Missing Bills? Missing Versions?

Andreu Casas

August 26, 2016

Goal of this report:

- To figure out if we have versions for all bills: do we have the entire population?
- To find out if we are missing some versions for some of the bills

---

When exploring the RH/RS versions and how much they amend the Introduced versions (`reported_versions` report), I realized that for some bills we were missing the Introduced versions (IH/IS): e.g. **111-HR-1275**.

A) Are we missing any bill?

Loading packages and metadata dataset:

```
library(dplyr)
library(ggplot2)
library(xtable)
load("../data/new_metadata_93_114.Rdata")
```

Checking the dimensions of the dataset:

```
dim(meta)
```

```
## [1] 139619    110
```

```
length(unique(meta$BillID))
```

```
## [1] 97457
```

The dataset has:

- 139,619 rows (bill versions)
- 97,457 unique bills
- 103-113 Congresses

```
bills_by_cong <- meta %>%
  group_by(Cong) %>%
  summarize(bills = length(unique(BillID)))
```

Table 1: A table showing the unique bills by Congress for which we have at least 1 version

Cong	bills
103	7878
104	5241
105	7009
106	8966
107	8944
108	8448
109	10554
110	11075
111	10439
112	10268
113	8635

Let's compare this data with data from the Congressional Bills Project.

Loading the last CBP dataset.

```
cbp <- read.table("../data/bills93-114.txt", sep = ",", header = TRUE)
```

Selecting congresses from 103 to 113 from the CBP dataset and then adding a column to the previous table showing the number of unique bills by congress in the CBP dataset.

```
cbp_bills_by_cong <- cbp %>%
  filter(Cong %in% c(103:113)) %>%
  group_by(Cong) %>%
  summarize(bills = length(unique(BillID)))
bills_by_cong <- rename(bills_by_cong, bills_meta = bills)
cbp_bills_by_cong <- rename(cbp_bills_by_cong, bills_cbp = bills)
bills_by_cong$Cong <- as.numeric(bills_by_cong$Cong)
cbp_bills_by_cong$Cong <- as.numeric(cbp_bills_by_cong$Cong)
all_bills <- left_join(bills_by_cong, cbp_bills_by_cong)
all_bills$meta_in_cbp_perc <- round(
  ((all_bills$bills_meta / all_bills$bills_cbp) * 100), 2)
```

Table 2: Comparing bills by Congress in the Versions and the CBP dataset

Cong	bills_meta	bills_cbp	meta_in_cbp_perc
103	7878	9814	80.27%
104	5241	7978	65.69%
105	7009	9141	76.68%
106	8966	10812	82.93%
107	8944	10788	82.91%
108	8448	10667	79.20%
109	10554	13071	80.74%
110	11075	14029	78.94%
111	10439	13674	76.34%
112	10268	12299	83.49%
113	8635	10604	81.43%

So it looks that one average we are missing around 20 to 25% of the bills. Why??

B) Are we missing any version?

This is a more complicated question to answer because we don't have a very clear reference of how many actual bill versions there are for each Congress. However, let's start with an easier question first:

Do we have IH/IS versions for all the unique bills in the Versions dataset?

```
bills_intr_vers <- meta %>%
  group_by(Cong, BillID) %>%
  filter(version_type %in% c("IH", "IS")) %>%
  summarize(introduced_n = n()) %>%
  group_by(Cong) %>%
  summarize(introduced_n = sum(introduced_n))
bills_intr_vers$Cong <- as.numeric(bills_intr_vers$Cong)
bills_intr_by_congress <- left_join(bills_by_cong, bills_intr_vers)
bills_intr_by_congress$missing_ih_is_perc <- round(((1 -
  (bills_intr_by_congress$introduced_n / bills_intr_by_congress$bills_meta)) * 100), 2)
```

Table 3: The percentage of unique bills by Congress for which we are missing the Introduced version (IH/IS)

Cong	bills_meta	introduced_n	missing_ih_is
103	7878	7746	1.68%
104	5241	4898	6.54%
105	7009	6769	3.42%
106	8966	8767	2.22%
107	8944	8800	1.61%
108	8448	8253	2.31%
109	10554	10360	1.84%
110	11075	10823	2.28%
111	10439	10313	1.21%
112	10268	10111	1.53%
113	8635	7667	11.21%

On average we are missing betwee 1 and 5% of the Introduced versions. Why?