On the Replicability of Data Collection Using Online News Databases
Appendix

Empirical Tests of Reliability in News Databases

Comparing News Databases

Conflict event data projects typically gather a corpus of articles from an online news database. We conduct a systematic review of two popular news databases—Factiva and Nexis-Uni—to document variation in source availability and compare database reliability.

Our comparison serves two purposes. First, it demonstrates the degree to which researchers can expect search results to vary over time. Second, it serves as a reliability comparison between Factiva and Nexis Uni. This comparison is particularly useful for informing decisions about coding procedure; the choice of database and timeframe over which sources are accessed; and the potential these choices have to affect the resulting corpus of articles. This comparison is also useful for users of event data who wish to know the limitations of the data generating process so that they can be appropriately cautious in their conclusions.

Procedure

From February through October of 2019, we accessed source articles on Factiva and Nexis-Uni using identical search strings drawn from the MID Project (Palmer et al. 2015).[1] We use these search strings as our test because the MID project is cited frequently in international relations research, meaning that the process of collecting the data is consequential, and because the MID project uses a variety of sources, making it a useful test of variation across many

[1] Searches were conducted roughly weekly, subject to coder availability. Note that the Nexis series begins later than the Factiva series because we adjusted our Nexis search procedure after the first weeks of searches.

sources.[2] We accessed source articles using the MID strings and recorded the number of resulting articles for each search. **Table 2** shows the specific news sources used in each search string. **Figures 1-4** show the number of results for each search.

