The Australian Economic Review
MELBOURNE INSTITUTE of Applied Economic and Social Research
The Australian Economic Review, vol. 50, no. 3, pp. 363–72

For the Student

An Introduction to Data Cleaning Using Internet Search Data

Matthew Greenwood-Nimmo and Kalvinder Shields*

Abstract

This article considers the issue of data cleaning. We use state-level data on internet search activity in the United States to illustrate several common data cleaning tasks, including frequency conversion and data scaling as well as methods for handling sampling uncertainty and accommodating structural breaks and outliers. We emphasise that data cleaning relies on informed judgement and so it is important to maintain transparency through careful documentation of data cleaning procedures.

* Greenwood-Nimmo and Shields: Department of Economics, The University of Melbourne, Victoria 3010 Australia. Corresponding author: Greenwood-Nimmo, email <matthew.greenwood@unimelb.edu.au>. The authors are grateful to the Editor and to two anonymous referees for their constructive feedback on an earlier draft of this article. They thank Victoria Baranov, David Johnston, Artur Tarassow and Michael Shields for their helpful comments. The raw data used here was originally gathered by David Johnston for a project on the mental health implications of the business cycle in the United States.

1. Introduction

Economists are trained to be sophisticated users of data. Techniques for data analysis feature prominently in a typical economics degree program, graduates of which will have studied a variety of techniques for the estimation, calibration and simulation of economic models. However, relatively little attention is paid to the construction of reliable datasets in taught courses. Indeed, much of the data that students encounter has either been professionally curated at source or carefully screened by the instructor. Such data is *clean* in the sense that it is reliable, reproducible and largely free from errors, omissions and biases. However, this is not necessarily true of all—or even most—data that may be of interest to economists. Outside of the classroom, economists will often be faced with *dirty* data—that is, data that suffers from various imperfections including measurement errors, missing observations, duplicate entries, errors in data coding and errors introduced by mismatching data when combining information from multiple sources. *Data cleaning* is the process of attempting to identify and correct such imperfections. However, despite its importance, data cleaning is a subject that receives little attention either in taught courses or in the academic literature (Wickham 2014).

In this article, we provide a practical illustration of a data cleaning problem using an interesting dataset on state-level internet search activity in the United States drawn from Google Trends. This data provides an excellent basis for our use as it exhibits several challenging features that require careful handling, not least of which are the variation in the index base period and the sampling frequency

© 2017 The University of Melbourne, Melbourne Institute of Applied Economic and Social Research
Published by John Wiley & Sons Australia, Ltd

across US states. Given that the dataset has a modest cross-section dimension and a somewhat larger time series dimension, this article is written largely from a time series perspective. However, many of the issues that we address are equally relevant for a broad class of cross-sectional and panel datasets.

This article proceeds as follows. Section 2 provides an introduction to the Google Trends data. Section 3 describes the issues involved in cleaning the dataset, including frequency conversion and redefining the base period used to construct index values, as well as methods for handling sampling uncertainty and accommodating structural breaks and outliers. Section 4 provides summary statistics for the cleaned data and Section 5 concludes the article.

2. Google Trends Data

Data on internet activity and social media usage offers economists insights into aspects of human behaviour that have previously been largely unobservable. Google Trends is one of the leading sources of such data, providing users with access to a summary of the relative frequency of popular internet searches undertaken on the Google search engine.[1] Google is the largest internet search engine by market share, accounting for almost two-thirds of the desktop explicit core search share in the United States as of March 2015 according to the analytics firm comScore.[2] Consequently, internet search activity on Google is likely to provide a good proxy for internet search activity across all search engines.

Several recent papers have used Google data to measure phenomena that may be unobservable or under-reported in traditional datasets. A good example is Tefft (2011), who uses information on the frequency of Google searches for depression and anxiety at the state level in the United States to investigate the relationship between mental health and the business cycle. Much of the official data on depression and anxiety is based on clinical diagnosis and therefore only measures the subset of sufferers who are receiving treatment. However Berger, Wagner and Baker (2005) show that sufferers of stigmatised illnesses—including depression and anxiety—often do not seek professional help but are more likely to use the internet to search for health information than the population at large. This suggests that Google Trends data is likely to capture a broader cross-section of the population suffering with mental health conditions than official datasets.

