RESEARCH ARTICLE

Quantifying the Search Behaviour of Different Demographics Using Google Correlate

Adrian Letchford*, Tobias Preis, Helen Susannah Moat

Data Science Lab, Behavioural Science, Warwick Business School, University of Warwick, CV4 7AL, Coventry, United Kingdom

* Adrian.Letchford@wbs.ac.uk

Abstract

Vast records of our everyday interests and concerns are being generated by our frequent interactions with the Internet. Here, we investigate how the searches of *Google* users vary across U.S. states with different birth rates and infant mortality rates. We find that users in states with higher birth rates search for more information about pregnancy, while those in states with lower birth rates search for more information about cats. Similarly, we find that users in states with higher infant mortality rates search for more information about credit, loans and diseases. Our results provide evidence that Internet search data could offer new insight into the concerns of different demographics.

Citation: Letchford A, Preis T, Moat HS (2016) Quantifying the Search Behaviour of Different Demographics Using Google Correlate. PLoS ONE 11(2): e0149025. doi:10.1371/journal.pone.0149025

Editor: Matjaz Perc, University of Maribor, SLOVENIA

Received: November 4, 2015

Accepted: January 26, 2016

Published: February 24, 2016

Data Availability Statement: The Google Correlate results for the birth rates and infant mortality rates as well as the random samples and their Google Correlate results constitutes the data for this analysis. This data is available as a Supporting Information file.

Funding: The authors acknowledge the support of the Research Councils UK Digital Economy Theme via grant EP/K039830/1.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Our everyday interactions with large technological systems are generating records of human behaviour on a colossal scale [1–9]. Data drawn from mobile phone calls [10], public transport smart cards [11], financial markets [12–17], usage of Internet services [18–29], and even immense digitised collections of books [30–33] are being exploited to gather new insights into human health [34–36], mobility [10, 37], economic decision making [23, 38, 39] and more.

Here, we focus on search queries submitted to the Internet search engine *Google*. *Google* makes aggregated data on what people search for online available via its service *Google Trends*, offering unprecedented insight into people's interests and concerns [4]. A number of studies have provided evidence that changes in the frequency with which *Google* users search for given terms across time not only correlate with changes in certain real world variables, such as unemployment rates, but may offer measurements of this behaviour before official data are released [40–45]. Further investigations have suggested that data on online information gathering may even anticipate future values of certain economic and behavioural indicators, such as box office movie revenue and financial market movements [38, 39, 46].

In this paper, we investigate whether or not we can identify a difference in online searches between people in different demographics. Instead of using data drawn from *Google Trends*, we use a service called *Google Correlate* [47, 48]. This service allows a user to input either a time series or data relating to U.S. states, and returns the search terms for which the number of searches is most strongly correlated across time or across states.

However, the correlation coefficients which are returned by *Google Correlate* need to be treated with care. Firstly, the system reports only the highest correlations out of potentially hundreds of millions which greatly increases the chances of finding spurious correlations. Secondly, the search data retrieved from neighbouring states may not be independent and the distribution of search volume may not be Gaussian, such that the data break the assumption of traditional correlation coefficient tests [49]. Thirdly, *Google Correlate* uses a hashing algorithm to improve the speed of searching millions of time series [48]. However, this applies to time series analysis only. Furthermore, *Google* restricts access to their full dataset hindering development of specialised statistical tests.

As a case study, we use *Google Correlate* to investigate how the searches of *Google* users vary across U.S. states with different birth rates and infant mortality rates. We seek to determine which correlations are significant among potentially hundreds of millions of correlations when the data do not follow traditional assumptions. To investigate the results from the case study, we develop a bootstrap statistical test.

Many different demographic variables are measured across US states. Further studies could use the approach we present to extend this investigation to other demographic variables.

Case study

We retrieve the number of births per 1,000 people in each U.S. state in 2012 from the *Centers for Disease Control and Prevention* on 27 May 2014 (http://wonder.cdc.gov/natality.html) (Fig 1A).

We retrieved the list of search terms for which search volume was most strongly positively correlated with birth rate by state by submitting the birth rate data to *Google Correlate* (http://www.google.com/trends/correlate) on 27 May 2014. On the left hand side of Fig 1B, we list the 31 terms for which search volume exhibits the strongest positive correlation with birth rate for a state. We retrieve the list of negatively correlated terms by multiplying the birth rate for each state by −1, before submission to *Google Correlate*. We list the 31 terms for which search volume exhibits the strongest negative correlation with birth rate for a state on the right hand side of Fig 1B.

We observe that particular topics emerge within the lists of terms that *Google Correlate* returns. For example, search terms for which searches are higher in states with higher birth rates include "pregnancy workout", "baby constipation" and "baby announcement". Search terms for which searches are higher in states with lower birth rates include "dry cat food", "older cats" and "cat not eating".

