Investigating Businesses' Responses to COVID via Web Crawling

Charlotte Chunming Meng
National Institute of Economic and Social Research (NIESR)
C.Meng@niesr.ac.uk

John Forth
City University of London

Rebecca Riley
King's College London

Data on the activities of businesses are typically collected using surveys or, in some cases, by interrogating administrative records. In the UK, the ONS' Business Impact of Coronavirus Survey and the Decision Maker Panel are two examples of surveys that have provided regular information on how businesses are responding to the COVID pandemic. Administrative data has also proved valuable. Examples include the HR1 forms that firms must submit to the Insolvency Service when planning collective redundancies, HRMC data on firms' use of the Coronavirus Job Retention Scheme and data from the British Business Bank on take-up of its various loan schemes.

These various sources are clearly of value. However, surveys are costly to administer and can suffer from low response rates, especially during crises. Administrative data, for their part, are necessarily partial, can be difficult to combine and often become available for research with a significant time lag.

In this research, we take a different approach, seeking to gather timely data on UK businesses' various responses to COVID-19 from information made publicly available on the web. Our motivations are two-fold. First, we wish to investigate the extent to which public information can replace, or complement, data obtained from traditional sources. Second, to the extent that web-based data prove to be of value, we wish to use these data to provide insights into businesses' responses to COVID through the various stages of the pandemic. Our approach is distinct from the

use of web sources to generate faster indicators of economic activity because we seek to use information from the web to identify responses to the pandemic at the level of the individual firm.

To meet these objectives, we crawl the web on a fortnightly basis to collect publicly available information on around 3,500 businesses operating in the UK. Web crawling is undertaken in partnership with glass.ai – a UK company that has developed artificial intelligence (AI) technology to read and interpret the content from the open web, reading millions of websites and news sources. We crawl the web to search for COVID-related information describing the responses of our 3,500 businesses to the pandemic, across four areas of business activity: employment, business operations, finance and investment and social responsibility.

We use text analysis to code the content of this web-crawled data and then validate the content by comparing with the results of a traditional survey of a subset of 310 firms. We find that information from these public web sources predicts firm actions to a reasonable accuracy for certain types of firms, i.e. 87% of cases for listed firms versus 53.6% for unlisted firms when considering the use of homeworking (see also column 1 of Table 1 which shows factors associated with prediction accuracy). We also identify in the data the publication bias – firms disclose actions that are likely to give them a positive image (e.g. homeworking) rather than a negative one (e.g. redundancies). Website posts and reports are more reliable than news reports.

With the data and methodology described above, we further investigate firm behaviours during the pandemic. For example, using these data we find that firms that are larger in size are more likely to adopt homeworking arrangements after controlling for factors associated with prediction accuracy (as is shown in column 2 of Table 1).

Table 1. Results for homeworking arrangements

	(1) Web text predicts survey response	(2) Firm has adopted homeworking arrangements
Listed	0.364***	0.353***
	(0.066)	(0.019)
Log(turnover)	0.016	0.062***
	(0.024)	(0.007)
Total number of documents (1,000)	0.034**	
	(0.012)	
Employment size dummies	Yes	Yes
Industry dummies	Yes	Yes
Number of observations	310	3,487
R-squared	0.167	0.178

Column 1: dependent variable = 1 if web text predicts survey response; 0 otherwise
Column 2: dependent variable = 1 if web text indicates that firm has adopted homeworking; 0 otherwise
Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Our research shows that data from public data sources (collected via web crawling) are of value when timely measures of the economy are lacking. In particular, we demonstrate the capability of web sources to generate firm-level data that can be used to identify and explain the behaviours of individual businesses in times of significant shocks (e.g. Brexit, COVID). Our research provides an assessment of when and when not public data might be used to substitute for or complement survey evidence and extends the use of web sources beyond conventional approaches to understanding economic activity, contributing to the increasing literature that uses texts as data during the pandemic (Cheema-Fox *et al.*, 2020; Hassan *et al.*, 2020; Kinne *et al.*, 2020; Sacerdote *et al.*, 2020).

References

Cheema-Fox, A., B. R. LaPerla, G. Serafeim and H. S. Wang (2020). 'Corporate resilience and response during covid-19', *Harvard Business School Accounting & Management Unit Working Paper*(**20-108**).

Hassan, T. A., S. Hollander, L. Van Lent, M. Schwedeler and A. Tahoun (2020) *Firm-level exposure to epidemic diseases: Covid-19, sars, and h1n1*, National Bureau of Economic Research.

Kinne, J., M. Krüger, D. Lenz, G. Licht and P. Winker (2020). 'Coronavirus pandemic affects companies differently', in (Editor Ed.)^Eds.), *Book Coronavirus pandemic affects companies differently*, City: ZEW–Leibniz Centre for European Economic Research Mannheim.s

Sacerdote, B., R. Sehgal and M. Cook (2020) *Why is all covid-19 news bad news?*, National Bureau of Economic Research.