Nicotine & Tobacco Research, 2018, 779–783
doi:10.1093/ntr/ntx143
Brief report
Received September 7, 2016; Editorial Decision June 12, 2017; Accepted June 19, 2017

Brief report

Google Searches for "Cheap Cigarettes" Spike at Tax Increases: Evidence from an Algorithm to Detect Spikes in Time Series Data

Theodore L. Caputi[1,2]

[1]The Wharton School, University of Pennsylvania, Philadelphia, PA; [2]Drug Policy Institute, Department of Psychiatry, College of Medicine, University of Florida, Gainesville, FL

Corresponding Author: Theodore L. Caputi, 3730 Walnut Street, Suite G95 Research Office, Philadelphia, PA 19104, USA, Telephone: 267-312-8471; Fax: 215-862-9894 E-mail: tcaputi@wharton.upenn.edu

Abstract

Introduction: Online cigarette dealers have lower prices than brick-and-mortar retailers and advertise tax-free status.[1–8] Previous studies show smokers search out these online alternatives at the time of a cigarette tax increase.[9,10] However, these studies rely upon researchers' decision to consider a specific date and preclude the possibility that researchers focus on the wrong date. The purpose of this study is to introduce an unbiased methodology to the field of observing search patterns and to use this methodology to determine whether smokers search Google for "cheap cigarettes" at cigarette tax increases and, if so, whether the increased level of searches persists.
Methods: Publicly available data from Google Trends is used to observe standardized search volumes for the term, "cheap cigarettes". Seasonal Hybrid Extreme Studentized Deviate and E-Divisive with Means tests were performed to observe spikes and mean level shifts in search volume.
Results: Of the twelve cigarette tax increases studied, ten showed spikes in searches for "cheap cigarettes" within two weeks of the tax increase. However, the mean level shifts did not occur for any cigarette tax increase.
Conclusion: Searches for "cheap cigarettes" spike around the time of a cigarette tax increase, but the mean level of searches does not shift in response to a tax increase. The SHESD and EDM tests are unbiased methodologies that can be used to identify spikes and mean level shifts in time series data without an *a priori* date to be studied. SHESD and EDM affirm spikes in interest are related to tax increases.
Implications:

- Applies improved statistical techniques (SHESD and EDM) to Google search data related to cigarettes, reducing bias and increasing power
- Contributes to the body of evidence that state and federal tax increases are associated with spikes in searches for cheap cigarettes and may be good dates for increased online health messaging related to tobacco

Introduction

Researchers have found that, traditionally, cigarette taxes are an effective policy to reduce cigarette consumption and/or increase state revenues.[1] However, in a time when online shopping gives the average consumer access to hundreds or thousands of alternative vendors (including those who advertise tax-free wares) with just a few clicks, smokers may seek to mitigate the burden of these tax increases by searching online for cheaper cigarettes. Researchers have adequately described the alternative markets for cigarettes available online.[2–8] Though searches for cheaper cigarettes do not necessarily translate into online cigarette purchases, the notion that tax increases may encourage smokers to (A) circumvent tax policies and/or (B) search out cheaper alternative channels concerns policy makers and piques

the interest of researchers. Indeed, previous studies[9,10] have used search volume data to show that searches for cigarette market alternatives spike at the time of state and federal tax increases for tobacco. However, a major limitation to these studies is that their methodologies require a decision by the researchers to focus on a specific date—the date of the tax increase. This decision precludes the possibility that other dates are more important predictors of search spikes for cheap cigarettes and/or mean level shifts in the search volumes for cheaper cigarettes, possibly biasing the results.

The present study contributes to the body of work relating tax increases to search volumes by introducing two statistical techniques not previously used in substance use research, the Seasonal Hybrid Extreme Studentized Deviate (SHESD)[11] and the E-Divisive with Means (EDM)[12] tests. The results of the SHESD and EDM tests are the identification of patterns referred to as "spikes" and "breakouts" (respectively). A spike is a sudden increase in searches for a specific term, whereas a breakout is a sudden shift in the mean level of searches. The literature[9,10] would suggest that "spikes" are more likely a result of cigarette tax increases, and so the analysis is focused on spikes. Previous tests using similar data have also sought to identify spikes and breakouts. However, the SHESD and EDM identify spikes and breakouts in time series data without requiring the researcher to specify a certain date of interest, while techniques such as least-square regression forecasting require researchers to define a specific date for analysis. Therefore, the SHESD and EDM tests can potentially reduce bias resulting from researchers choosing to study a date that matches their *a priori* hypothesis.