To allow us to generalise beyond single keywords and interpret these datasets in an objective fashion, we conduct an online survey using *Amazon Mechanical Turk*. *Amazon Mechanical Turk* is a service which allows users to post tasks that they wish other users to complete, in exchange for a small fee. For each list of 31 terms, we ask participants, "What is the most prominent topic in these phrases?" Responses are limited to one word, and each participant is only allowed to respond once to each question. In total, we analyse 40 responses received from *Amazon Mechanical Turk* users, with 23 responses for the positively correlated terms, and 17 responses for the negatively correlated terms. Details of the survey can be found in the *Supporting Information*.

In Fig 1C, we depict all survey responses which account for more than 5% of submitted responses, along with the percentage and number of respondents who gave each response. We find that 74% of respondents judge that the search terms for which the number of searches is higher in states with higher birth rates relate to "pregnancy". Conversely, we find that 88% of respondents judge that the search terms for which the number of searches is higher in states

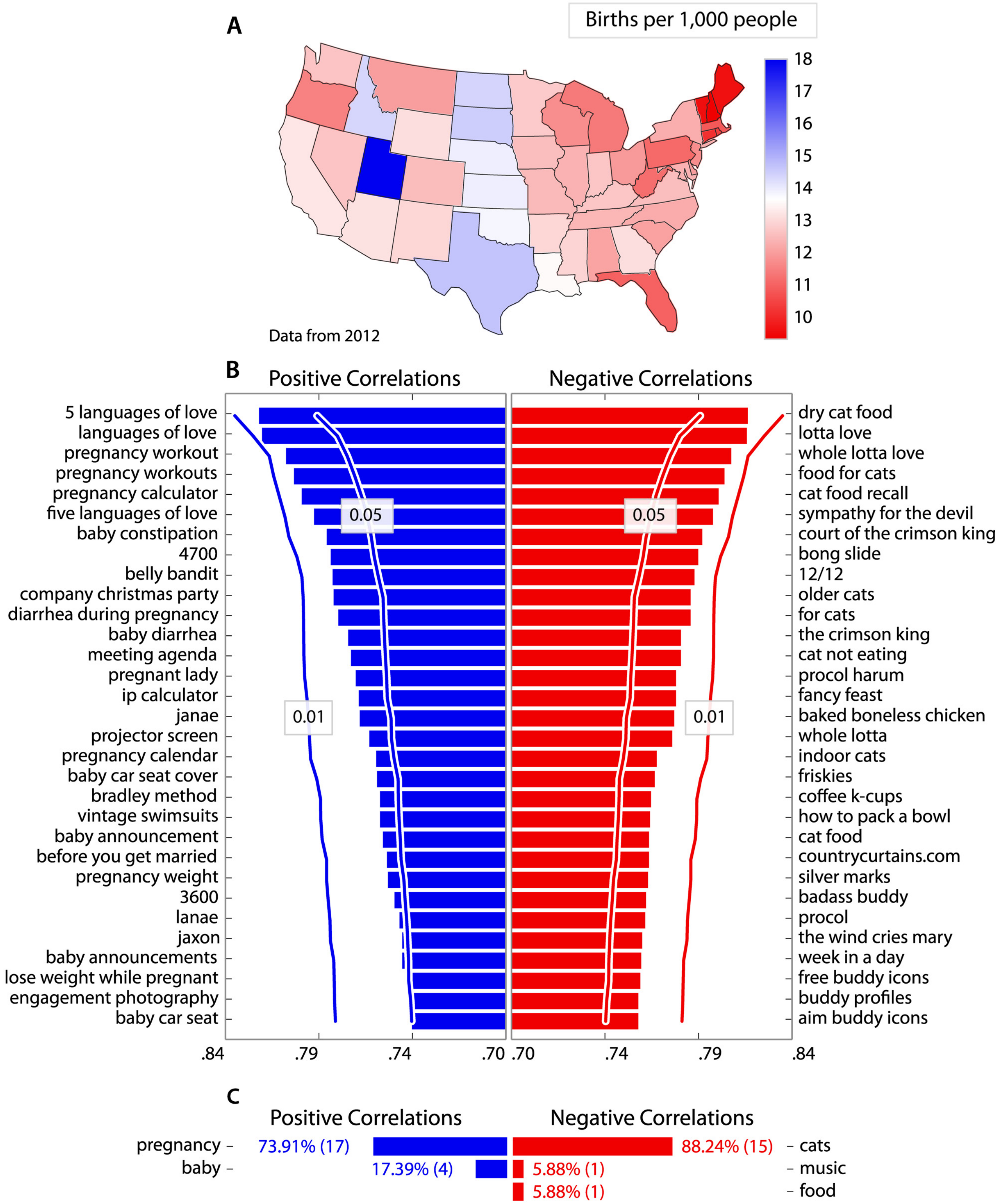

Fig 1. How do *Google* queries vary with birth rate? (**A**) The number of births for 1,000 people in each US state. Birth rate is defined as the number of births for 1,000 people. (**B**) We use *Google Correlate* to find terms for which the number of searches is higher in U.S. states with higher birth rates. Similarly, we identify terms for which the number of searches is higher in states with lower birth rates. Here, we list the 31 terms which showed the strongest positive correlation (left) and negative correlation (right) with state wide birth rate. To determine the significance of these correlations, we generate 1,000 random samples from a multivariate Gaussian distribution where states which are closer together tend to have a similar value. We submit these samples to *Google*