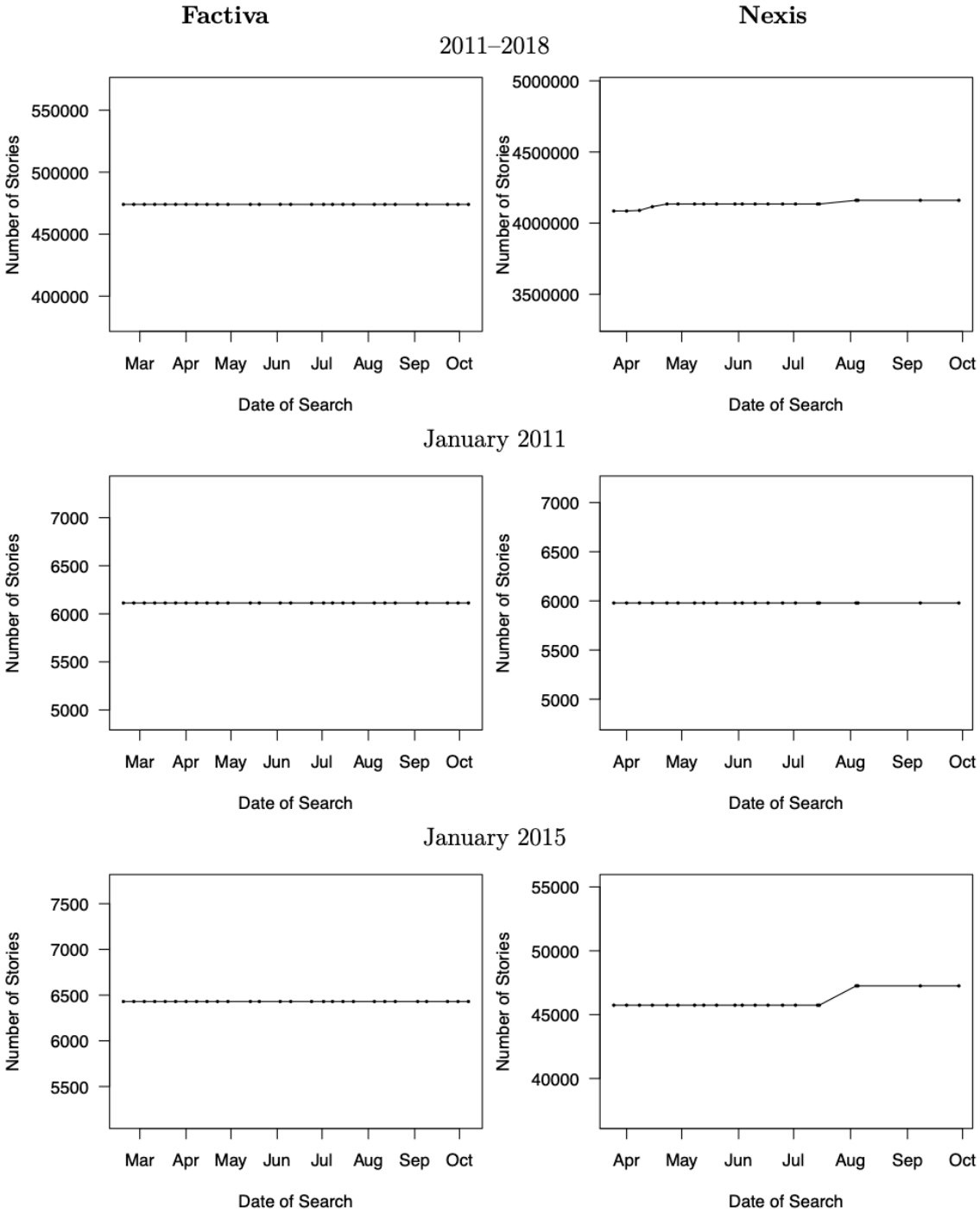
Table 2: MID 5 Source Strings

Search String	Sources
BBC String	BBC – All Sources
AP String	The Associated Press – All Sources
Multi-Source String 1	The New York Times – All Sources
	CNN – All Sources
	ITAR TASS
	Agence France Press – All Sources
Multi-Source String 2	Xinhua – All Sources
	The Times (U.K.)
	dpa International Service – All Sources
	Montreal Gazette
	Interfax – All Sources
	The Jerusalem Post – All Sources

NOTE: Source names are shown exactly as they appear in the Factiva source database as of March 26, 2019, at 4:35pm.