Beyond the improvement to study bias, the SHESD and EDM are designed to exclude other anomalies in the dataset during the analysis through a recursive data reduction process (see the methods section), increasing the power of the tests. Previously used techniques (e.g., relative mean increases,[10] least squares forecasting[9]) lack the ability to detect anomalies in otherwise "noisy" data.

The increased precision of the SHESD and EDM tests relative to tests used in previous studies adds to the power of statistical analysis relative to previously used tests. In this domain—search volumes related to cigarette tax increases—the SHESD and EDM make possible analysis of state-level (rather than just federal) cigarette tax increases, which by virtue of having fewer "searchers" tend to have increased variance. There have been relatively few federal tax increases in the United States; the vast majority of cigarette tax policies are enacted at the state level. Therefore, an analysis of the search volume data with SHESD and EDM can determine whether search queries spike or "breakout" in the state where a cigarette tax is imposed and can be used directly by state lawmakers, who are more likely to consider cigarette tax policy in the near future.

Methods

The analysis is performed on Google.com searches for "cheap cigarettes" in 11 states that had major cigarette tax increases (>$1.00 per pack) and the United States as a whole, which had a $0.62 cigarette tax increase in 2009 (see table 1). The states were selected based on Google Trends data availability. Certain states that experienced a major cigarette tax increase (e.g., Iowa in 2007, South Dakota in 2007, Utah in 2010) are excluded from the analysis Google Trends data was not available for these states, presumably due to low search volumes. For states with more than one major tax increase after 2004, I analyzed the later tax increase to mitigate the possibility of over-reliance on certain states. The time period for each analysis is given as one year before and after each respective cigarette tax increase.

The data used for this study is publicly available on http://www.google.com/trends. Google Trends provides weekly data dating back to January 2004 on Google search volumes by search term, geography, and timeframe.

The raw number of searches is not made available by Google. Instead, the search volume data made available is standardized on a scale of 0 to 100 such that search volumes are reported weekly as a percentage of the highest search volume for a given time period in a given geographic region. For this reason, search volumes should only be considered on a relative basis by both geographic region and search term. However, this does not limit the analysis in this study, as the purpose of the study is to measure relative spikes in search volume.

The Seasonal Hybrid Extreme Studentized Deviate test (SHESD) with an alpha of 0.05 is used to determine anomalies or spikes in the dataset. SHESD combines the Student t-distribution with a generalized form of the Extreme Standardized Deviate test (ESD), along with an adjustment for seasonality (see[13] for a complete description).

The purpose of the SHESD is to identify anomalies in the data. Before one can identify a point as an anomaly, however, we must assume the distribution of the dataset. Because of the relatively weak assumptions required, the SHESD assumes a Students t-distribution, a sample of normally distributed data with unknown variance. The distribution is described mathematically as:

$$f(x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v*\pi}*\Gamma(\frac{v}{2})}*(1+\frac{1}{v}*t^2)^{\frac{-v+1}{2}}$$

Table 1. Tax Increases and Observation Periods

State	Effective Date	Tax Increase Amount	Beginning of Observation Period	End of Observation Period
Texas	1/1/2007	$1.00	12/1/2005	2/1/2008
Maryland	1/1/2008	$1.00	12/1/2006	2/1/2009
Wisconsin	1/1/2008	$1.00	12/1/2006	2/1/2009
Washington DC	10/1/2008	$1.00	9/1/2007	11/1/2009
United States	4/1/2009	$0.62	3/1/2008	5/1/2010
Florida	7/1/2009	$1.00	6/1/2008	8/1/2010
Connecticut	10/1/2009	$1.00	9/1/2008	11/1/2010
Washington	5/1/2010	$1.00	4/1/2009	6/1/2011
New York	7/1/2010	$1.60	6/1/2009	8/1/2011
Illinois	6/24/2012	$1.00	5/24/2011	7/24/2013
Minnesota	7/1/2013	$1.60	6/1/2012	8/1/2014
Massachusetts	7/31/2013	$1.00	6/30/2012	8/31/2014

SHESD first applies ESD using the student's t-distribution to determine if there is a single outlier in the data. The furthest point from the mean of the sample is taken and given a value G.