The Google Trends data is broken down according to:

(i) *Query search terms*. Query search terms are specified by the user and define the Google search activity for which one wishes to obtain data.

(ii) *Query time*. The frequency with which the data is reported (that is, the resolution at which query time is reported) varies according to the frequency with which the search terms have been used in practice—data on less popular searches is reported monthly while data on more popular searches may be available at higher frequency. Consequently, one may obtain a cohort of cross-section units with weekly data and a separate cohort with monthly data, for example.

(iii) *Query location*. The level of spatial disaggregation—that is, whether data is reported by state or by country, for example—varies depending on the region being studied, with disaggregate data coverage for the United States being among the richest available.

(iv) *Query category*. Queries are grouped into categories such as *Health* or *Automotive* for reporting purposes.

Following Tefft (2011), we query the Google Trends database for the search terms 'Depression + Anxiety' within the 'Health' category for the United States. This returns the search frequency for all searches containing 'Depression' and/or 'Anxiety' on a state-by-state basis. The three most common searches containing these terms are 'Anxiety Symptoms', 'Anxiety Disorder' and 'Depression Symptoms'. The search frequency is computed using a random

© 2017 The University of Melbourne, Melbourne Institute of Applied Economic and Social Research

sample of Google searches originating from a specified geographical location within a specified timeframe. The sampling routine excludes repeat searches and searches containing special characters. In addition, 'unpopular' (that is, very infrequent) searches are also excluded from the sample—consequently, the usage of a search term must exceed a threshold if it is to be included in the sample. The value of this threshold is not explicitly reported.

The reported search frequency has been normalised and scaled at source. Let the geographical region be indexed by $i = 1, 2, ..., N$ and the time period by $t = 1, 2, ..., T$. The relative popularity of searches for 'Depression + Anxiety' in region i at time t is expressed as:

$$\tilde{k}_{it} = \frac{K_{it}}{G_{it}} \tag{1}$$

where K_{it} denotes the number of searches for 'Depression + Anxiety' in the sample for region i at time t and G_{it} is the total number of Google searches in the sample for region i at time t. This is then scaled such that:

$$k_{it} = 100 \times \frac{\tilde{k}_{it}}{\max_{t}\left(\tilde{k}_{it}\right)} \tag{2}$$

where $\max\left(\tilde{k}_{it}\right)$ is the maximum value of $\tilde{k}_{it}$ in state i over the time periods $t = 1, 2, ...T$. Hence, k_{it} is an index taking values in the range [0,100]. However, recall that searches that occur at low frequency are excluded from the sample. Consequently, $k_{it} = 0$ does not necessarily imply that no searches for the specified query search terms occurred in region i at time t but rather that $0 \leq k_{it} < c$, where c is the unknown threshold required for inclusion in the sample. Nonetheless, assuming that c is sufficiently small, then $k_{it} = 0$ implies that the search frequency of the specified query search terms is either zero or negligible.

3. Cleaning the Google Trends Data

The raw data extracted from the Google Trends database covers 50 US states plus the District of Columbia (for convenience we will henceforth refer to the District of Columbia as a state) from January 2004 to February 2015 at either weekly or monthly frequency, depending on the amount of search activity in the sample. In addition to this frequency mismatch, the data for several states shows evidence of outlying observations and apparent structural breaks. Outliers and structural breaks both introduce discontinuities into the data—in the case of an outlier, this is usually a one-off deviation from the pattern defined by the remaining observations while a structural break is typically a sustained phenomenon, such as a change in the mean or trend of the data. Consequently, the dataset provides many challenges in terms of data cleaning and we outline our approach on a step-by-step basis below.

3.1 Frequency Conversion

Most common time series regression techniques require that all series share a common sampling frequency. Consequently, we have two options:

(i) Convert the monthly data to weekly frequency by interpolation. Interpolation involves inserting additional synthetic data points in between a set of discrete observations. This is typically achieved by using a smooth function such as a polynomial of a given order to approximate the unobserved variation between the respective observations (that is, the within-month variation in this case). The resulting series is typically rather smooth and its dynamic properties will depend on the choice of the function used in interpolation. As a result, interpolation should be used with care.

