Procedures for Assembling the Third Data Set

Congress, the Press, and Political Accountability

R. Douglas Arnold

The third data set includes information about the volume and timing of coverage for a large sample of newspapers and representatives. This data set shows how 67 local newspapers covered 187 representatives during 1993 and 1994, with a total of 242 representative/newspaper dyads. The 61,084 citations — headline, date, section, page, and byline, but not full-text — allow one to analyze how the amount and timing of coverage depend on the newsworthiness of individual representatives, the competitiveness of elections, and the resources and constraints that face individual newspapers.

Sample

The third data set is not a random sample of all newspapers. Indeed, it is closer to being the universe of all newspapers that were available for electronic searches in 1993 and 1994. The data set was assembled incrementally. First, I collected information about how the 25 newspapers in the first data set covered all 91 representatives in their primary circulation areas (22,175 citations). Then I collected information about how the 6 comparison newspapers in the second data set covered the 33 representatives in their primary circulation areas (5,718 citations). Then I gradually added information from other newspapers. Since the greatest puzzle was determining how the number of representatives in a newspaper's circulation area affected its coverage of individual representatives, I first selected newspapers according to the number of representatives in their primary circulation areas (starting with papers with the most representatives) and continued downloading articles until resources ran out. By that time, I had added 36 newspapers, 96 representatives, and 33,191 citations to the data set. Although the third data set is not a random sample of newspapers, it contains the 25 randomly-selected newspapers from the first data set. By analyzing separately how these 25 newspapers covered the 91 legislators within their primary circulation areas, one can determine if the larger but less representative sample differs significantly from the smaller but more representative sample.

Procedures

The procedures used for collecting and cleaning the third data set differed in three ways from the procedures used to create the first and second data sets. First, the search routine set was less exhaustive. For the third data set, I searched for any mention of a representative's first name and last name (or nick name and last name), but I avoided searching for last-name-only references to the representative. Since I was downloading the citations but not the articles, there was no way to separate correct and incorrect last-name references. The citation counts for eight representatives in the first data set are different for the same representatives in the third data set because of the different search routines. Second, I adopted a different mechanism for eliminating duplicate articles. For the first two data sets, human coders read all of the articles and then discarded any duplicate articles that appeared in the various regional or timed editions of the same newspaper. For the third data set, human coders searched the citation lists for similarly-titled articles published on the same day or adjacent days and discarded any duplicate citations. Third, I used a different procedure for eliminating references to people who shared a representative's first and last names. For the first data set, human coders read all articles and discarded any articles that did not mention the local representative. For the third data set, I read a sample of all articles (on-line), searching for wrong citations, and then estimated the percentage of all citations that were in error. The range was from zero to 52 percent. The citation lists were then adjusted according to the error rate for individual representatives.

Original Data

The third data set contains 23 monthly citation counts for each representative/newspaper dyad from January 1993 to November 1994 (the last is a partial month). Raw monthly counts can be found in variables C9301 to C9411. Corrected monthly counts — the raw counts times variable WRONG — can be found in variables C9301C to C9411C. The corrected counts are rounded to two decimal places. The third data set also contains 23 monthly counts for a representative's name appearing in a newspaper headline (variables H9301 to H9411). These counts appear to be highly accurate and have not been corrected. There are three variables that sum the corrected monthly counts for 1993, 1994, and both years (CITE93, CITE94, CITE), as well as three variables that sum the monthly headline counts (HEAD93, HEAD94, HEADLINE).

Although the procedures used to collect and assemble the original data in the third data set were not quite as painstaking as the procedures used to create the first and second data sets, I believe the data are of high quality.

Data from Secondary Sources

Most of the other variables in the third data set were created from various published sources.