Problem Set #1

2022-09-23

Note: Please prepare your answers using Rmarkdown and submit a pdf via Canvas. Each submission has to include all code and R output used to answer the questions. I encourage you to work on the assignments together, but each of you have to type up their responses individually. Identical submissions will not be accepted. Late submissions may only receive partial credit at my discretion.

Penney (2016) explored whether the widespread publicity about NSA/PRISM surveillance (i.e., the Snowden revelations) in June 2013 was associated with a sharp and sudden decrease in traffic to Wikipedia articles on topics that raise privacy concerns. If so, this change in behavior would be consistent with a chilling effect resulting from mass surveillance. The approach of Penney (2016) is sometimes called an interrupted time series design, and it is related to the approaches described in section 2.4.3. of Salganik (2019).

To choose the topic keywords, Penney referred to the list used by the US Department of Homeland Security for tracking and monitoring social media. The DHS list categorizes certain search terms into a range of issues, i.e., "Health Concern," "Infrastructure Security," and "Terrorism." For the study group, Penney used the 48 keywords related to "Terrorism" (see appendix table 8). He then aggregated Wikipedia article view counts on a monthly basis for the corresponding 48 Wikipedia articles over a 32-month period from the beginning of January 2012 to the end of August 2014. To strengthen his argument, he also created several comparison groups by tracking article views on other topics. Now, you are going to replicate Penney (2016). All the raw data that you will need for this activity is available from Wikipedia (https://dumps.wikimedia.org/other/pagecounts-raw/ (Links to an external site.)). Or you can get it from the R-package wikipediatrend. When you write up your responses, please note which data source you used. This activity will give you practice in data wrangling and thinking about discovering natural experiments in big data sources. It will also get you up and running with a potentially interesting data source for future projects.

a) Read Penney (2016) and replicate his figure 2, which shows the page views for "Terrorism"-related pages before and after the Snowden revelations. Interpret the findings.

```r
library(foreign)
library(MASS)
library(ggplot2)
library(tidyverse)
library(wikipediatrend)

dat <- wp_trend("Al-Qaeda", lang = "en", from = "2012-01-01", to = "2014-12-12")
```

b) Next, replicate figure 4A, which compares the study group ("Terrorism"- related articles) with a comparator group using keywords categorized under "DHS & Other Agencies" from the DHS list (see appendix table 10 and footnote 139). Interpret the findings.

c) In part (b) you compared the study group with one comparator group. Penney also compared with two other comparator groups: "Infrastructure Security"–related articles (appendix table 11) and popular Wikipedia pages (appendix table 12). Come up with an alternative comparator group, and test whether the findings from part (b) are sensitive to your choice of comparator group. Which choice makes most sense? Why?