2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8327

Digital Vapor Trails: Using Website Behavior to Nowcast Entrepreneurial Activity

Slaper, Timothy F.[a]; Bianco, Alyssa[b] and Lenz, Peter E.[c]
[a]Indiana Business Research Center, Indiana University, United States, [b]Dstillery, United States, [c]Dstillery, United States.

Abstract

Following recent research, we explore virtually contemporaneous, and geographically granular, user online activity related to entrepreneurship. In this paper, we present evidence that data harvested by Dstillery can complement efforts of, and data collected by, government agencies and organizations advocating for entrepreneurship, business formation and economic growth, e.g., the Kauffman Foundation. Our website-based behavior data is close to real time and at a geographically granular level. We find that the concentration of a region's visits to website resources for entrepreneurship and business development are statistically related to business start-up and, particularly, growth activity. Visits to websites related to entrepreneurship are more strongly associated with growth entrepreneurship, in contrast to start-up entrepreneurship. While data capture and analysis related to entrepreneurship website activity is in its infancy, this analysis points to the potential of this data source to nowcast business formation and growth at a regional level.

Keywords: *Nowcasting; entrepreneurship; start-ups; business formation; website behavior.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

Data based on website activity—or, in this case, the behavior of economic agents—raise the possibility of enabling researchers and policymakers with a means to augment, and make more timely, official government statistics at a geographically granular level. Recently, Google Trends has gained increasing popularity in the practice of "nowcasting" economic variables, as well as other social and health outcomes. Google Trends data have been used to predict the outbreaks and spread of disease (Carneiro & Mylonakis, 2009), tourist flows (Siliverstovs & Wochner, 2018), consumer behaviors (Vosen & Schmidt, 2012), and unemployment (Askitas & Zimmermann, 2009; Pavlicek & Kristoufek, 2015; Naccarato, Pierini, & Falorsi, 2015; Vicente, López-Menéndez, & Pérez, 2015; D'Amuri & Marcucci, 2017).

In short, research on the potential of using digital vapor trails to predict economic outcomes, or nowcast economic activity, is growing (e.g., Choi & Varian, 2012; Wu & Brynjolfsson, 2015; Goel, Hofman, Lahaie, Pennock, & Watts, 2010). As noted by Glaeser, Kim, and Luca (2017), online data sources make it possible to measure new activities and outcomes that, heretofore, were outside the scope of official, traditional data sources.

In this paper, we explore virtually contemporaneous and geographically granular user online activity related to entrepreneurship. We present evidence that data harvested by Dstillery can complement efforts of, and data collected by, government agencies and organizations advocating for entrepreneurship, business formation and economic growth—such as the Kauffman Foundation—using a proxy measure of entrepreneurial activity. We extend the existing nowcasting literature based on digital vapor trails by aligning data new to economic research, namely from Dstillery, with data developed by an entrepreneurship advocacy organization—the Kauffman Foundation. Our key contribution is to evaluate the potential (potential because capturing the data is still emerging) of whether website behavior data tracks with time-tested and traditional measures of business formation activity at a regional geographic scale, based on metropolitan statistical areas (MSAs).

2. Data and Method

For over two decades, the Kauffman Foundation has published the Kauffman Index on Entrepreneurship. In many ways, it is the state-of-the-art in measuring business start-ups, formation and growth (Fairlie, Morelix, Reedy, & Russell, 2015; Kauffman Foundation, 2017a, 2017b). Over time, and with much research and analysis, the index differentiated start-up activity and early business growth activity. In other words, what may signal the creation of a new business is not the same as the signals that a new business is growing and prospering. Given the time and research focus devoted to developing the three measures for the two entrepreneurial categories—start-up and growth—we embrace the six measures that comprise the two indexes as the best available. These data are available for the top 40 metro areas in the country. The Kauffman Foundation does not publish an index for other metro areas.

While Dstillery is well established in the digital marketing ecosystem, the application of their data linking individual website behavior to economic activity is novel. Dstillery is a predictive marketing intelligence firm that anonymously collects, classifies, and disseminates behavioral data. Dstillery's digital data is collected as a proxy for real-world behaviors. Marketers use these data to optimize which devices see ads at times relevant to a consumer (Raeder, Stitelman, Dalessandro, Perlich, & Provost, 2012).

Dstillery can create an analytical category based on the websites a marketer or researcher may think pertains to a particular constituency. Dstillery captures a representative sample of devices that visit these websites and finds the highest scoring features to create a model that scores devices based on their affinity to the behavior of interest (Ibarra & Lenz, 2016), in this case the E-ship audience. The score of that set of devices is compared to a random sample of devices across the internet. Device scores are aggregated at the ZIP code level, based on the predicted "home" location of each device according to a probabilistic model that takes into account time of day and frequency of visit to a discrete location.