Correlate and build a distribution of correlation coefficients for each of the 31 top most search terms. We depict the strength of correlation required for the correlation to be significant at the $p < 0.05$ and $p < 0.01$ level, given this null hypothesis distribution. (**C**) To allow us to generalise beyond individual search terms, we conduct an online survey asking participants to identify the main topic in each list of 31 terms. Here, we depict all survey responses which account for more than 5% of submitted responses. Our results suggest that users in states with higher birth rates search for more information about pregnancy, while those in states with lower birth rates search for more information about cats ("baby car seat", $p = 0.051$, all remaining ps <0.05).

doi:10.1371/journal.pone.0149025.g001

with lower birth rates relate to "cats". We investigate the statistical significance of these correlations in the following section.

We repeat this process using data on the number of infant deaths per 1,000 births for each state in 2010 downloaded from the *Centers for Disease Control and Prevention* on 27 May 2014 (http://wonder.cdc.gov/lbd.html) (Fig 2A). An infant is defined as any person one year old or younger. In Fig 2B, we list the 31 terms for which search volume exhibits the strongest positive correlation with infant mortality rate for a state (left), and the strongest negative correlation with infant mortality rate for a state (right).

Again, we note that certain topics are apparent within these lists. For example, search terms for which searches are higher in states with higher infant mortality rates include "loan for bad credit" and "people with bad credit", as well as "abnormal pap smear" and "transmitted diseases". Search terms for which searches are higher in states with lower birth rates include "red cabbage salad", "simple frosting" and "carob chips".

Once more, we ask *Amazon Mechanical Turk* users to identify the most prominent topic in each of these lists of terms. In total, we analyse 46 responses received from *Amazon Mechanical Turk* users, with 15 responses for the positively correlated terms, and 31 responses for the negatively correlated terms.

In Fig 2C, we depict all survey responses which account for more than 5% of submitted responses, along with the percentage and number of respondents who gave each response. We find that 80% of respondents judge that the search terms for which the number of searches is higher in states with higher infant mortality rates relate to "credit" or "loans", and 20% of respondents judge that these terms relate to "s.t.d" or "diseases". Conversely, we find that 48% of respondents judge that the search terms for which the number of searches is higher in states with lower infant mortality rates relate to "food", with 10% of users suggesting "frosting", and 6% suggesting "gluten".

Methods

We construct a method to test whether the strength of the correlations for the most correlated search terms for birth rates and infant mortality rates is statistically significant. We note that in Fig 1A the birth rates are not independently distributed. Visual inspection indicates that states which are closer together tend to have similar birth rates. The traditional statistical test for Pearson's correlation coefficient explicitly requires the observations to be independent [49]. To overcome this problem, we perform a bootstrapped statistical test of the correlation between the birth rates and search data. We assume that the birth rates and infant mortality rates are drawn from a multivariate Gaussian distribution. We generate random samples from this distribution and submit each one to *Google Correlate* and build a distribution of the highest correlation coefficients returned by this system.

We set the multivariate Gaussian distribution with a mean of zero and a covariance matrix **K** which accounts for potential covariance between US states. Denoting the geographic data as a vector **y**, we write our model as:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}) \tag{1}$$