$$G = \frac{|Y_i - \overline{Y}|}{s}$$

The researcher decides on a level of alpha (in this analysis, alpha = 0.05), which refers to the maximum probability that an observation would exist in the dataset by chance in order for it to be considered an anomaly. The given alpha corresponds to a "critical value", which is the threshold value above which an observation would be considered an anomaly. If G is larger than a critical value, then we describe the point as an outlier (as described mathematically below):

$$G > \frac{N-1}{\sqrt{N}} * \sqrt{\frac{t^2_{\frac{\alpha}{2N},N-1}}{N-2+t^2_{\frac{\alpha}{2N},N-1}}}$$

Note: $t^2_{\frac{\alpha}{2N},N-1}$ is the upper critical value with N-2 degrees of freedom.

However, ESD only gives a singular outlier. A generalized ESD (GESD) runs the ESD as an algorithm to find up to a pre-determined number of outliers, *r*. In GESD, ESD is run on the whole dataset to determine if the point furthest from the dataset is an outlier. If it is, then that number is removed from the dataset and a second ESD is run to determine if the point furthest from the mean of the new dataset is an outlier. The algorithm continues until the GESD finds the remaining point furthest from the mean is not an outlier or *r* outliers are found. In this way, the GESD is a robust statistical technique used to determine if several outliers exist in the data.

However, searches can occur with seasonality, and so the GESD must be adjusted. The SHESD decomposes the data into seasons, a time trend, and a local regression using the Loess method. The Loess method (explained in detail by Fox[14]) is a nonparametric form of regression whereby subsets of data are determined by a "nearest-neighbor" algorithm and then each subset is smoothed to either a linear or quadratic function through a weighted least-squares regression, resulting in a smooth curve mapping the trend of the data without the potential for model selection bias by the researcher. In SHESD, the GESD is run within the decomposed data.

All SHESD analyses were run in R using the Open Source package "AnomalyDetection". Code is available upon request.

In addition to the SHESD, an E-Divisive with Medians (EDM) test was conducted to find "breakouts" or mean shifts in the data. The EDM test evaluates a dataset to distinguish spikes (or anomalies) from mean changes, i.e. a change in the steady state of the time series, regardless of the noise. This test was performed on the same dataset using the R package "BreakoutDetection".

From this study, it becomes clear how future researchers can leverage the unbiased advantage of SHESD and EDM to improve time series analysis in several health-related domains. From a more generalized perspective, researchers can use this technique with time series data to discover events of consequence without knowing of the events' existence before the analysis.

Results

For 10 of the 12 analyses, a positive anomaly or spike was detected within two weeks of the cigarette tax increase (see Table 2). Spikes before the tax increase are also considered in the analysis because cigarette smokers may not only respond to an actual increase in prices but also increased risk of a tax increase if it is being deliberated or increased press of a tax increase that may precede the enactment date. From visual inspection of the plots (see Figure 1), we can see that search spikes are typically quite large at the time of a tobacco tax increase. Spikes were not detected for Washington D.C. or Connecticut, though this may be due to occasional gaps (see Figure 1) in the data, likely due to a relatively small population size.

Table 2. Spikes Occurring Near Tax Increase Dates

State	Spike Date	Tax Increase	Difference (days)
Florida	7/4/2009	6/30/2009	3
Florida	7/11/2009	6/30/2009	10
Illinois	6/30/2012	6/23/2012	6
Illinois	7/7/2012	6/23/2012	13
Maryland	12/22/2007	12/31/2007	−10
Maryland	12/29/2007	12/31/2007	−3
Maryland	1/5/2008	12/31/2007	4
Maryland	1/12/2008	12/31/2007	11
Massachusetts	8/3/2013	7/30/2013	3
Massachusetts	8/10/2013	7/30/2013	10
Minnesota	6/29/2013	6/30/2013	−2
Minnesota	7/6/2013	6/30/2013	5
New York	7/3/2010	6/30/2010	2
New York	7/10/2010	6/30/2010	9
Texas	1/6/2007	12/31/2006	5
Texas	1/13/2007	12/31/2006	12
United States	3/21/2009	3/31/2009	−11
United States	3/28/2009	3/31/2009	−4
United States	4/4/2009	3/31/2009	3
United States	4/11/2009	3/31/2009	10
Washington	5/8/2010	4/30/2010	7
Wisconsin	12/29/2007	12/31/2007	−3
Wisconsin	1/5/2008	12/31/2007	4

The robustness of these results is important. Spikes consistently happen in states where there is a cigarette tax increase and not nearly as frequently at times when there is no cigarette tax increase. Further, comparing graphs that overlap in time indicates that spikes occur in states where there is a cigarette tax increase and not in states where there is no cigarette tax increase. For example, Massachusetts experienced a tax increase on July 31, 2013 and Minnesota on July 1, 2013, and we see search volume spikes around each of those dates in their respective states. We do not, however, see spikes around July 31, 2013 in Minnesota or spikes around July 1, 2013 in Massachusetts.

