GDA of the share of women researchers (using **eurostat**)

Background The **eurostat** package offers tools to access data from Eurostat databases. As International Women's Day falls on March 8th each year, data about the share of women researchers were chosen.

Aims How does the share of women researchers vary over time and by sector and country?

Source Eurostat (http://ec.europa.eu/eurostat), downloaded 2016-03-29 12:20:07

Structure 2256 observations on 4 variables (2 factors, 1 date variable, 1 continuous variable)

The data are first downloaded from Eurostat. Variables have been added to make the country names and the date variable more readable.

```
library(eurostat)
datw <- get_eurostat("tsc00005")
library(countrycode)
library(dplyr)
datw$c1 <- countrycode(datw$geo, "iso2c", "country.name", warn=TRUE)
datw[datw$geo=="EA19", "c1"] <- "Eurozone"
datw[datw$geo=="EU28", "c1"] <- "EU"
datw[datw$geo=="EL", "c1"] <- "Greece"
datw[datw$geo=="UK", "c1"] <- "United Kingdom"
library(tidyr)
datw <- datw %>% separate(time, c("Year", "month", "day"), remove=FALSE)
```

You get an overall view of the share of women researchers by plotting the series for TOTAL for each country and for the EU and the Eurozone aggregates.

```
ggplot(na.omit(datw %>% filter(sectperf=="TOTAL", !(c1 %in% c("EU", "Eurozone")))),
  aes(Year, values)) + ylim(0, 60) + geom_line(aes(group=c1, colour=c1)) +
  theme(legend.position="none") + labs(x=NULL, y=NULL)
ggplot(na.omit(datw %>% filter(sectperf=="TOTAL", c1 %in% c("EU", "Eurozone"))),
  aes(Year, values)) + ylim(0, 60) + geom_line(aes(group=c1, colour=c1)) +
  theme(legend.position="none") + labs(x=NULL, y=NULL)
```

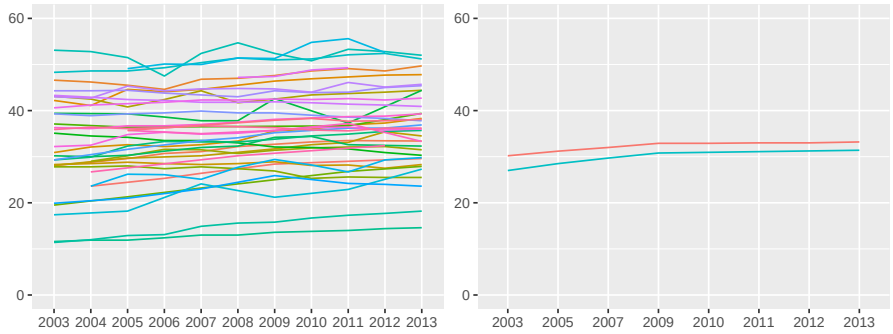

Fig 1: (Left) Shares of women researchers over the years 2003 to 2013 for the 28 EU members and 10 other countries. Two countries have much lower shares than the others, although they are rising, (these turn out to be Japan and Korea). (Right) Plots of shares of TOTAL for the EU as a whole (red) and the Eurozone. Some of the bigger countries must be dragging the aggregate averages down.

You get a more detailed view by plotting the five series for each country in a facetted plot. The sectors are Business (BES), Government (GOV), Higher Education (HES), Private non-profit (PNP), and all together (TOTAL). Missing values are now included to highlight the gaps in the data.

```
ggplot(datw, aes(Year, values)) + geom_line(aes(group=sectperf, colour=sectperf)) +
  geom_point(aes(group=sectperf, colour=sectperf), size=0.5) + facet_wrap(~c1) +
  theme(axis.text.x=element_blank(), legend.position="bottom",
  legend.title=element_blank()) + labs(x=NULL, y=NULL)
```

Fig 2: Shares of women researchers in different sectors over the years 2003 to 2014 for the 28 EU members and 10 other countries. The first thing to notice is the large number of missing values. There are hardly any data points for Montenegro and some countries have data for only some years and not for all sectors (e.g., Germany, Greece, Switzerland).