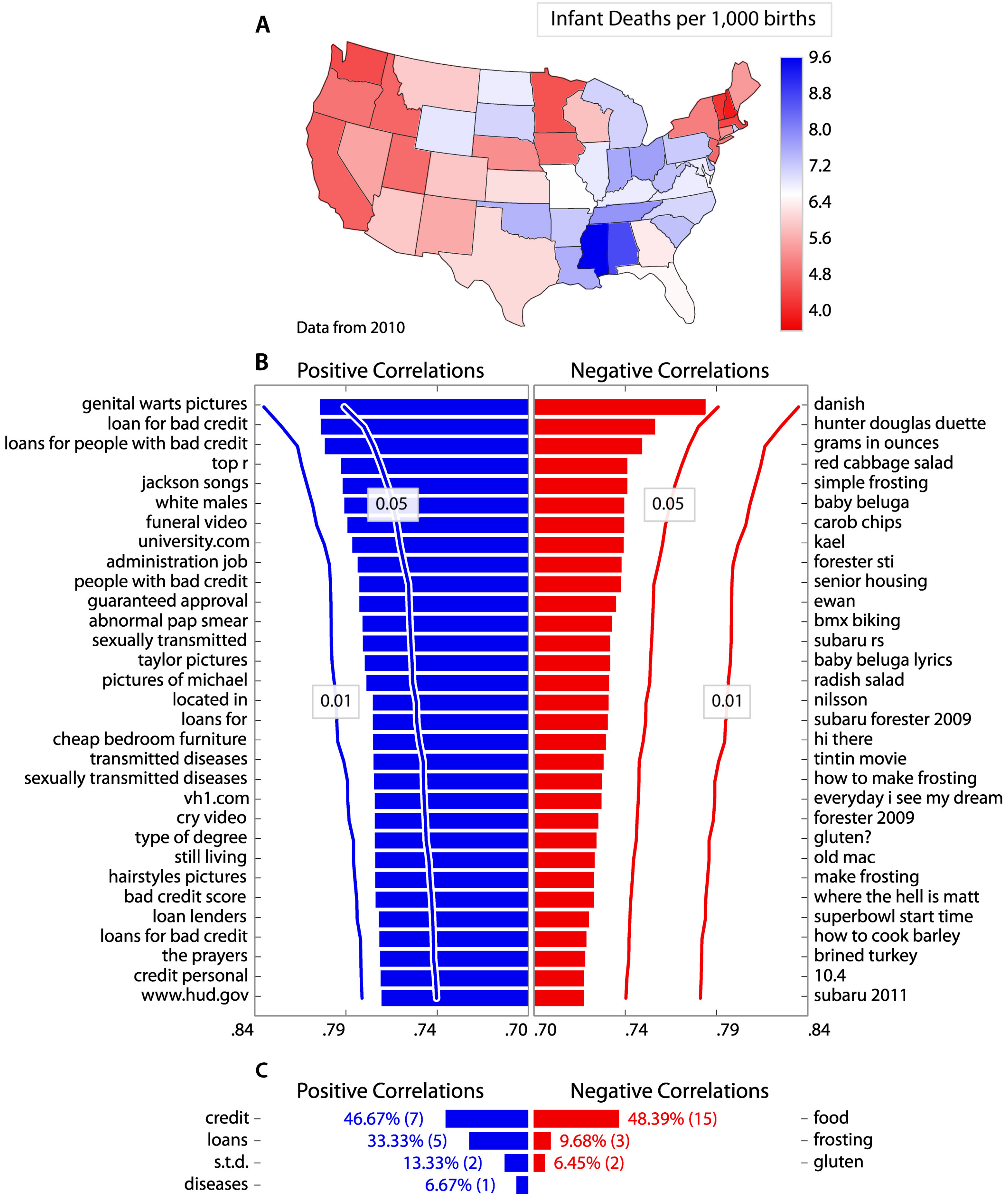

Fig 2. How do *Google* queries vary with infant mortality rate? (**A**) Infant mortality rates for each state in the US. An infant is defined as any person one year old or younger. Infant mortality rate is defined as the number of infant deaths per 1,000 births. (**B**) In a similar fashion to our investigation of birth rates (Fig 1), we use *Google Correlate* to find terms for which the number of searches is higher in U.S. states with higher infant mortality rates, and with lower infant mortality rates. We list the 31 terms for which differences in search volume across U.S. states shows the strongest positive correlation (left) and negative correlation (right) with state wide infant mortality rate. Again, we generate 1,000 random samples from a multivariate Gaussian distribution where states

which are closer together tend to have a similar value. We submit these samples to *Google Correlate* and build a distribution of correlation coefficients for each of the 31 top most search terms. We depict the strength of correlation required for the correlation to be significant at the $p < 0.05$ and $p < 0.01$ level, given this null hypothesis distribution. (**C**) Again, we ask *Amazon Mechanical Turk* users to identify the most prominent topic in each of these lists of terms. We depict all survey responses which account for more than 5% of submitted responses, along with the percentage and number of respondents who gave each response. Our results suggest that users in states with higher infant mortality rates search for more information about credit and loans, as well as sexually transmitted diseases (all search terms $p < 0.05$).

doi:10.1371/journal.pone.0149025.g002

In our model, two states are more dependent on each other if they are physically closer together. We imagine that the states are like vertices in a graph and the edges represent shared borders. The distance between any two states is the length of the shortest path between them. For example, California and Nevada have a distance of 1 because they share a border, that is, they are connected on the network. California and Utah have a distance of 2 because the shortest path between them is two edges long. Each state's distance from itself is zero. The distance between all US states and Alaska and Hawaii is set to ∞.

