Big Data Measures of Environmental Concern

Agol Wai Ming Ho
Hong Kong Metropolitan University

Kevin Chi Keung Li
Hong Kong Metropolitan University

Environmental concern is a subjective state of society and researchers have typically relied on survey data to measure it. However, survey-based methods only capture a snapshot of it at the time and place the surveys were conducted. To overcome these problems, we develop an observable indicator that allows us to study environmental concern over time and across territories based on big data. The indicator composes of keyword groups that fit with the environmental concern measures revealed by a large-scale survey. We find that keywords associated with climate change, water pollution and waste management are the strongest predictors of environmental concern. To the best of our knowledge, our paper is the first to use online search data to capture subjective environmental concern.

Keywords: environmental concern, big data, online activities

INTRODUCTION

Most researchers refer environmental concern to attitudes about environmental issues or perceptions that such issues are important (Cruz, 2017). There are several reasons why we construct an indicator of environmental concern among the citizens of a country or a geographical area. First, environmental concern is an important variable, and many policy and business decisions depend on it. Unfortunately, it a subjective state of society that cannot be measured directly. Second, researchers have typically relied on survey data to measure environmental concern (e.g., Teoh and Gaur 2019; Ahmad 2018; Li & Chen 2018). However, survey-based methods only capture a snapshot of environmental concerns at the time and place the surveys were conducted while sample limitations or the high cost involved make a comprehensive analysis infeasible. In order to overcome these problems, we develop an observable indicator that allows us to study the actual environmental behavior and attitude of citizens over time and across geographical areas based on big data. Search frequency of keywords on internet are the big data employed in this paper.

In respect to the definition of big data, there is no uniform point of view. Following Manyika et al. (2011), this paper defines big data as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze". Since big data contains massive information related to how people, firms and other entities interact, it could help reveal economic trends and monitor social behaviors (Blazquez & Domenech, 2018). There is an expanding set of studies that uses online search activity to study subjective states of society. Stephens-Davidowitz (2014) uses online search data to model racial animus in the United States. Algan et al. (2016) uses online search activity to build an indicator of well-being in the

United States. Likewise, in the finance literature Da et al. (2011), Vlastakis and Markellos (2012) and Chen et al. (2016) use online search activity to study investor attention or investor information demand. Many studies (Kornellia & Syakurah, 2023) even used online search activity to predict positive cases and deaths due to COVID-19 during the pandemic. To the best of our knowledge, our paper is the first to use online search data to capture subjective environmental concern.

METHODOLOGY

The basic idea of this paper is that we use search frequencies on internet for related keywords to measure subjective environmental concern among citizens of a territory. We adopt the approach of Algan et al. (2016) to construct an observable indicator of environmental concern.

First, we scanned through a large number of publications of some renowned environmental organizations, such as World Wide Fund for Nature (Annual Review), United Nations Environment Programme (Global Environment Outlook) and the Organization for Economic Cooperation and Development (Environment at a Glance) to identify a long list of keywords on environmental concern. A long list of possible keywords will avoid the problem of cherry picking (Algan et al, 2016). Keywords that are included represent searches for causes, consequences and remedial actions to environmental problems, such as "water pollution", "global warming" and "recycling", etc. Some of the keywords were slightly adjusted because of internet searching habits of people were considered. For example, "recycled plastic" was changed to "recycle plastic".

Second, we obtained the search frequencies for each keyword from Google. We choose Google for our paper as it is the most used online search engine, with over 70,000 search queries per second (http://www.Internetlivestats.com/). Google provides data on "Search Volume Index" through a service called "Google Trends". In this paper, US cross-sectional data of the keywords from Google Trends were collected. According to the geographical information of the Google Trends, they could provide search information up to metropolitan area in the US, which is called the designated market area (DMA). There are 210 DMAs in the US.

Third, we combined individual keywords into composite categories to deal with the issue of high-dimensionality that is pervasive in big data analyses. In our setting, the issue of high-dimensionality is caused by a large number of potential explanatory variables and the creation of composite categories greatly reduces the number.

