both statistically and economically relevant association with lower records of confirmed cases in Iran. This finding refers to the importance of the free flow of information and removal of restriction on internet and social media, as well as increasing the affordability of access to information, as potentially effective tools to dampen outbreak of Corona.

Funding: Not applicable

Conflict of interest: The authors declare that they have no conflict of interest.

Informed consent: All individual participants included in this study provided informed consent.

References

Aguilera, A. M., Fortuna, F., Escabias, M., & Di Battista, T. (2019). Assessing Social Interest in Burnout Using Google Trends Data. *Social Indicators Research*, 1-13.

Althouse, B.M., Ng, Y.Y., & Cummings, D.A. (2011). Prediction of dengue incidence using search query surveillance. *PLoS Negl. Trop. Dis. 5, e1258*.

Arellano, M., & Bover, O. (1995). Another look at the instrumental variables estimation of error components models. *Journal of Econometrics*, 68(1), 29-51.

Baltagi, B. H. (2008). Econometric analysis of panel data (4th ed.). Chichester: *Wiley*.

Baltagi, B. H., Demetriades, P. O., & Law, S. H. (2009). Financial development and openness: Evidence from panel data. *Journal of Development Economics*, 89(2), 285–296.

Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1), 115–143.

Carneiro, H.A., & Mylonakis, E. (2009). Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks. *Clinical Infectious Diseases*, 49, 1557–1564.

Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with Google Trends in an emerging market. *Journal of Forecasting*, *32*(4), 289-298.

Chan, E.H., Sahai, V., Conrad, C., & Brownstein, J.S. (2011). Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS Negl. Trop. Dis*, 5(5), 1-6.

Cho, S., Sohn, C.H., Jo, M.W., Shin, S.Y., Lee, J.H., Ryoo, S.M., Kim, W.Y. & Seo, D.W. (2013). Correlation between national influenza surveillance data and google trends in South Korea. *PloS one*, *8*(12), 1-7.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*, 2-9.

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), 1-8.

Dergiades, T., Mavragani, E., & Pan, B. (2018). Google Trends and tourists arrivals: emerging biases and proposed corrections. *Tourism Management*, 66, 108–120.

Di Bella, E., Leporatti, L., & Maggino, F. (2018). Big data and social indicators: Actual trends and new perspectives. *Social Indicators Research*, 135(3), 869-878.

Dugas, A.F., Hsieh, Y.H., Levin, S.R., Pines, J.M., Mareiniss, D.P., Mohareb, A., Gaydos, C.A., Perl, T.M. & Rothman, R.E. (2012). Google Flu Trends: correlation with emergency department influenza rates and crowding metrics. *Clinical Infectious Diseases*, *54*(4), 463-469.

Fetzer, T., Hensel, L., Hermle, J., & Roth, C. (2020). Perceptions of Coronavirus Mortality and Contagiousness Weaken Economic Sentiment. *arXiv preprint arXiv:2003.03848*.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., & Brilliant, L., (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012–1014.

Hand, C., & Judge, G. (2012). Searching for the picture: forecasting UK cinema admissions using Google Trends data. *Applied Economics Letters*, *19*(11), 1051-1055.

Hu, D., Lou, X., Xu, Z., Meng, N., Xie, Q., Zhang, M., Zou, Y., Liu, J., Sun, G.P. & Wang, F. (2020). More Effective Strategies are Required to Strengthen Public Awareness of COVID-19: Evidence from Google Trends. *Available at SSRN 3550008.*

Husnayain, A., Fuad, A., & Yu Su, E.C. (2020). Applications of google search trends for risk communication in infectious disease management: A case study of COVID-19 outbreak in Taiwan. *International Journal of Infectious Diseases (In Press).*

Husnayain, A., Fuad, A., & Lazuardi, L. (2019). Correlation between Google Trends on dengue fever and national surveillance report in Indonesia. *Glob Health Action*, 12, 1-8.

Kang, M., Zhong, H., He, J., Rutherford, S., & Yang, F. (2013). Using google trends for influenza surveillance in South China. *PloS one*, *8*(1), 1-6.

Kristoufek, L., Moat, H. S., & Preis, T. (2016). Estimating suicide occurrence statistics using Google Trends. *EPJ Data Science*, *5*(32), 1-12.

Li, C., Chen, L.j., Chen, X., Zhang, M., Pang, C.P., & Chen, H. (2010). Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020, *Euro Surveill*, 25(10), 1-5.

Lu, Y., Wang, S., Wang, J., Zhou, G., Zhang, Q., Zhou, X., Niu, B., Chen, Q. & Chou, K.C. (2019). An epidemic avian influenza prediction model based on google trends. *Letters in Organic Chemistry*, *16*(4), 303-310.

Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., & Simonsen, L. (2013). Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Computational Biology*, *9*(10), 1-11.

Preis, T., Moat, H. S., & Stanley, H.E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, *3(*1684), 1-6.

Silva, E.S., Hassani, H., Madsen, D.Ø., & Gee, L. (2019). Googling fashion: forecasting fashion consumer behaviour using google trends. *Social Sciences*, 8(4), 111.

StatCounter Global Stats (2020). Search Engine Market Share Islamic Republic of Iran. Available at https://gs.statcounter.com/search-engine-market-share/all/iran/#monthly-201902-202002

Teng, Y., Bi, D., Xie, G., Jin, Y., Huang, Y., Lin, B., An, X., Feng, D., & Tong, Y. (2017). Dynamic forecasting of Zika epidemics using Google Trends. *PloS one*, *12*(1), 1-10.

Verma, M., Kishore, K., Kumar, M., Sondh, A.R., Aggarwal, G. & Kathirvel, S. (2018). Google search trends predicting disease outbreaks: An analysis from India. *Healthcare Informatics Research*, 24(4), 300-308.

Yang, S., Santillana, M., & Kou, S. C. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, *112*(47), 14473-14478.

Yu, L., Zhao, Y., Tang, L., & Yang, Z. (2019). Online big data-driven oil consumption forecasting with Google trends. *International Journal of Forecasting*, *35*(1), 213-223.

Zhang, Y., Bambrick, H., Mengersen, K., Tong, S., & Hu, W. (2018). Using Google Trends and ambient temperature to predict seasonal influenza outbreaks. *Environment International*, 117, 284-291.