We specify a distance matrix, **D**, where each cell is the squared distance between two states. We then write the covariance matrix as a Gaussian function of **D**:

$$\mathbf{K} = b \cdot e^{-a\mathbf{D}} + c\mathbf{I} \tag{2}$$

We select the parameters a, b and c by maximising the likelihood function of both the birth rates and infant mortality rates. The log likelihood function of the parameters given Eq (1) is:

$$\log p(\mathbf{y}|a,b,c) = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \tag{3}$$

and the log likelihood of the parameters given both the birth rates and infant mortality rates is:

$$\log p(\mathbf{y}_b|a,b,c) + \log p(\mathbf{y}_m|a,b,c) \tag{4}$$

where $\mathbf{y}_b$ represents the birth rates and $\mathbf{y}_m$ represents the infant mortality rates. We use the downhill simplex algorithm [50] to maximise Eq (4). We run the algorithm 10 times with the starting values for a, b, and c drawn from a standard Gaussian distribution and select the result which maximises the log likelihood. The values we find are $a = 0.0816628$, $b = 0.75687803$, $c = 0.44543885$.

Previous studies have used other methods of quantifying the relationship between geographic regions. For example, the spread of epidemics can be modelled with the air traffic connecting global regions [51, 52]. Future analysis could investigate whether a different metric of distance between states would improved the fit of the multivariate Gaussian to the demographic data.

We generate 1,000 random samples from this multivariate Gaussian distribution and submit each one to *Google Correlate*. For each sample, *Google Correlate* returns a maximum of 100 terms for which search volume is most correlated with the sample and where the Pearson's correlation coefficient is equal to or above 0.6. The left panel of Figure C in S1 Text depicts the distribution of Pearson's correlations for the random samples. We compile the cumulative distribution function (CDF) of the correlation coefficient for each k^{th} most correlated search term. These distributions represent the distribution of the correlation coefficient we would expect under the null hypothesis that the submitted dataset is drawn from a multivariate Gaussian distribution with mean 0 and covariance **K**, with no relationship to the k^{th} search term. The right panel of Figure C in S1 Text shows the CDF of the first ($k = 1$) search term. We use these distributions to statistically test the correlation coefficient of the search volume of each search term with both the birth rate and infant mortality rate data.

Results

We find that all search terms that *Google Correlate* lists as both positively and negatively correlated with birth rates are statistically significant at the $p < 0.05$ level. Only the least most positively correlated term, "baby car seat", is not significant at the $p < 0.05$ level. All terms that are most positively correlated with infant death rates are all significant at the $p < 0.05$ level. The most negatively correlated terms with infant death rates are not significantly correlated.

Discussion

In this study, we investigate how searches of *Google* users vary across U.S. states with different birth rates, by using the service *Google Correlate*. We find that as the number of babies born per 1,000 inhabitants increases, the number of searches for information about pregnancy also increases, as one might expect. However, as birth rate decreases, our analysis reveals increases in the number of searches about cats.

In a second analysis, we consider differences in search activity in states with different infant mortality rates. We find that as the proportion of babies who do not live until the age of one increases, the number of searches for information about credit, loans and sexually transmitted diseases also increases.

Previous studies have demonstrated how data on *Google* usage retrieved from the *Google Trends* interface can reveal interesting relationships between online behaviour and various measures of behaviour in the real world, such as reports of infections of influenza like illnesses [44, 53], fluctuations in stock markets [38, 39, 54, 55], measures of risk of investment [14, 56] and unemployment claims [42].

However, search volume data can only be retrieved from *Google Trends* if the user specifies the search terms of interest. Researchers interested in the link between *Google* usage and real world behaviour may select a set of terms which they believe to be related to the behaviour of interest, or generate lists of terms which cover a range of different topics [38]. It is not possible to submit data relating to real world behaviour and automatically retrieve search terms where search behaviour reflects the submitted real world data.

Comparison of search behaviour across geographic areas is also challenging when using the *Google Trends* interface. When data is requested for a specific geographic area, *Google Trends* scales the maximum value in the retrieved data to 100. For this reason, data retrieved for multiple geographic areas cannot be directly compared, unless data for two keywords is retrieved simultaneously and the ratio between these two keywords is calculated [29, 57].

The *Google Correlate* service offers a solution to both of these problems. Users are able to input data which varies across time or across US states, and retrieve search terms for which the frequency of searches is most correlated with the input data. The interface also returns the strength of correlation for each of these search terms. However, no method has been proposed to determine whether correlations of the observed strength might be expected simply as a result of *Google Correlate* evaluating search volume data for an extremely large number of search terms.

In this paper, we demonstrate how *Google Correlate* can be used to identify the search terms for which search activity is most correlated with the real world data provided—for example, per state birth rates or infant mortality rates. Crucially, we develop a statistical test to determine how likely it would be to observe correlations of this strength under a null hypothesis of no relationship between the search term and the real world data. According to this method, all but one of the terms for which search volume is most positively and negatively correlated with birth rates are significantly correlated at the $p < 0.05$ level. The terms for which search volume is most positively correlated with infant mortality rates are significant at the 0.05% level.