The Indiana Business Research Center team provided a list of relevant organizations and websites from the Kauffman Foundation website, plus several entrepreneurial website guides, to generate a list of 100 websites for Dstillery to observe and measure traffic. These data for E-ship affinity was, in turn, translated or aggregated from ZIP code geographic units of analysis into MSAs. Concentration of activity values were scaled by the relative population of each ZIP code within an MSA. (Unscaled data yields results that were less robust.)

The first tranche of Dstillery data for entrepreneurship—E-ship—from the website list was captured in January 2018. These data are new and exploratory. The Kauffman Index data are from the latest iteration of the index (2017), but some of the data used to calculate the index are based on business formation relationships, or data, from 2014. We assume one important regional characteristic: A region’s internet and website behavior in the late months of 2017 is consistent with that region’s experience in 2016 and years previous. That is, people’s tastes and interests don’t change dramatically in a place/location over time unless there is some titanic event. Another way to view this assumption is that a region’s culture does not change quickly; thus, a region’s propensity to express interest in activities, entrepreneurial or not, will not change dramatically compared to other regions over the course of a few years.

One may view this as a corollary of the work of Obschonka et al. (2015): Personality profiles of a region can help explain entrepreneurial activity. One would not expect that the psychological cultural characteristics of a region would change significantly from one year to the next.

3. Results

Because these data are so new—Dstillery only recently started to capture E-ship data—we only have one snapshot of E-ship–related web behavior. (Over time, however, Dstillery is expected to collect these so that researchers will be able to track E-ship concentration changes over time and how they relate or possibly predict start-up and entrepreneurship growth.)

First, we performed a simple correlation between the three sub-measures of the start-up activity and the entrepreneurship growth indexes—six component measures in all (namely the rate, the share and the density) and the Dstillery E-ship measure for website traffic for the 40 metropolitan statistical areas covered by the Kauffman Index data. (Please refer to the Kauffman Foundation reports (2017a, 2017b) for more detail on what the rate, share and density measures capture.)

The correlations are not particularly strong, but there are interesting differences in the relationship between Dstillery E-ship data and the two types of entrepreneurial activity tracked by the Kauffman Foundation. For the start-up index, the correlations are 0.32, 0.00 and 0.19 for rate, share and density components, respectively. The correlations for the entrepreneurial growth index components are 0.49, 0.27 and 0.42 for rate, share and density, respectively. This would indicate that entrepreneurs in the growth phase of their new companies utilize web-based resources to a much greater extent to help them grow their businesses, in contrast to the utilization of entrepreneurs who are just getting started.

We then considered the relationship between Dstillery E-ship and the higher correlated index measures for both start-up rate, growth rate and growth density. We used a simple OLS model to assess the degree to which the variation in the index component values for these three concepts may be explained by the regional/MSA concentration of E-ship website traffic as captured by Dstillery. The explained variation—adjusted R-squares—for the three dependent variables were not particularly strong: Start-up Rate of New Entrepreneurs – 0.08; Rate of Startup Growth – 0.22; and High-Growth Company Density – 0.16. The coefficients are positive with p-values of 0.048, 0.001 and 0.007, respectively.

The modest association suggested by the statistical results is both good news and bad news. That E-ship web traffic can provide some explanatory power across metropolitan areas suggests that this concept and measure may be a good candidate to use as one piece of an entrepreneurship nowcasting data set and deserves further study and development. Part and parcel of that development would be to ensure that the E-ship website universe is complete and captures all relevant web-based resources. The bad news is that this simple exercise has compared differences in two snapshots of metro areas. It doesn't measure change over time. Moreover, the foundational characteristics of each region has not been explored sufficiently in this analysis. The Kauffman measures treat each start-up the same, whether a food truck or a medical testing clinic or a high-tech systems integration company. Arguably, those regions that dominate in the digital and technology space would also have greater connectivity and more devices (digital reach) and would likely have a greater proportion of their denizens who would access web-based resources to gain knowledge, know-how or seek enterprise funding. Regional industry characteristics and opportunities may explain many behaviors.