Fourth, we searched for US cross-sectional surveys on environmental concern for the calibration of our environmental concern indicator. We selected the Cooperative Congressional Election Study (CCES) survey for this paper, because it is most representative survey in the US with over 60,000 respondents. The CCES is a national online survey administered by YouGov which is a global public opinion and data company. Its goal is to survey voters in the US presidential and midterm elections. This paper made use of the 2016 CCES survey which contains a question related to environmental concern. The question is "How important are each of these issues (environment, CC16_301h) to you?". The respondents' geographical information from the selected survey were matched into various DMAs using their zip codes.

Fifth, we ran a regression of survey environment concern on the composite categories of keywords to determine their importance. A subset of composite categories that best represents environmental concern was selected and used to compile a time-series of environmental concern indicator. We call it a big data measure of environmental concern.

Finally, we compared our big data measure of environmental concern with a traditional measure from a time-series survey to evaluate the effectiveness of the resulting model. The time-series survey data on environmental concern were taken from the Gallup Poll Social Series of Gallup Analytics, which is a famous survey company in the US. The survey data were collected by annual telephone interviews with a random sample of about 1,000 adults, ages 18 and older, living in all 50 U.S. states and the District of Columbia. The time span of the survey covers 15 years from 2006 to 2020. The relevant survey question is "How much do you personally worry about the quality of the environment?"

RESULTS

After scanning through the publications of three renowned environmental organizations for the period of 2010-2020, we selected 512 keywords that would be possibly correlated with environmental concern. To refine this long list of keywords, the frequency of keywords appeared in the publications were counted using the R software. The top 50 keywords were selected for the construction of environmental concern indicator. Search frequencies of these keywords were downloaded from Google Trends for all DMAs in the US. These keywords were later classified into five categories, which are climate change, air quality, water pollution, waste management and biodiversity. Table 1 lists out the keywords selected for all categories.

TABLE 1
KEYWORD GROUPS FOR ENVIRONMENTAL CONCERN

Climate change	Air pollution	Water pollution	Waste management	Biodiversity
climate change	air quality	water quality	recycling	conservation
renewable	ozone	wastewater	waste management	redd
global warming	smog	freshwater fish	landfill	ecology
carbon cycle	clean air	runoff	reuse	biodiversity
energy efficient	respiratory infection	water pollution	hazardous waste	biodegradable
greenhouse effect	air pollution	dead zone	composting	endanger species
carbon footprint	aerosol	water disease	food waste	deforestation
fuel efficient	clean energy	bci	recycle plastic	oil spill
extreme weather	acid rain	water crisis	plastic pollution	resource management
sea level rise	greenhouse gas	desalination	3r	poaching

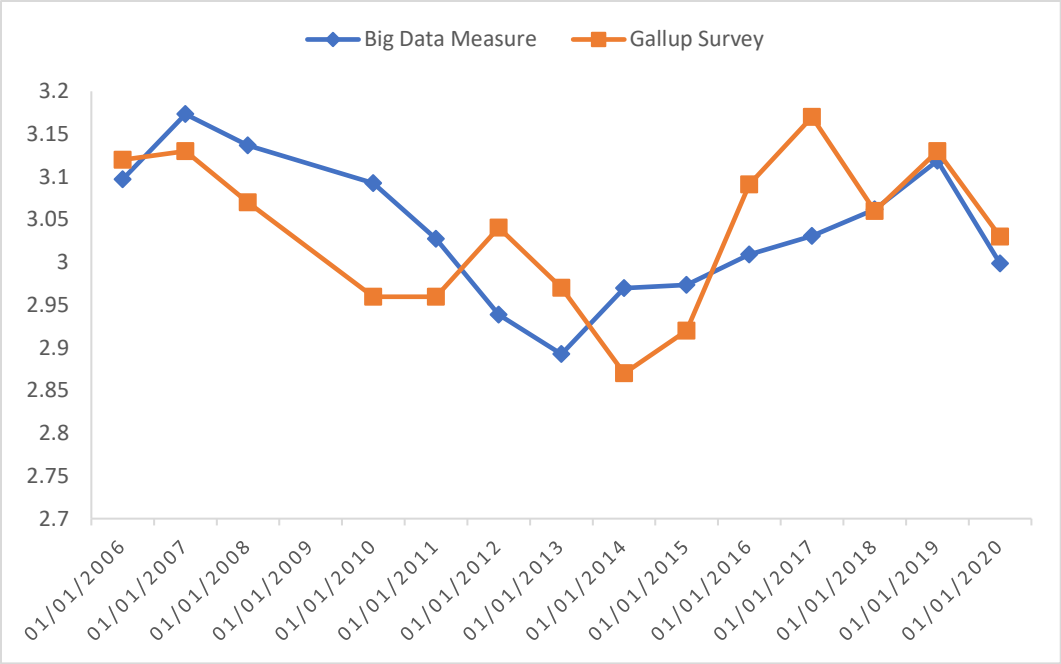
We constructed the latent variables of these categories by using the factor loadings of an exploratory factor analysis.