(ii) Convert the weekly data to monthly frequency using an appropriate aggregation algorithm. This is not a trivial exercise because a given week may start in one month but end in the next, which implies that not all weeks can be uniquely assigned to a single month. A common approach to overcome this issue is to first convert the weekly data to daily frequency under the

© 2017 The University of Melbourne, Melbourne Institute of Applied Economic and Social Research

simplifying assumption that there is no within-week variation.[3] The daily values can then be converted to monthly frequency without difficulty because each day can be uniquely assigned to a single month. For example, one could compute monthly values by taking the sum or the mean over all days within each month. This approach discards some of the informational content of the higher frequency data but its underlying assumptions are relatively uncontroversial.

The choice between interpolation and aggregation will typically depend on the economic question one seeks to address and the type of model one plans to build. For the purpose of illustration, we pursue the latter option and compute the monthly observations as the mean of the daily observations within each month.

3.2 Rebasing the Data

Recall that the raw Google Trends data for the i th state is normalised and scaled following (1) and (2) such that it takes the value of 100 in the period with the highest search frequency in that state. When comparing across states, there is no reason that these maxima should coincide—that is, the data for each state has a different base period. Furthermore, the normalisation applied by Google is not preserved during our frequency conversion routine. Consequently, we re-base the data to ensure that it is measured on a common scale for all states. This is achieved as follows:

$$d_{it} = 100 \times \frac{k_{it}}{k_{iT}} \tag{3}$$

where d_{it} is the re-based data for state i which takes a value of 100 in the final period, that is, $d_{iT} = 100$.

3.3 Sampling Uncertainty

The data is now comparable across states—specifically, it is now expressed on the same scale and at the same sampling frequency for every state. However, before we present time series plots of the data, we undertake one additional step. Given that we have little information about the sampling routine that Google uses to construct the Google Trends data, we consider 30 different versions of the dataset, each of which was downloaded on a different day between 6 March 2015 and 15 May 2015.[4] Each of these downloads, which we shall index by $j = 1, 2, ...J$, represents a separate draw extracted from the underlying population of Google searches by Google's sampling routine. A comparison of the draws for selected states—once they have been converted to monthly frequency and re-based following the procedures described in Subsections 3.1 and 3.2—is shown in Figure 1.

The variation across draws is non-negligible in all cases and is particularly substantial in some cases, especially among the less populous states. Given this variation, it may be prudent to combine information from multiple draws rather than simply focusing on data from a single draw. To this end, the median across draws, denoted $\tilde{d}_{it}$, is shown by the heavy dashed black line in each panel of Figure 1. We use the median for this purpose rather than the mean as it is more robust in the presence of extreme values.

3.4 Trimming the Sample

Comparing the upper and lower panels of Figure 1 suggests that states can be grouped into three categories. First, for the more populous states (for example, New York, Pennsylvania and Texas), the reported search frequency is typically non-zero throughout the sample period. By contrast, for the least populous states (for example, Vermont), the data contains many zeros over much of the sample, which indicates that the frequency of anxiety and depression searches is below the threshold required for inclusion in Google's sampling routine. Many states (for example, Connecticut and Oregon), however, occupy an intermediate position, with many zeros early in the sample and largely complete data later in the sample.

For this intermediate group, the majority of zero observations occur in 2004 and 2005, with

©2017 The University of Melbourne, Melbourne Institute of Applied Economic and Social Research