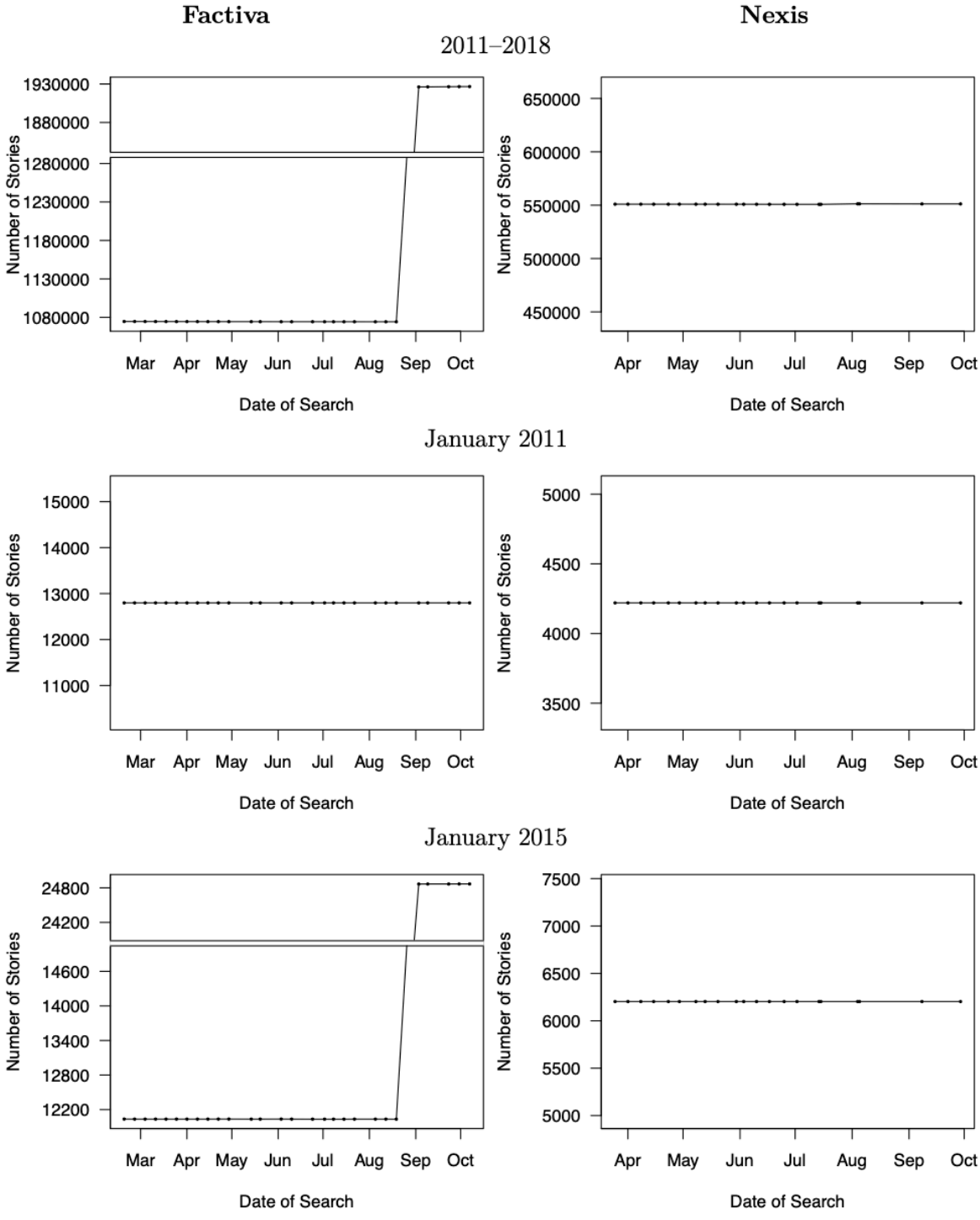
[2] We also replicated our tests with the SCAD search string (Salehyan et al. 2012) as a robustness test to ensure that the MIDs string was not uniquely prone to variation. Results from the SCAD searches are similar.