Using time-series Google data of environmental keywords in the US and the model we developed in the last stage, we can construct a time-series measure of environmental concern in the US.

To evaluate the effectiveness of our big data measure of environmental concern, we compare the time-series of the big data measure with the time-series survey result conducted by Gallup Analytics, which is a famous survey company in the US.

Figure 1 shows the time trend of environmental concern in the US over 15 years from 2006 to 2020. Note that, we put environmental concern on the vertical axis, and time on the horizontal axis. We use a four-point scale to measure environmental concern. “4” means very concern; and “1” means not concern at all. The blue line is our big data measure of environmental concern, while the red line is the survey result conducted by Gallup.

FIGURE 1
TREND OF ENVIRONMENTAL CONCERN IN THE US



It is observed that our big data measure tracks the Gallup survey result quite well. This prove that our bid data measure is closely correlated with traditional survey measures while being available in real time.

CONCLUSION

Environmental concern is a subjective state of society and evolve over time and varies across geographical areas. In this paper, we compile an observable and timely indicator of the subjective environmental concern in a territory using big data. Online searching activities is the big data used in this paper. With the new measure of environmental concern, people can determine the influence of environment concern on various government policies, social issues and financial markets.

We also find that keywords associated with climate change, water pollution and waste management are the strongest predictors of subjective environmental concern.

ACKNOWLEDGEMENT

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS16/B17/19).

REFERENCES

Ahmad, I., Syed, F., Naseer, S., & Rasool, G. (2018). Environmental Concern as an Underlying Mechanism between Environmental Beliefs and Green Purchase Intentions. *South Asian Journal of Management Sciences*, *12*(1), 93–115.

Algan, Y., Beasley, E., Guyot, F., Higa, K., Murtin, F., & Senik, C. (2016). *Big Data Measures of Well-Being: Evidence From a Google Well-Being Index in the United States* (No. 2016/3). OECD Publishing.

Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*, 99–113 https://doi.org/10.1016/j.techfore.2017.07.027

Chen, J., Liu, Y.J., Lu, L., & Tang, Y. (2016). Investor attention and macroeconomic news announcements: Evidence from stock index futures. *Journal of Futures Markets*, *36*(3), 240–266.

Cruz, S.M. (2017). The relationships of political ideology and party affiliation with environmental concern: A meta-analysis. *Journal of Environmental Psychology*, *53*, 81–91.

Da, Z.H.I., Engelberg, J., & Gao, P. (2011). In Search of Attention. *The Journal of Finance*, *66*, 1461–1499.

Davidowitz, S.S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics*, *118*, 26–40.

Kornellia, E., & Syakurah, R.A. (2023). Use of Google Trends database during the COVID-19 pandemic: Systematic Review. *Multidisciplinary Reviews*, (Accepted Articles). Retrieved from https://malque.pub/ojs/index.php/mr/article/view/855

Li, W., & Chen, N. (2018). Absolute income, relative income and environmental concern: Evidence from different regions in China. *Journal of Cleaner Production*, *187*, 9–17.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A.H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Retrieved from https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/big%20data%20the%20next%20frontier%20for%20innovation/mgi_big_data_full_report.pdf

Teoh, C.W., & Gaur, S.S. (2019). Environmental concern: An issue for poor or rich, *Management of Environmental Quality: An International Journal*, *30*(1), 227–242. https://doi.org/10.1108/MEQ-02-2018-0046

Vlastakis, N., & Markellos, R.N. (2012). Information demand and stock market volatility. *Journal of Banking and Finance*, *36*, 1808–1821.