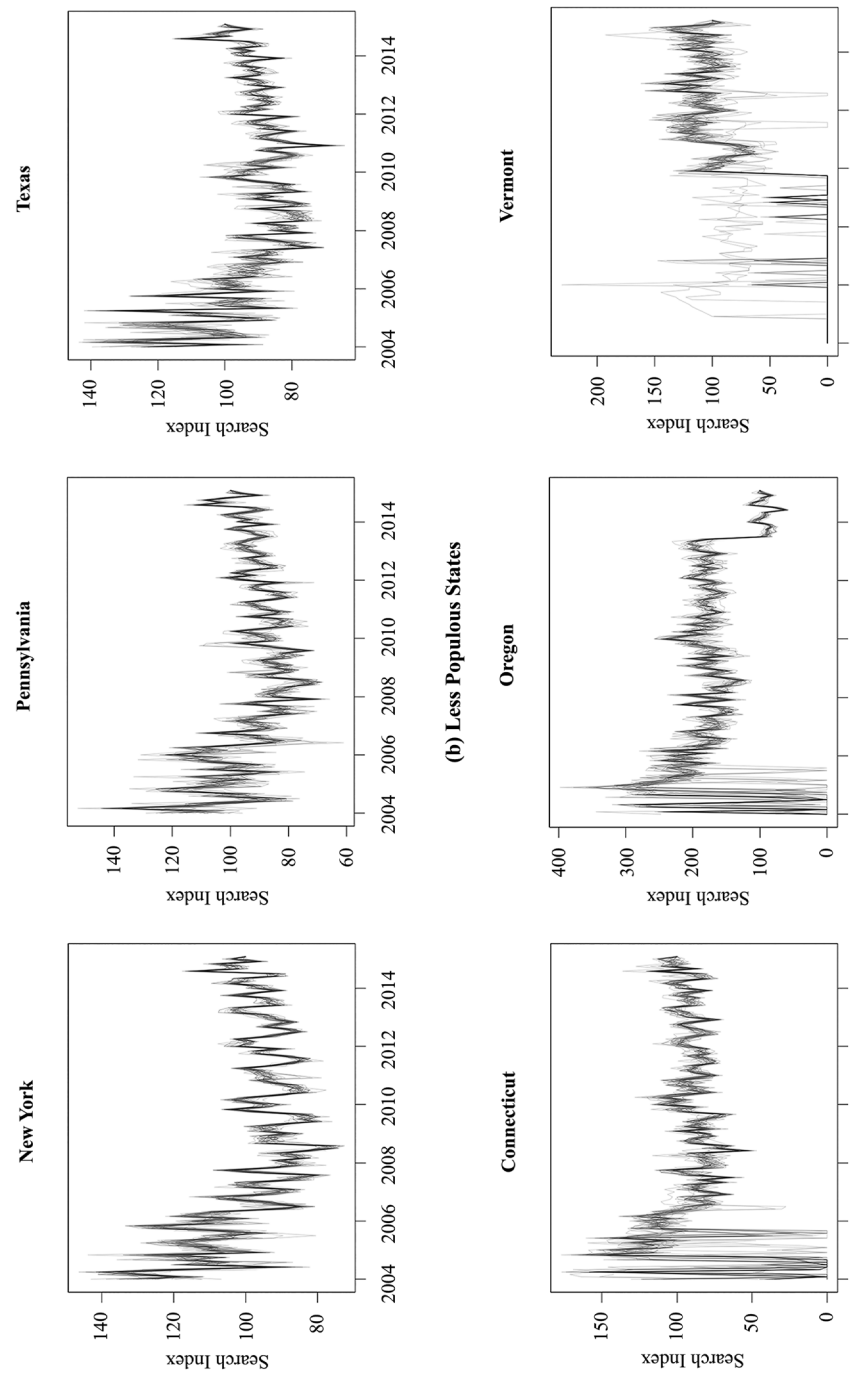
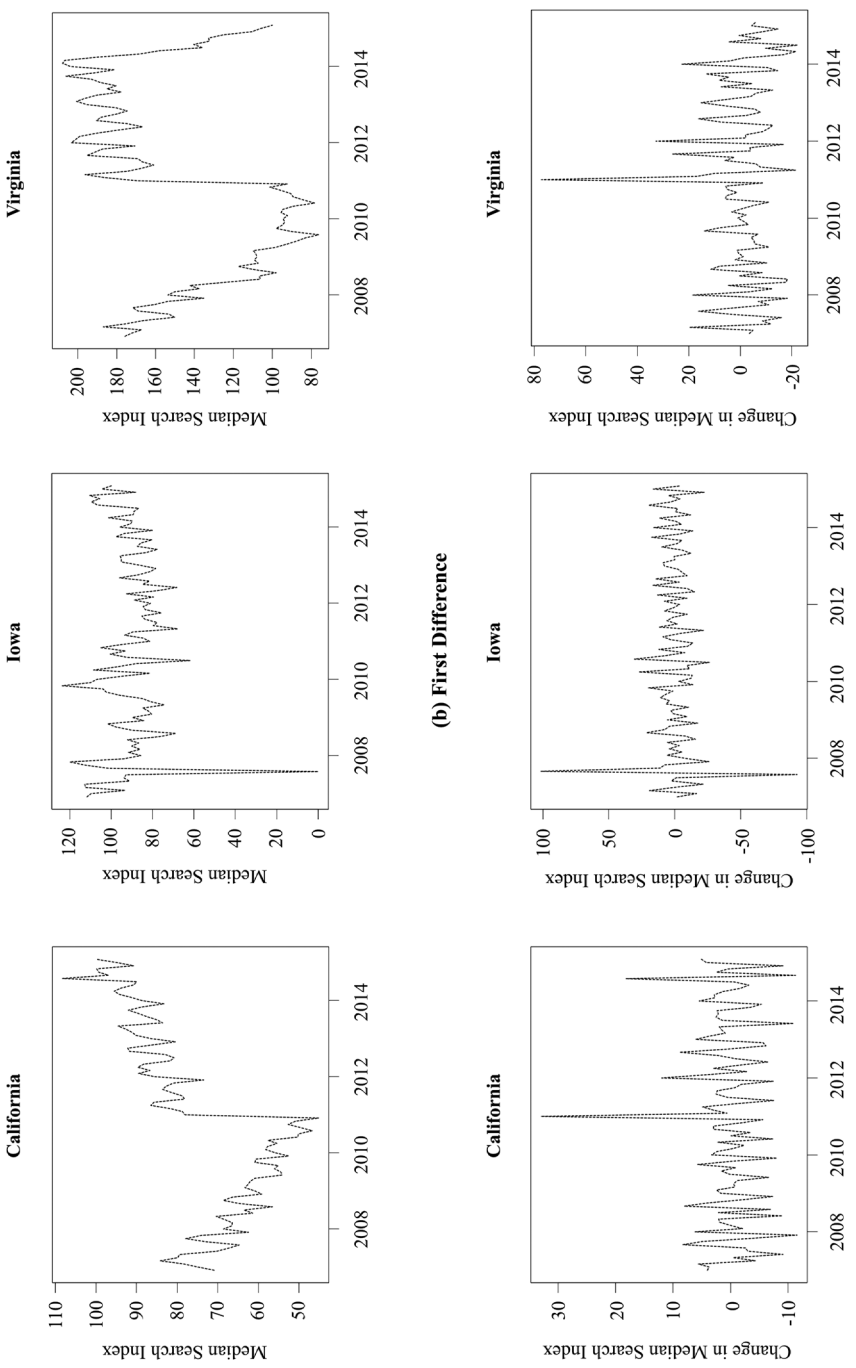