However, we find no evidence that the strength of the correlations for the terms most negatively correlated with infant mortality rates is significant.

We highlight that the presence of relationships at the aggregate level does not imply the presence of similar relationships at the individual level. For example, our finding that Internet users in states with lower birth rates search for more information about cats does not allow us to conclude that individuals with lower birth rates search for more information about cats. Furthermore, while our statistical method allows us to demonstrate a significant correlation between interest in certain search terms and demographics, our analysis does not imply causation. For example, poor education and low wages might be a factor that causes a decrease in infant survival rates as well as interest in credit, loans and sexually transmitted diseases.

In this paper, we propose a method to statistically evaluate search behaviour data provided by *Google Correlate*. The results of our two case studies suggest that appropriate analyses of Internet search data could offer new insights into the concerns of different demographics. Combined with data on real world economic and health variables, search engine data may allow us to gain a better understanding of the different worlds experienced by different sectors of society.

Supporting Information

S1 Text. Contains details on the *Amazon Mechanical Turk* survey and the bootstrapped statistical test. This document describes a survey conducted on the online *Amazon Mechanical Turk* system. It includes the raw responses and the methods we used to clean the data. We also include in this document figures of the distribution of Pearson's correlation as returned by *Google Correlate*.
(PDF)

S1 Data. The data. Contains all results from *Google Correlate* used in this study.
(TXT)

Acknowledgments

The authors acknowledge the support of the Research Councils UK Digital Economy Theme via grant EP/K039830/1.

The maps in Figs 1A and 2A were created using *matplotlib* [58] with geographic data obtained from *Natural Earth* [59]. Permission to use *matplotlib* granted by licence http://matplotlib.org/users/license.html accessed on December 18th 2015. Permission to use *Natural Earth* geographic data is granted by a Public Domain CC licence which can be found at http://www.naturalearthdata.com/about/terms-of-use/.

The results of calculations displayed in Figs 1B and 2B are derived from data returned by *Google Correlate* (http://correlate.googlelabs.com).

Author Contributions

Conceived and designed the experiments: AL TP HSM. Performed the experiments: AL TP HSM. Analyzed the data: AL TP HSM. Contributed reagents/materials/analysis tools: AL TP HSM. Wrote the paper: AL TP HSM.

References

1. Conte R, Gilbert N, Bonelli G, Cioffi-Revilla C, Deffuant G, Kertész J, et al. Manifesto of computational social science. Eur Phys J ST. 2012; 214(1):325–346. doi: 10.1140/epjst/e2012-01697-8