Figure 1: BBC Search String Results



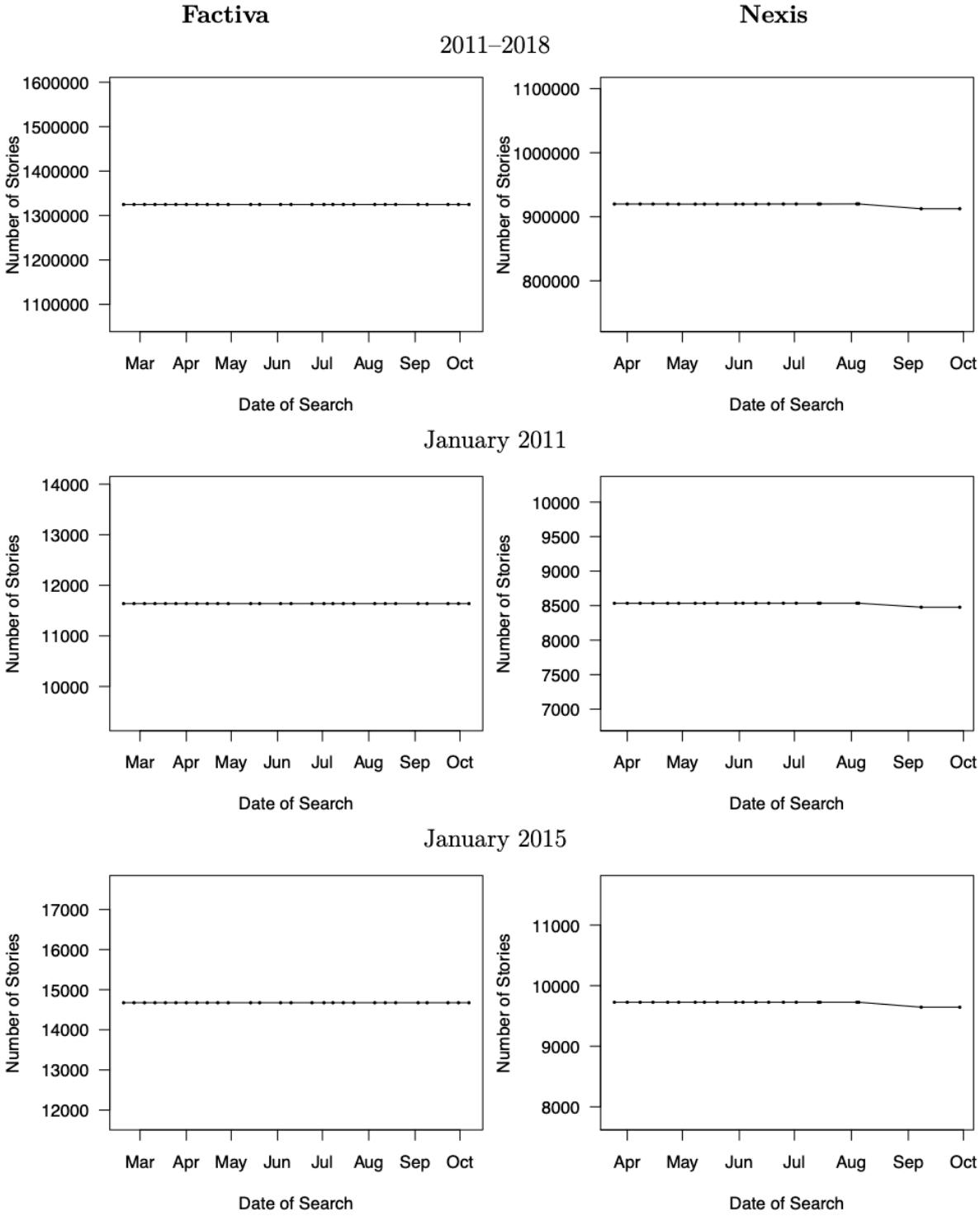
Each plot demonstrates the number of search results returned by the MID string for this source over time. Y-axes are scaled to preserve variance: they range from 20% below to 20% above the mean number of stories for the search over the time period. Results that differ by more than 20% from the mean are shown with a broken y-axis.