Figure 1 Re-based and Frequency-Adjusted Draws for Selected States

Note: For the ith state, the fine gray lines plot the re-based and frequency-adjusted data, d_{ijt} for $j = 1, 2, ..., 30$, as well as the median across draws, $\tilde{d}_{it}$, which is shown as a heavy dashed black line.

©2017 The University of Melbourne, Melbourne Institute of Applied Economic and Social Research

Figure 2 Structural Breaks and Outliers in the Median Search Frequency for Selected States



©2017 The University of Melbourne, Melbourne Institute of Applied Economic and Social Research

several also falling in 2006. These zeros are mixed in among large non-zero observations—the plots for Connecticut and Oregon in Figure 1 are representative examples of this phenomenon. In practice, it is unlikely that the frequency of depression and anxiety searches in a given state fluctuates from a negligible level in one month to a very high level in the next month, which suggests that the signal arising from the data may be unreliable over this time period. Consequently, we exclude data from the highly volatile period prior to December 2006.

Even after trimming the start date of the sample, the data for several of the smaller states is still somewhat patchy. A good example is Vermont, where the median search frequency is zero for almost the entire period prior to 2010 (see the bottom right panel of Figure 1). Given that variation in the data is required to achieve accurate estimation in regression models, we elect to exclude states where more than one observation over the period December 2006 to February 2015 takes a value of zero.[5] This leads to the exclusion of Arkansas, Hawaii, Idaho, Maine, Mississippi, Nebraska, Nevada, New Hampshire, New Mexico, Rhode Island, South Dakota and Vermont, leaving us with 39 states.