Given the robust results of the SHESD, it would seem reasonable that there would be a mean shift in the time series data before and after the tax increase. If people search for cheaper sources for cigarettes at the time of a cigarette tax increase, it is plausible that those consumers may continue to purchase cigarettes from the cheaper online outlet, given that the tax and subsequent price increases are not temporary. However, no breakouts were detected. Searches for "cheap cigarettes" spiked at the time of a tax increase, but the mean level of searches did not shift upwards.

Discussion

The results show clearly that searches for "cheap cigarettes" spiked at the time of cigarette tax increases. However, the mean level of

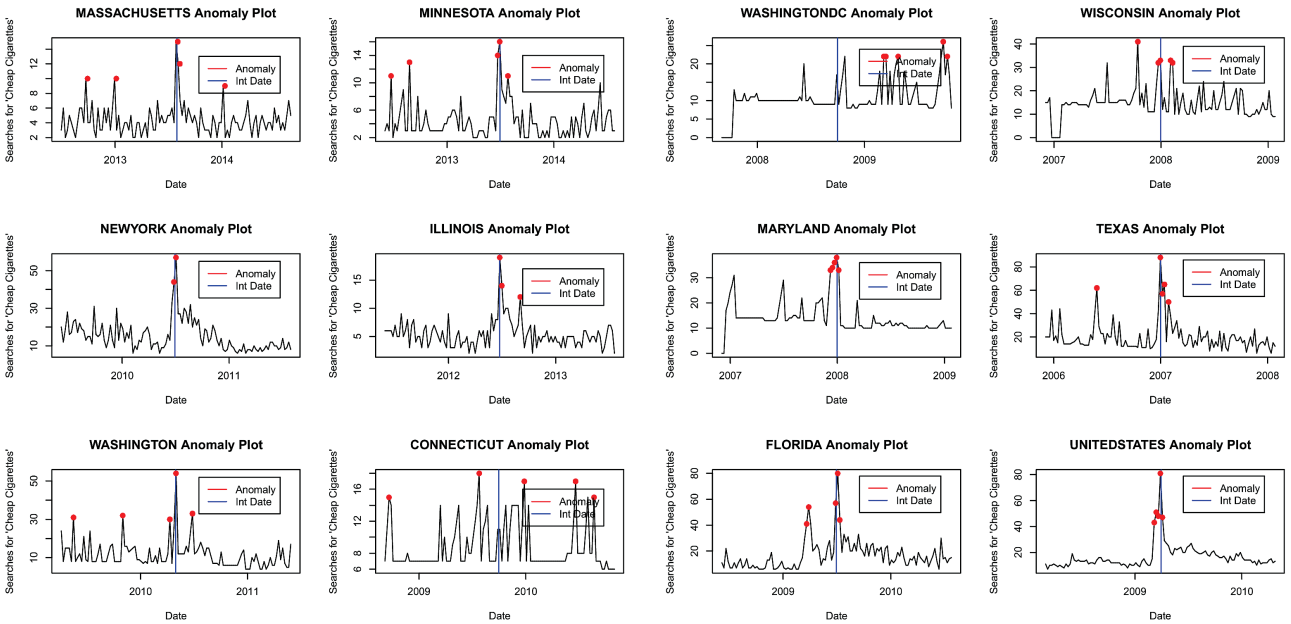

Figure 1. Anomaly plots for search data.

searches for "cheap cigarettes" did not increase meaningfully between the periods before and after a cigarette tax increase.

As previous studies have noted, search volume data does not provide information on the motivation for searches. Therefore, there are several possible interpretations for these results:

The most reasonable conclusion is that cigarette tax increases incentivize people to search for cheaper options for cigarettes. Tax-free cigarettes are widely available online, and smokers searching for "cheap cigarettes" find a way to circumvent the tax increase. However, for some reason, this alternative marketplace does not "stick" for most smokers – perhaps because cigarette delivery is inconvenient or because shipping costs are high for those unwilling to purchase cigarettes in bulk.

