EPJ.org

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE   **Open Access**

CrossMark

Estimating suicide occurrence statistics using Google Trends

Ladislav Kristoufek[1,2]*, Helen Susannah Moat[1] and Tobias Preis[1]

*Correspondence:
ladislav.kristoufek@fsv.cuni.cz
[1]Data Science Lab, Behavioural
Science, Warwick Business School,
University of Warwick, Coventry,
CV4 7AL, UK
[2]Institute of Economic Studies,
Faculty of Social Sciences, Charles
University, Opletalova 26, Prague,
110 00, Czech Republic

Abstract

Data on the number of people who have committed suicide tends to be reported with a substantial time lag of around two years. We examine whether online activity measured by *Google* searches can help us improve estimates of the number of suicide occurrences in England before official figures are released. Specifically, we analyse how data on the number of *Google* searches for the terms 'depression' and 'suicide' relate to the number of suicides between 2004 and 2013. We find that estimates drawing on *Google* data are significantly better than estimates using previous suicide data alone. We show that a greater number of searches for the term 'depression' is related to fewer suicides, whereas a greater number of searches for the term 'suicide' is related to more suicides. Data on suicide related search behaviour can be used to improve current estimates of the number of suicide occurrences.

Keywords: nowcasting; search data; Google Trends; official statistics

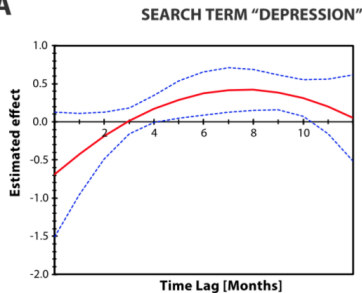**Figure 1 Official data on suicide occurrences and *Google* searches for 'depression' and 'suicide'.** We analyse monthly data from official records of suicide occurrences in England (left *y*-axis). We investigate whether monthly data on *Google* searches for 'depression' and 'suicide' in England (right *y*-axis) can help estimate counts of suicide occurrences before the official data are available. Note that search data retrieved from *Google* are normalised to create an index which takes integer values between 0 and 100. Higher values indicate that a higher proportion of the total searches in England in a given month were for the term of interest.
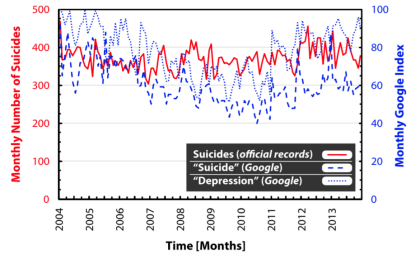
Figure 2 Estimating official suicide counts using *Google* search data. We investigate the relationship between the volume of searches for **(A)** 'depression' and **(B)** 'suicide' and the number of suicide occurrences, with a lag between the search data and the suicide data of up to 12 months (*x*-axis). Solid red lines represent the estimated effect and the dashed blue lines illustrate the 90% confidence intervals of these effects. At a lag of 0 months, we find that a higher number of searches for the term 'depression' corresponds to a lower number of suicides. However, this effect is not statistically significant. For lags of 5 to 10 months, a greater number of searches for the term 'depression' corresponds to a greater number of suicides. Conversely, at a lag of 0 months, we find that a greater number of searches for the term 'suicide' corresponds to a higher number of suicides. The effect vanishes after approximately 2 months. For lags of 6 to 11 months, a greater number of searches for the term 'suicide' corresponds to a lower number of suicides. However, it should be noted that the changes between positive and negative effects for both search terms may be due to the polynomial shape induced by the Almon model specification.

A

SEARCH TERM "DEPRESSION"

Estimated effect / Time Lag [Months]

B

SEARCH TERM "SUICIDE"

Estimated effect / Time Lag [Months]