3.5 Structural Breaks and Outliers

Our final step is to address the presence of structural breaks and extreme observations that may otherwise exert an abnormal influence on the estimated parameters of any models that are subsequently fitted to the data. Figure 2 reports both the level and first difference of the median search frequency, $\tilde{d}_{it}$ and $\Delta\tilde{d}_{it}$, for a selection of states. Two different types of behaviour can be seen in the figure—a level shift occurs in the data for both California and Virginia, while an apparent outlier causes a one-off spike in the data for Iowa.

There are many methods to account for such phenomena. One option is to model them explicitly in the estimation stage by means of dummy variables, for example. Alternatively, in the presence of outliers, one could employ a procedure such as the least absolute deviations estimator, which attaches less weight to extreme observations than many other popular estimators, including ordinary least squares. Another solution is to clean the data of structural breaks and outliers prior to estimation. Given that our focus here is on data cleaning rather than estimation, we implement a simple approach to purge extreme observations from the sample.

Our first step is to research the properties of the data in search of operational factors that may contribute to the structural breaks and outliers that we observe. Stephens-Davidowitz (2013) notes that Google improved its geographic definitions for both California and Virginia as of January 2011, which coincides precisely with the apparent structural breaks for these states in Figure 2. Figure 1 reveals a similar phenomenon in the case of Oregon in July 2013. For each of these three cases, we set $\Delta\tilde{d}_{it} = 0$ in the month of the change, which is consistent with the assumption that the true unobserved data-generating process is a simple random walk.[6] Recall that, if desired, the level of an index can be easily reconstructed by accumulating the first difference and setting an appropriate base period.

Having accounted for these three structural breaks, a small number of extreme observations remain. We are therefore faced with the issue of whether or not an observation can be considered an outlier. This inevitably involves an element of judgement. By inspecting the data for each state, we identify two cases—Iowa and Oklahoma—where large outlying observations contribute to an enlarged range of the data and to very high values of the excess kurtosis of $\Delta\tilde{d}_{it}$.[7] The excess kurtosis of $\Delta\tilde{d}_{it}$ for Iowa and Oklahoma is 15.96 and 7.30, respectively, which contrasts with values in the range $[-0.74, 2.27]$ for the remaining states. We therefore set two outlying observations of $\Delta\tilde{d}_{it}$ for Iowa and one for Oklahoma to zero.

4. Summary Statistics for the Cleaned Data

Table 1 reports a selection of standard sum- mary statistics for the cleaned data in first differences. By virtue of our use of first differences, the mean and median values reported in the table are

© 2017 The University of Melbourne, Melbourne Institute of Applied Economic and Social Research

Table 1 Summary Statistics for the Cleaned Data in First Differences, 2007m1–2015m2