2. King G. Ensuring the Data-Rich Future of the Social Sciences. Science. 2011; 331(6018):719–721. doi: 10.1126/science.1197872 PMID: 21311013
3. Lazer D, Pentland A, Adamic L, Aral S, Barabási AL, Brewer D, et al. Computational Social Science. Science. 2009; 323(5915):721–723. doi: 10.1126/science.1167742 PMID: 19197046
4. Moat HS, Preis T, Olivola CY, Liu C, Chater N. Using big data to predict collective behavior in the real world. Behav Brain Sci. 2014; 37(1):92–93. doi: 10.1017/S0140525X13001817 PMID: 24572233
5. Vespignani A. Predicting the Behavior of Techno-Social Systems. Science. 2009; 325(5939):425–428. doi: 10.1126/science.1171990 PMID: 19628859
6. Watts DJ. A twenty-first century science. Nature. 2007; 445(7127):489. doi: 10.1038/445489a PMID: 17268455
7. Santi P, Resta G, Szell M, Sobolevsky S, Strogatz SH, Ratti C. Quantifying the benefits of vehicle pooling with shareability networks. Proc Natl Acad Sci USA. 2014; 111(37):13290–13294. doi: 10.1073/pnas.1403657111 PMID: 25197046
8. Letchford A, Moat HS, Preis T. The advantage of short paper titles. Royal Society Open Science. 2015; 2(8):150266. doi: 10.1098/rsos.150266 PMID: 26361556
9. Letchford A, Preis T, Moat HS. The advantage of simple paper abstracts. Journal of Informetrics. 2016; 10(1):1–8. doi: 10.1016/j.joi.2015.11.001
10. Gonzalez MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. Nature. 2008; 453(7196):779–782. doi: 10.1038/nature06958 PMID: 18528393
11. Roth C, Kang SM, Batty M, Barthélemy M. Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. PLOS ONE. 2011; 6(1):e15923. doi: 10.1371/journal.pone.0015923 PMID: 21249210
12. Alanyali M, Moat HS, Preis T. Quantifying the Relationship Between Financial News and the Stock Market. Sci Rep. 2013; 3:3578. doi: 10.1038/srep03578 PMID: 24356666
13. Gabaix X, Gopikrishnan P, Plerou V, Stanley HE. A theory of power-law distributions in financial market fluctuations. Nature. 2003; 423(6937):267–270. doi: 10.1038/nature01624 PMID: 12748636
14. Kristoufek L. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Sci Rep. 2013; 3:3415. doi: 10.1038/srep03415 PMID: 24301322
15. Podobnik B, Horvatic D, Petersen AM, Stanley HE. Cross-correlations between volume change and price change. Proc Natl Acad Sci USA. 2009; 106(52):22079–22084. doi: 10.1073/pnas.0911983106 PMID: 20018772
16. Preis T, Schneider JJ, Stanley HE. Switching processes in financial markets. Proc Natl Acad Sci USA. 2011; 108(19):7674–7678. doi: 10.1073/pnas.1019484108 PMID: 21521789
17. Botta F, Moat HS, Stanley HE, Preis T. Quantifying Stock Return Distributions in Financial Markets. PLoS ONE. 2015; 10(9):e0135600. doi: 10.1371/journal.pone.0135600 PMID: 26327593
18. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. J Comput Sci. 2011; 2(1):1–8. doi: 10.1016/j.jocs.2010.12.007
19. Ciulla F, Mocanu D, Baronchelli A, Gonçalves B, Perra N, Vespignani A. Beating the news using social media: the case study of American Idol. EPJ Data Sci. 2012; 1:8. doi: 10.1140/epjds8
20. Gonçalves B, Perra N, Vespignani A. Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. PLOS ONE. 2011; 6(8):e22656. doi: 10.1371/journal.pone.0022656 PMID: 21826200
21. Halu A, Zhao K, Baronchelli A, Bianconi G. Connect and win: The role of social networks in political elections. EPL. 2013; 102(1):16002.
22. Mestyán M, Yasseri T, Kertész J. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLOS ONE. 2013; 8(8):e71226. doi: 10.1371/journal.pone.0071226 PMID: 23990938
23. Moat HS, Curme C, Avakian A, Kenett DY, Stanley HE, Preis T. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. Sci Rep. 2013; 3:1801. doi: 10.1038/srep01801
24. Mocanu D, Baronchelli A, Perra N, Gonçalves B, Zhang Q, Vespignani A. The Twitter of Babel: Mapping World Languages through Microblogging Platforms. PLOS ONE. 2013; 8(4):e61981. doi: 10.1371/journal.pone.0061981 PMID: 23637940
25. Yasseri T, Sumi R, Rung A, Kornai A, Kertész J. Dynamics of Conflicts in Wikipedia. PLOS ONE. 2012; 7(6):e38869. doi: 10.1371/journal.pone.0038869 PMID: 22745683
26. Barchiesi D, Preis T, Bishop S, Moat HS. Modelling human mobility patterns using photographic data shared online. R Soc Open Sci. 2015; 2(8). doi: 10.1098/rsos.150046 PMID: 26361545
27. Barchiesi D, Moat HS, Alis C, Bishop S, Preis T. Quantifying International Travel Flows Using Flickr. PLOS ONE. 2015; 10(7):e0128470. doi: 10.1371/journal.pone.0128470 PMID: 26147500

28. Botta F, Moat HS, Preis T. Quantifying crowd size with mobile phone and Twitter data. R Soc Open Sci. 2015; 2(5):150162. doi: 10.1098/rsos.150162 PMID: 26064667

29. Preis T, Moat HS, Stanley HE, Bishop SR. Quantifying the Advantage of Looking Forward. Sci Rep. 2012; 2:350. doi: 10.1038/srep00350 PMID: 22482034

30. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, The Google Books Team, et al. Quantitative Analysis of Culture Using Millions of Digitized Books. Science. 2011; 331(6014):176–182. doi: 10.1126/science.1199644 PMID: 21163965

31. Petersen AM, Tenenbaum J, Havlin S, Stanley HE. Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death. Sci Rep. 2012; 2:313. doi: 10.1038/srep00313 PMID: 22423321

32. Perc M. Evolution of the most common English words and phrases over the centuries. J R Soc Interface. 2012; 9:3323–3328. doi: 10.1098/rsif.2012.0491 PMID: 22832364

33. Petersen AM, Tenenbaum JN, Havlin S, Stanley HE, Perc M. Languages cool as they expand: allometric scaling and the decreasing need for new words. Sci Rep. 2012; 2:943. doi: 10.1038/srep00943 PMID: 23230508

34. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci USA. 2009; 106(51):21484–21489. doi: 10.1073/pnas.0906910106 PMID: 20018697