A second explanation is that a cohort of price-sensitive smokers performs searches for "cheap cigarettes" around the time of a tax increase. Once they find a suitable marketplace website, however, they no longer need to search Google. Instead, they can navigate directly to the marketplace website, i.e. www.cheap-cigarettes.com.

Less likely explanations include: (1) press coverage of the tobacco tax increase causes people to think about cigarettes and to consider using them and (2) smokers search for "cheap cigarettes" to learn more about the tax increase. Search volume data does not provide information on the searcher's intentions, so we cannot rule out these explanations, however unlikely they seem.

With the results of this study in mind, tobacco control advocates should consider that many smokers may be inclined to circumvent tobacco tax increases using the Internet. Tobacco control advocates and agencies can ameliorate this situation by increasing online messages regarding both smoking cessation and cigarette health risks at the time of tobacco increases. Indeed, tobacco control agencies can prioritize resources to fund anti-smoking advertisements targeting those searching Google for "cheap cigarettes" and similar terms within the weeks before and after a tobacco tax increase.

This study contributes to the literature by introducing a new, powerful test for time series analysis that reduces bias in that researchers do not have to identify a date to study. This technique can be applied to time series analysis in several health domains to improve our understanding of what events and interventions change people's behavior. Researchers can use this technique to discover consequential events or dates that they may never had considered before. Researchers are likely removed from many of the causes that would inspire people to engage in observable behaviors—whether that is searching for cheaper cigarettes on Google, asking for cessation advice on Twitter, or calling into government help lines—and researchers can use the SHESD and EDM techniques to help discern what those events may be from the observable time series data alone.

Declaration of Interests

None declared.

Funding

None.

References

1. Goolsbee A, Lovenheim MF, Slemrod J. Playing with fire: Cigarettes, taxes, and competition from the internet. *AEJ Econ Policy*. 2010;2(1):131–154.
2. Hrywna M, Delnevo CD, Staniewska D. Prevalence and correlates of internet cigarette purchasing among adult smokers in New Jersey. *Tob Control*. 2004;13(3):296–300.
3. Cohen JE, Sarabia V, Ashley MJ. Tobacco commerce on the internet: a threat to comprehensive tobacco control. *Tob Control*. 2001;10(4):364–367.
4. Graff SK. State taxation of online tobacco sales: Circumventing the archaic bright line penned by Quill. *Fla L Rev*. 2006;58:375.
5. Williams RS, Derrick J, Ribisl KM. Electronic cigarette sales to minors via the internet. *JAMA Pediatr*. 2015;169(3):e1563.
6. Bryant JA, Cody MJ, Murphy ST. Online sales: profit without question. *Tob Control*. 2002;11(3):226–227.
7. Ribisl KM, Williams RS, Kim AE. Internet sales of cigarettes to minors. *JAMA*. 2003;290(10):1356–1359.

8. Hall MG, Williams RS, Gammon DG, Ribisl KM. Internet cigarette vendors make tax-free claims and sell cigarettes cheaper than retail outlets. *Tob control*. 2015;25(6):616–618.
9. Ayers JW, Ribisl K, Brownstein JS. Using search query surveillance to monitor tax avoidance and smoking cessation following the United States' 2009 "SCHIP" cigarette tax increase. *PLoS One*. 2011;6(3):e16777.
10. Ayers JW, Althouse BM, Ribisl KM, Emery S. Digital detection for tobacco control: online reactions to the United States' 2009 cigarette excise tax increase. *Nicotine Tob. Res*. 2014;16(5):576–583.
11. Vallis OS, Hochenbaum J, Kejariwal A. (2014). *AnomalyDetection: Anomaly Detection Using Seasonal Hybrid Extreme Studentized Deviate Test*. R Package Version 1.0. https://www.rdocumentation.org/packages/AnomalyDetection/versions/1.0. Accessed June 1, 2016.
12. NAIR KR. The distribution of the extreme deviate from the sample mean and its studentized form. *Biometrika*. 1948;35(Pts 1–2):118–144.
13. Anansi's Calabash. Problem of the month: Anomaly Detection. 2015. https://web.archive.org/web/20160704231229/https:/warrenmar.wordpress.com/2015/02/10/problem-of-the-month-anomaly-detection/. Accessed July 4, 2016.
14. Fox, J. (2015). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: Sage Publications.