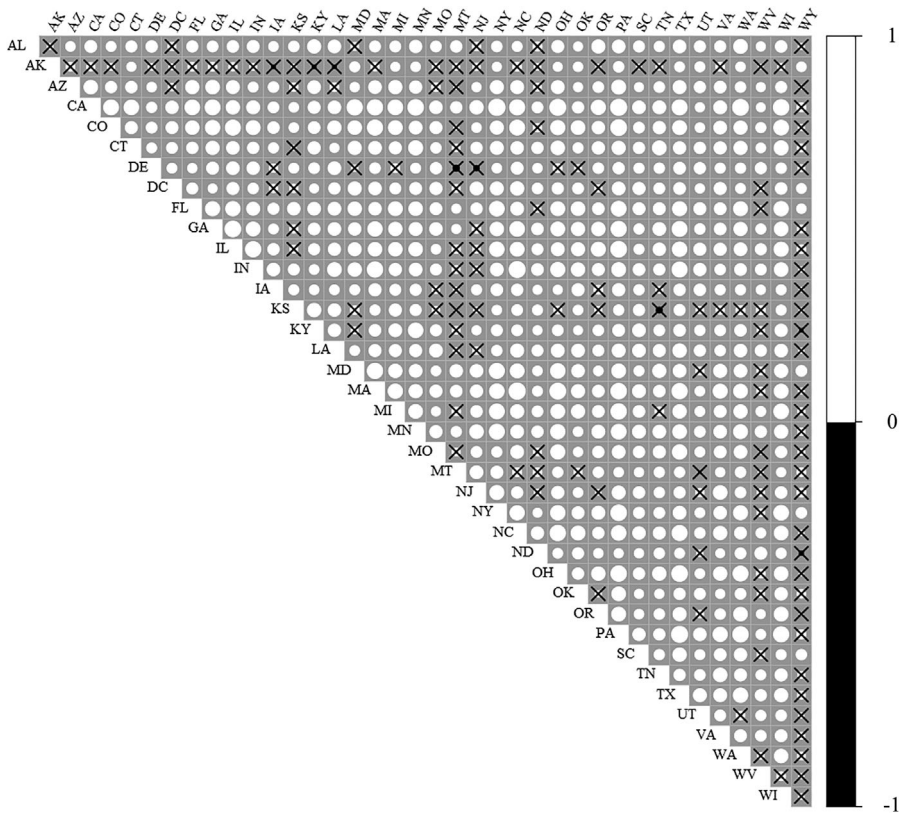
	Mean	*Median*	*Minimum*	*Maximum*	*Skew*	*Kurtosis*
Alabama	–0.03	0.44	–31.38	22.88	–0.36	0.09
Alaska	0.31	–2.33	–31.70	35.95	0.37	–0.31
Arizona	0.22	0.60	–18.82	20.73	0.12	0.07
California	–0.04	0.45	–11.58	18.24	0.10	1.19
Colorado	0.19	–0.35	–18.22	18.62	0.28	–0.18
Connecticut	0.31	0.41	–30.48	25.84	–0.12	0.40
Delaware	0.29	1.70	–34.68	31.07	–0.15	0.31
District of Columbia	0.15	–0.72	–20.81	22.75	0.40	0.05
Florida	0.27	–0.20	–18.72	19.92	0.12	0.06
Georgia	0.14	0.00	–19.29	16.20	–0.23	0.70
Illinois	0.09	0.70	–18.30	18.93	0.04	0.17
Indiana	0.24	1.25	–18.28	17.88	–0.28	–0.74
Iowa	–0.21	–0.19	–26.21	30.56	0.11	–0.03
Kansas	0.14	–0.09	–31.04	26.03	–0.13	0.56
Kentucky	0.23	0.67	–24.30	27.49	0.11	0.55
Louisiana	0.09	–0.12	–28.13	29.25	–0.17	0.31
Maryland	0.20	0.50	–17.38	22.21	0.23	–0.35
Massachusetts	0.10	0.39	–20.09	20.16	0.04	0.72
Michigan	0.20	0.97	–17.22	16.21	0.03	–0.59
Minnesota	0.34	0.58	–15.92	20.41	0.14	0.23
Missouri	0.28	0.30	–20.62	18.09	–0.09	0.41
Montana	0.26	1.38	–47.52	58.78	0.19	0.93
New Jersey	0.10	–0.18	–19.38	20.87	0.46	0.17
New York	0.05	0.19	–11.84	21.92	0.66	1.37
North Carolina	0.07	–0.14	–23.30	21.54	0.05	0.86
North Dakota	0.00	–1.06	–47.98	49.66	0.18	0.63
Ohio	0.25	–0.78	–22.97	19.72	0.15	0.03
Oklahoma	0.26	–0.78	–26.51	34.99	0.40	0.63
Oregon	0.12	–0.10	–57.84	47.07	–0.11	0.19
Pennsylvania	0.13	0.94	–16.93	18.49	0.09	0.26
South Carolina	–0.04	–0.17	–18.80	22.44	–0.08	–0.46
Tennessee	0.12	0.00	–20.87	27.19	0.08	0.47
Texas	0.16	0.44	–17.20	18.20	–0.22	0.61
Utah	0.16	0.74	–23.08	22.16	–0.20	–0.05
Virginia	–1.56	–2.72	–22.34	32.72	0.56	0.49
Washington	0.24	0.02	–18.32	27.78	0.46	1.07
West Virginia	0.14	1.07	–51.16	51.12	–0.13	0.68
Wisconsin	0.19	–0.47	–16.17	24.35	0.24	0.18
Wyoming	–0.08	0.88	–44.48	46.18	–0.44	2.27