Figure 2: Associated Press Search String Results



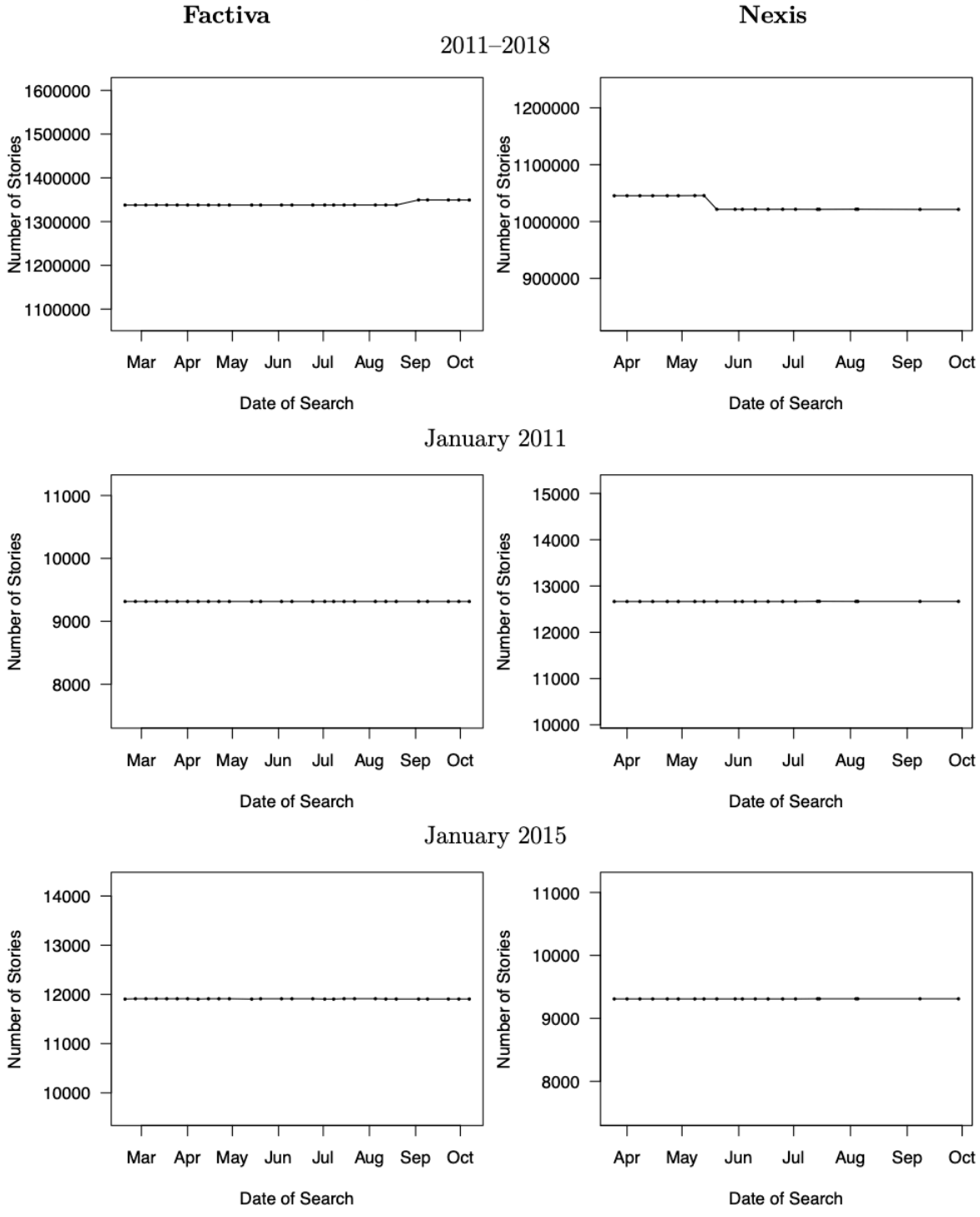
Each plot demonstrates the number of search results returned by the MID string for this source over time. Y-axes are scaled to preserve variance: they range from 20% below to 20% above the mean number of stories for the search over the time period. Results that differ by more than 20% from the mean are shown with a broken y-axis.

Figure 3: Multi-Source Search String 1 Results (NYT, CNN, AFP, ITAR TASS)



Each plot demonstrates the number of search results returned by the MID string for this source over time. Y-axes are scaled to preserve variance: they range from 20% below to 20% above the mean number of stories for the search over the time period. Results that differ by more than 20% from the mean are shown with a broken y-axis.

Figure 4: Multi-Source Search String 2 Results (Xinhua, The Times (U.K.), DPA, Montreal Gazette, Jerusalem Post, Interfax)



Each plot demonstrates the number of search results returned by the MID string for this source over time. Y-axes are scaled to preserve variance: they range from 20% below to 20% above the mean number of stories for the search over the time period. Results that differ by more than 20% from the mean are shown with a broken y-axis.

Observations

1. Large Changes in Factiva

We find that search strings in both sources display some amount of variability, even when using identical search strings, over time. Factiva generally displays fewer changes in the number of results for a search, but also experienced a massive increase in the number of Associated Press stories in August/September. Around this time, Factiva began including AP Photostream – a news-in-photos service provided by the Associated Press – in its results when searching for AP stories. This change increased the number of Associated Press search results for 2015 onward, but not for earlier years. We were unable to find explanations for the smaller variations in number of stories in either service. We verified that the availability of entire sources did not change from week to week.

