Missing Bills? Missing Versions?

Andreu Casas

August 26, 2016

Goal of this report:

- To find out if we are missing some bills.
- To find out if we are missing some bill versions.

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**. Then we decided to actually find out how many bills and bill versions we are missing.

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), BillType %in% c("hr", "s")) %>%
  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	7872	100.08%
104	5241	6535	80.20%
105	7009	7529	93.09%
106	8966	8943	100.26%
107	8944	8945	99.99%
108	8448	8466	99.79%
109	10554	10558	99.96%
110	11075	11073	100.02%
111	10439	10621	98.29%
112	10268	10439	98.36%
113	8635	8873	97.32%

On average we are missing around 1 to 2% of the bills.

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 %>%
  dplyr::select(Cong, BillID, version_type) %>%
  unique() %>% # SOME DUPLICATED VERSIONS IN THE MEATADATA DATASET!
  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

We finally decided to go to the GPO's website and scrape all bill versions they have available for the 103rd to the 113th Congress. The GPO have a sitemap for all the bill versions they have in their website for each year. So I wrote a `python` program (`scrapping_gpo_list_bill_versions.py`) that scrapes the sitemap for each year between 1993 and 2014 (both included), and parses the `url` of each version in order to get the versions': `Cong`, `BillType`, `BillNum`, `version_type`, and `url`. See for example the sitemap for 1993.

I saved the resulting data in a `csv` file that I included it in the `data` directory: `all_gpo_versions.csv`.

Loading that datset and creating a table with number of versions type by Congress

```
gpo_versions <- read.csv("../data/all_gpo_versions.csv")
gpo_vtype_by_cong <- gpo_versions %>%
  filter(BillType %in% c("HR", "S")) %>%
  group_by(Cong, version_type) %>%
  summarize(n = n()) %>%
  arrange(Cong, desc(n))
```

Creating the same table (number of version-type by Congress) for our Versions dataset

```
vtype_by_cong <- meta %>%
  dplyr::select(Cong, BillID, version_type) %>%
  unique() %>%
  group_by(Cong, version_type) %>%
  summarize(n = n()) %>%
  arrange(Cong, desc(n))
```

Merging the 2 datasets

```
vtype_by_cong <- rename(vtype_by_cong, OURS_n = n)
gpo_vtype_by_cong <- rename(gpo_vtype_by_cong, GPO_n = n)
vtype_by_cong$Cong <- as.character(vtype_by_cong$Cong)
gpo_vtype_by_cong$Cong <- as.character(gpo_vtype_by_cong$Cong)
vtype_by_cong$version_type <- as.character(vtype_by_cong$version_type)
gpo_vtype_by_cong$version_type <- as.character(gpo_vtype_by_cong$version_type)
both_vtype_by_cong <- left_join(vtype_by_cong, gpo_vtype_by_cong)
both_vtype_by_cong$missing_perc <- round((1 -
  (both_vtype_by_cong$OURS_n / both_vtype_by_cong$GPO_n)) * 100, 2)
both_vtype_by_cong <- arrange(both_vtype_by_cong, Cong, desc(OURS_n))
both_vtype_by_cong <- both_vtype_by_cong %>% mutate(Dif = GPO_n - OURS_n)
# write.csv(both_vtype_by_cong, file = "./data/gpo_our_dataset_comparison.csv",
#           row.names = FALSE)
```

The resulting table is too large to put in here. Go to this link to take a look at it.

Trying to visualize this huge table

```
pdf("./images/gpo_ours_comparison.pdf", width = 12, height = 7)
ggplot(both_vtype_by_cong2,
       aes(x = factor(version_type), y = missing_perc)) +
  geom_bar(stat = "identity") +
  geom_text(aes(x = factor(version_type),
                y = missing_perc, label = GPO_n), vjust=-2) +
  geom_text(aes(x = factor(version_type),
                y = missing_perc,
                label = paste0("(", Dif, ")")), col = "blue4", vjust=-.75) +
  ylim(0, 120) +
  facet_wrap(~ Cong, scales = "free") +
  xlab("Version Type") +
  ylab("% versions we are missing") +
  theme(axis.text.y= element_text(size = 10),
        axis.text.x= element_text(size = 10),
        strip.text.x = element_text(size = 14),
        panel.background = element_rect("white"),
        panel.border = element_rect("black", fill = NA),
        strip.background = element_rect("white"))
dev.off()
```

So it seems we are not missing a lot of bills. The only exceptions are the 104th and 105th Congress.

I will now write a program to download all the versions that we are missing from GPO

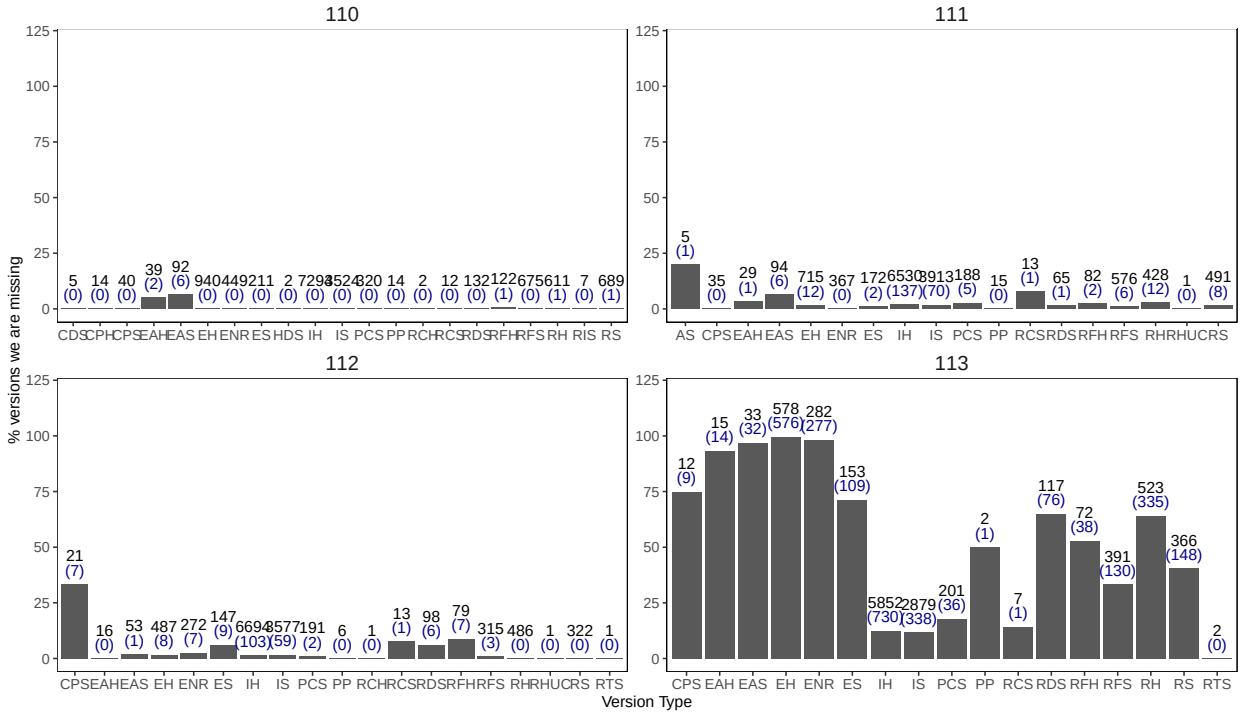

Figure 1: Black = Versions in the GPO's website | Blue = Versions we don't have