Note: 'Kurtosis' denotes excess kurtosis.

relatively close to zero in all cases. The range and standard deviation display some variation across states but no state stands out in this regard. Similarly, there is little evidence of skewness or excess kurtosis, which suggests that the distribution of $\Delta\tilde{d}_{it}$ for each state is relatively symmetric and does not exhibit excess tail mass. Finally, Figure 3 reports the correlation matrix for $\Delta\tilde{d}_{it}$ in the form of a heatmap. The figure reveals widespread positive correlation in the cross-section, which indicates that the frequency of Google searches for depression and anxiety comoves across states. This is an intuitive finding, as we conjecture that many of the factors contributing to mental health disorders in the United States are likely to act nation-wide and are therefore common to all states. Ultimately, however, identifying the factors that explain the positive cross-section cor-relation in the data is an empirical task—the purpose of data cleaning is to ensure that such empirical analysis is based on a solid foundation.

©2017 The University of Melbourne, Melbourne Institute of Applied Economic and Social Research

Figure 3 Cross-section Correlation of the Cleaned Data in First Differences



Notes: Positive correlations are represented by white circles and negative correlations by black circles. The size of each circle is proportional to the strength of the correlation. Values marked X are insignificant at the 5 per cent level.

5. Concluding Remarks

In this article, we provide an illustration of data cleaning based on a dataset from Google Trends. Data cleaning is an important but under-appreciated aspect of empirical work in economics and, as with many aspects of empirical work, it requires an element of judgement. Naturally, where judgement is involved there is the potential for disagreement as judgement differs from person to person. When undertaking empirical work, it is therefore important to maintain transparency in relation to data handling by carefully documenting one's data cleaning procedures and by applying them systematically.

May 2017

Endnotes

1. Google Trends can be freely accessed via <https://www.google.com/trends/>.

2. For more information, see <https://www.comscore.com/Insights/Rankings/comScore-Releases-March-2015-US-Desktop-Search-Engine-Rankings>.

3. This is an important point of contrast relative to the interpolation procedure outlined above, where one uses an arbitrary smooth function to infer the within-period variation.

4. In principle, working with a larger number of draws will yield an improved understanding of Google's sampling routine. In practice, there is a substantial time-cost associated with collecting the data as each draw is obtained on a separate day. Our choice to work with 30 draws represents a balance between these two considerations.

5. Given that the time dimension of our sample is $T = 98$ months, this rule excludes states where more than ≈ 1 per cent of the data takes a zero value. As such, it is relatively

© 2017 The University of Melbourne, Melbourne Institute of Applied Economic and Social Research

conservative—a more lenient rule would allow for more states to be included in the sample at the cost of more noise.

6. If a variable, x_t, follows a simple random walk and if there is no disturbance in period t, then $x_t = x_{t-1}$ by definition which implies that $\Delta x_t = 0$. The assumption that data is generated by a random walk is relatively common in time series analysis but more sophisticated data-generating processes could be used. In addition, one could account for different types of structural break, including trend shifts, for example.

7. Kurtosis is a measure of the mass in the tails of a distribution—the higher the kurtosis, the greater the likelihood that one will observe extreme values. It is common to present the excess kurtosis, which expresses the kurtosis of a distribution relative to the kurtosis of the normal distribution.

References

Berger, M., Wagner, T. and Baker, L. 2005, 'Internet use and stigmatized illness', *Social Science and Medicine*, vol. 61, pp. 1,821–7.

Stephens-Davidowitz, S. 2013, 'Essays using Google Data', PhD thesis, Harvard University.

Tefft, N. 2011, 'Insights on unemployment, unemployment insurance, and mental health', *Journal of Health Economics*, vol. 30, pp. 258–64.

Wickham, H. 2014, 'Tidy data', *Journal of Statistical Software*, vol. 59, pp. 1–23.

© 2017 The University of Melbourne, Melbourne Institute of Applied Economic and Social Research