35. Pastor-Satorras R, Vespignani A. Epidemic Spreading in Scale-Free Networks. Phys Rev Lett. 2001; 86(14):3200–3203. doi: 10.1103/PhysRevLett.86.3200 PMID: 11290142

36. Seresinhe CI, Preis T, Moat HS. Quantifying the Impact of Scenic Environments on Health. Scientific Reports. 2015; 5:16899. doi: 10.1038/srep16899 PMID: 26603464

37. Song C, Qu Z, Blumm N, Barabási AL. Limits of Predictability in Human Mobility. Science. 2010; 327 (5968):1018–1021. doi: 10.1126/science.1177170 PMID: 20167789

38. Curme C, Preis T, Stanley HE, Moat HS. Quantifying the semantics of search behavior before stock market moves. Proc Natl Acad Sci USA. 2014; 111(32):11600–11605. doi: 10.1073/pnas.1324054111 PMID: 25071193

39. Preis T, Moat HS, Stanley HE. Quantifying Trading Behavior in Financial Markets Using Google Trends. Sci Rep. 2013; 3:1684. doi: 10.1038/srep01684 PMID: 23619126

40. Askitas N, Zimmermann KF. Google Econometrics and Unemployment Forecasting. Appl Econ Quart. 2009; 55(2):107–120. doi: 10.3790/aeq.55.2.107

41. Brownstein JS, Freifeld CC, Madoff LC. Digital Disease Detection—Harnessing the Web for Public Health Surveillance. N Engl J Med. 2009; 360(21):2153–2157. doi: 10.1056/NEJMp0900702 PMID: 19423867

42. Choi H, Varian HAL. Predicting the Present with Google Trends. Econ Rec. 2012; 88:2–9. doi: 10.1111/j.1475-4932.2012.00809.x

43. Ettredge M, Gerdes J, Karuga G. Using Web-based Search Data to Predict Macroeconomic Statistics. Commun ACM. 2005; 48(11):87–92. doi: 10.1145/1096000.1096010

44. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457(7232):1012–1014. doi: 10.1038/nature07634 PMID: 19020500

45. Preis T, Reith D, Stanley HE. Complex dynamics of our economic life on different scales: insights from search engine query data. Phil Trans R Soc A. 2010; 368:5707–5719. doi: 10.1098/rsta.2010.0284 PMID: 21078644

46. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ. Predicting consumer behavior with Web search. Proc Natl Acad Sci USA. 2010; 107(41):17486–17490. doi: 10.1073/pnas.1005962107 PMID: 20876140

47. Mohebbi M, Vanderkam D, Kodysh J, Schonberger R, Choi H, Kumar S. Google Correlate Whitepaper. Google; 2011.

48. Vanderkam D, Schonberger R, Rowley H, Kumar S. Nearest Neighbor Search in Google Correlate. Google; 2013.

49. Fisher RA. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika. 1915; 10(4):507–521. doi: 10.2307/2331838

50. Nelder JA, Mead R. A Simplex Method for Function Minimization. Comput J. 1965; 7(4):308–313.

51. Colizza V, Barrat A, Barthelemy M, Vespignani A. The role of the airline transportation network in the prediction and predictability of global epidemics. Proc Natl Acad Sci USA. 2006; 103(7):2015–2020. doi: 10.1073/pnas.0510525103 PMID: 16461461

52. Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena. Science. 2013; 342(6164):1337–1342. doi: 10.1126/science.1245200 PMID: 24337289

53. Preis T, Moat HS. Adaptive nowcasting of influenza outbreaks using Google searches. R Soc Open Sci. 2014; 1:140095. doi: 10.1098/rsos.140095 PMID: 26064532

54. Moat HS, Curme C, Stanley HE, Preis T. Anticipating Stock Market Movements with Google and Wikipedia. In: Matrasulov D, Stanley HE, editors. Nonlinear Phenomena in Complex Systems: From Nano to Macro Scale. NATO Science for Peace and Security Series C: Environmental Security. Springer Netherlands; 2014. p. 47–59.

55. Preis T, Moat HS. Early Signs of Financial Market Moves Reflected by Google Searches. In: Gonçalves B, Perra N, editors. Social Phenomena. Computational Social Sciences. Springer International Publishing; 2015. p. 85–97.

56. Kristoufek L. Can Google Trends search queries contribute to risk diversification? Sci Rep. 2013; 3:2713. doi: 10.1038/srep02713 PMID: 24048448

57. Noguchi T, Stewart N, Olivola CY, Moat HS, Preis T. Characterizing the Time-Perspective of Nations with Search Engine Query Data. PLoS ONE. 2014; 9(4):e95209. doi: 10.1371/journal.pone.0095209 PMID: 24736725

58. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007; 9(3):90–95. doi: 10.1109/MCSE.2007.55

59. Natural Earth; 2014. Accessed: 2014-09-02. Available from: http://www.naturalearthdata.com/.