Our conversations with newspaper database representatives suggest to us that large changes of the kind observed in our AP results have two potential sources: licensing agreements and changes to the database backend. During our time period, Factiva either decided that AP Photostream should be included in the "Associated Press – All Sources" category or secured a licensing agreement to include the previously-omitted AP Photostream results in its database. We were unable to confirm the cause of this specific change, but personal communication with a news database developer suggested that most sourcing decisions are driven by licensing arrangements.

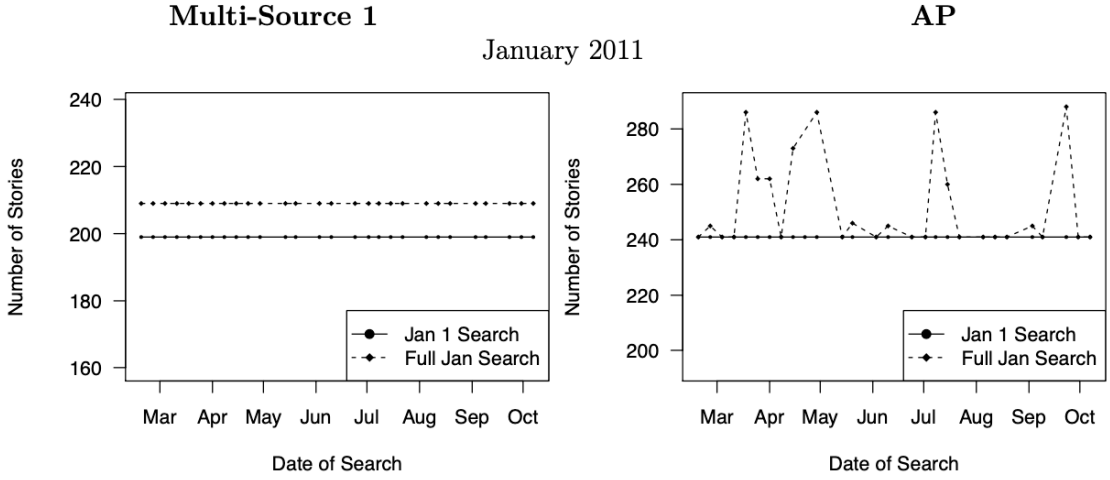
2. Date Range Irregularities

Over the course of our testing, Factiva consistently exhibited irregularities in search results by date range. When sorting by date, Factiva results were sometimes out of order; stories labeled "January 2" would appear among stories labeled "January 1." Searches that began in January

would sometimes contain stories labeled December 31 of the previous year. Finally, the number of Factiva stories from a particular date would change depending on whether the search was conducted in isolation or whether it was narrowed from a larger set of results. For example, entering a particular search for January 1 would sometimes return a different number of results than searching for all of January and narrowing to include only stories from the first day of the month.

Figure 5 demonstrates an example of this phenomenon. When searching for stories from January 1, 2011 in Factiva, we get a different number of results depending on whether we (1) specify that we want results only within January 1 or (2) search for all of January, sort by date, and then count only the stories before the first story from January 2. For the first multi-source string, we consistently have a few extra stories when searching the wider date range and then sorting. When searching for Associated Press stories, the results for searching January 1 by itself are stable, but the results for searching all of January and then narrowing the date range are inconsistent.