4. Conclusion

In this paper, we have used a novel data set to test the hypothesis that regional differences in E-ship–related web traffic may help to explain differences in regional business start-ups and new business growth. The data are so novel, that there is only one snapshot for the concentration of E-ship–related website usage. We compared two snapshots: the Kauffman Index of Entrepreneurship Activity and Dstillery E-ship data for 40 metropolitan areas in the United States. We found that E-ship web activity is more closely associated with growth entrepreneurship than with start-ups. While this difference may be attributed to differences in regional characteristics across metro areas, it may also signal that those in the start-up phase of creating a new business are not as inclined to utilize online resources to learn how to run a business or expand their knowledge base. If predominantly the latter, these findings may point to the need for a policy or resource response to better serve those in the very early stages of starting a business.

Given the one-off nature of the analysis, we cannot currently advocate for the Dstillery E-ship data to be considered a viable data source for nowcasting entrepreneurship and business formation. That said, as the data series is captured over time, these data may be valuable for policymakers, economic development practitioners and even government economic statisticians to watch in the future.

Acknowledgment

This publication was prepared by Tim Slaper, Alyssa Bianco, and Peter Lenz using Federal funds awarded to the Trustees of Indiana University and as a subcomponent under award number ED17HDQ3120040 from the U.S. Economic Development Administration, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of the Economic Development Administration or the U.S. Department of Commerce.

References

Askitas, N., & Zimmermann, K. (2009). *Google econometrics and unemployment forecasting* (DIW Berlin Discussion Paper 899). Berlin: German Institute for Economic Research. doi:10.2139/ssrn.1465341

Baker, S. R., & Fradkin, A. (2017). The impact of unemployment insurance on job search: Evidence from Google search data. *The Review of Economics and Statistics, 99*(5), 756–768.

Carneiro, H. A., & Mylonakis, E. (2009). Google Trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases, 49*(10), 1557–1564.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record, 88*(s1), 2–9.

D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting, 33*(4), 801–816.

Fairlie, R. W., Morelix, A., Reedy, E. J., & Russell, J. (2015). The Kauffman Index 2015: Startup activity national trends. Kansas City, MO. doi: 10.2139/ssrn.2613479

Glaeser, E. L., Kim, H., & Luca, M. (2017). *Nowcasting the local economy: Using Yelp data to measure economic activity* (NBER Working Paper 24010). Cambridge, MA: National Bureau of Economic Research.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences, 107*(41), 17486-17490.

Ibarra, P. & Lenz, P. (2016, March). Using digital signals to measure audience brand engagement at major sports events: The 2015 MLB season. Paper presented at *MIT Sloan Sports Analytics Conference*, Boston.

Kauffman Foundation (2017a). Kauffman Index 2017: Growth entrepreneurship metropolitan area and city trends. Kansas City, MO. doi:10.2139/ssrn.3080714

Kauffman Foundation (2017b). 2017 Kauffman Index of Startup Activity: Metropolitan area and city trends. Kansas City, MO. doi:10.2139/ssrn.2974544

Naccarato, A., Pierini, A., & Falorsi, S. (2015). *Using Google Trend data to predict the Italian unemployment rate* (Department of Economics Working Paper 203). Rome: University Roma Tre.

Obschonka, M., Stuetzer, M., Gosling, S. D., Rentfrow, P. J., Lamb, M. E., Potter, J., & Audretsch, D. B. (2015). Entrepreneurial regions: Do macro-psychological cultural characteristics of regions help solve the "knowledge paradox" of economics? *PLOS ONE*, *10*(6). doi:10.1371/journal.pone.0129332

Pavlicek, J., & Kristoufek, L. (2015). Nowcasting unemployment rates with Google searches: Evidence from the Visegrad Group countries. *PLOS ONE, 10*(5). doi:10.1371/journal.pone.0127084

Raeder, T., Stitelman, O., Dalessandro, B., Perlich, C., & Provost, F. (2012). Design principles of massive, robust prediction systems. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD

'12), 1357-1365. New York: Association for Computing Machinery. doi:10.1145/2339530.2339740

Siliverstovs, B., & Wochner, D. S. (2018). Google Trends and reality: Do the proportions match? Appraising the informational value of online search behavior: Evidence from Swiss tourism regions. *Journal of Economic Behavior & Organization 145*, 1–23.

Vicente, M. R., López-Menéndez, A. J., & Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change, 92*(Supplement C), 132–139.

Vosen, S., & Schmidt, T. (2012). A monthly consumption indicator for Germany based on Internet search query data. *Applied Economics Letters, 19*(7), 683–687.

Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In A. Goldfarb, S. M. Greenstein, & C. E. Tucker (Eds.), *Economic Analysis of the Digital Economy* (pp. 89-118). Chicago: University of Chicago Press.