Figure 5: Factiva Results for January 1, 2011



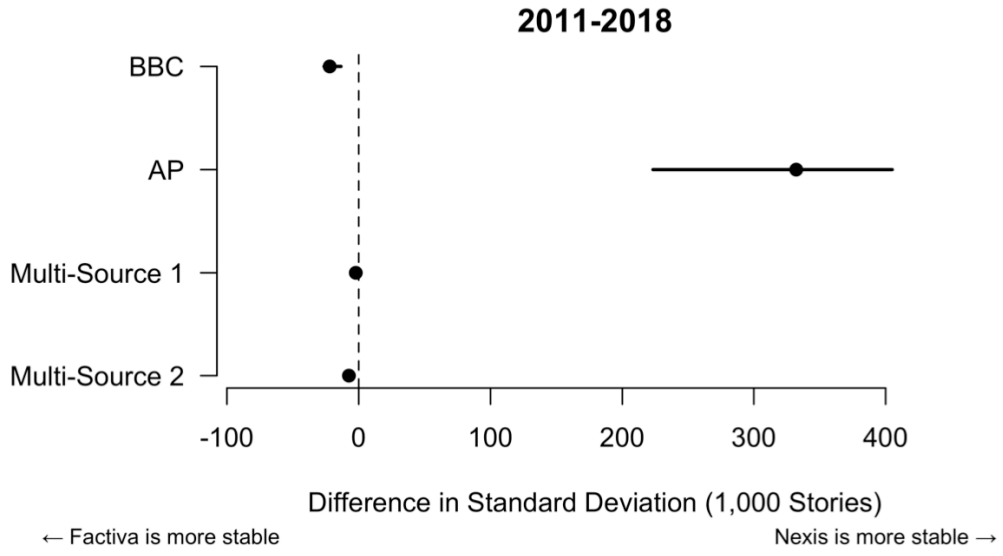
Each plot demonstrates the number of Factiva search results returned by the MID string for January 1, 2011, when searching only within January 1 or searching all of January and narrowing results. Y-axes are scaled to preserve variance: they range from 20% below to 20% above the mean number of stories for the search over the time period.

One possible explanation for this observation is that Factiva uses different dates for searching and sorting purposes. On the backend, Factiva is finding stories from the specified date range and including them in the search results. On the front end, Factiva is sorting the articles in ways inconsistent with the backend date labels, resulting in inconsistencies and apparent mis-sorting of stories. We are unable to confirm the causes of these discrepancies without looking into the "black box" of the Factiva backend, and it is thus an example of a stochastic source of error.

3. General Stability Comparison

We augment our comparison of the raw search results by comparing the amount of variation in each search over the time period of our study. We compare the standard deviation (in thousands of stories) for each search string between Factiva and Nexis-Uni to assess which database is more stable. The results of this test are shown in **Figure 6.**

Figure 6: Difference in Variation by Search String



Points represent the difference in standard deviation (in thousands of stories) between Factiva and Nexis for the given search string over the time period of our study. Lines indicate bootstrapped 95% confidence intervals.

Our tests indicate that Factiva displays less variation in search results for the majority of our search strings. Factiva, however, displays an extreme amount of variation for the Associated Press search string due to the nearly 50% increase in the number of search results in August/September.

Neither Factiva nor Nexis-Uni is free from random variation that can affect the corpus of articles gathered by researchers. Nexis-Uni displays significant random variation, while Factiva appears to have changed the function of its Associated Press search backend during the time period of our study. The difference between the maximum and minimum number of stories produced for each search string in Nexis-Uni is fairly substantial. For instance, while the BBC search string produced upwards of 4.1 million stories for the entire date range, the difference between the most and least number of stories produced was 74,873. The first multi-source search string produced as many as approximately 92,000 stories and the biggest gap in stories was

7,606. For the second multi-source search string, as many as about 1,045,000 stories were generated from the search and the variation was as wide as 24,357 stories. Finally, the AP search string produced approximately 416,000 stories at its height, with the biggest single difference between two searches being 453 stories.

References

Palmer, Glenn, Vito d'Orazio, Michael Kenwick, and Matthew Lane. 2015. "The MID4 dataset, 2002–2010: Procedures, coding rules and description." *Conflict Management and Peace Science*, *32* (2): 222-242.

Salehyan, Idean, Cullen S. Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams. 2012. "Social conflict in Africa: A new database." *International Interactions*, *38* (4): 503-511.