The populations of the countries in the EU differ a great deal and it is helpful to bear their relative sizes in mind. (The search_eurostat function in **eurostat** was used to identify a data source.)

```
popw <- get_eurostat("tps00005")
pop1 <- popw %>% filter(time=="2013-01-01", geo!="EU28")
ggplot(pop1, aes(reorder(geo, -values), values)) + geom_bar(stat="identity") +
  labs(x=NULL, y="Share of EU population (%)")
```

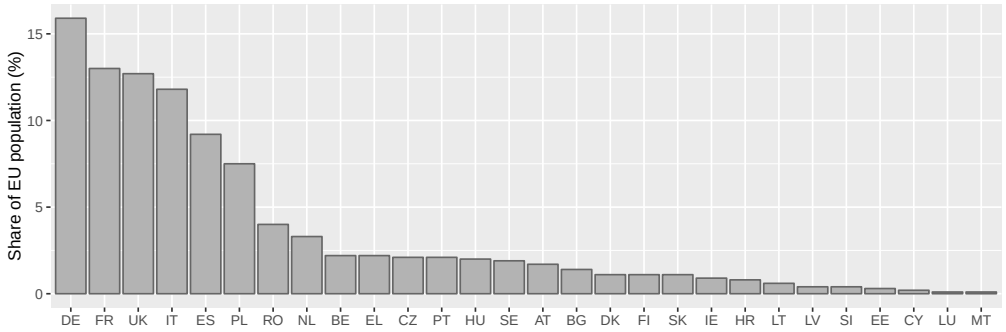

Fig 3: Percentage share of total EU population for the 28 member States in 2013. The six biggest countries make up well over half the population and are each a lot bigger than any of the other 22.

To get an overview of missing values by sector, restructure the dataset and display the patterns of missings.

```
library(extracat)
d4 <- spread(datw, sectperf, values)
visna(d4)
```

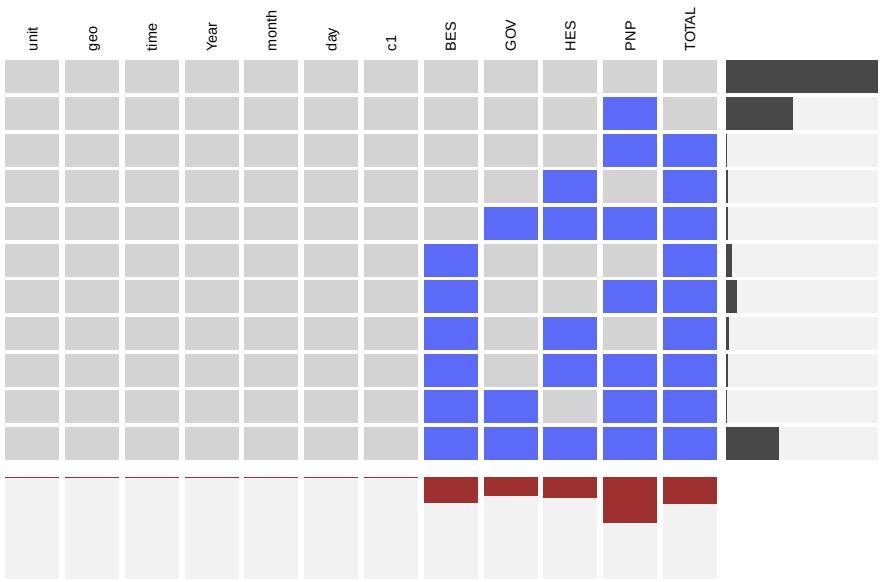

Fig 4: Missing value patterns by sector. There are three frequent patterns: no missings, only PNP missing, and all sectors missing.

It is curious that TOTAL is not missing when only PNP is missing. This suggests that some countries do not report the data for PNP and their TOTAL is based only on the other three sectors or that PNP data are included in another sector. (The spread function from **tidyr** used above fills in implicit missing values with NAs. This means that where there are no entries for a country in the original long dataset an NA value is added in the wide version.)

A missing patterns plot by Year similar to Figure 4 shows mainly that a large number of values are missing for the last year, 2014, so that year will be excluded from further analyses.

To check which countries have no PNP values and to see how many values each country has, the original dataset is filled out with missings. The numbers of PNP missings are then counted and plotted in Figure 5.

```
daty <-complete(datw, c1, Year, sectperf)
dy <- daty %>% filter(sectperf=="PNP") %>% group_by(c1) %>%
  summarise(ms=sum(!is.na(values)))
ggplot(dy, aes(x=c1, weight=ms)) + geom_bar() + labs(x=NULL, y=NULL) +
  ylim(0,15) + coord_flip()
```

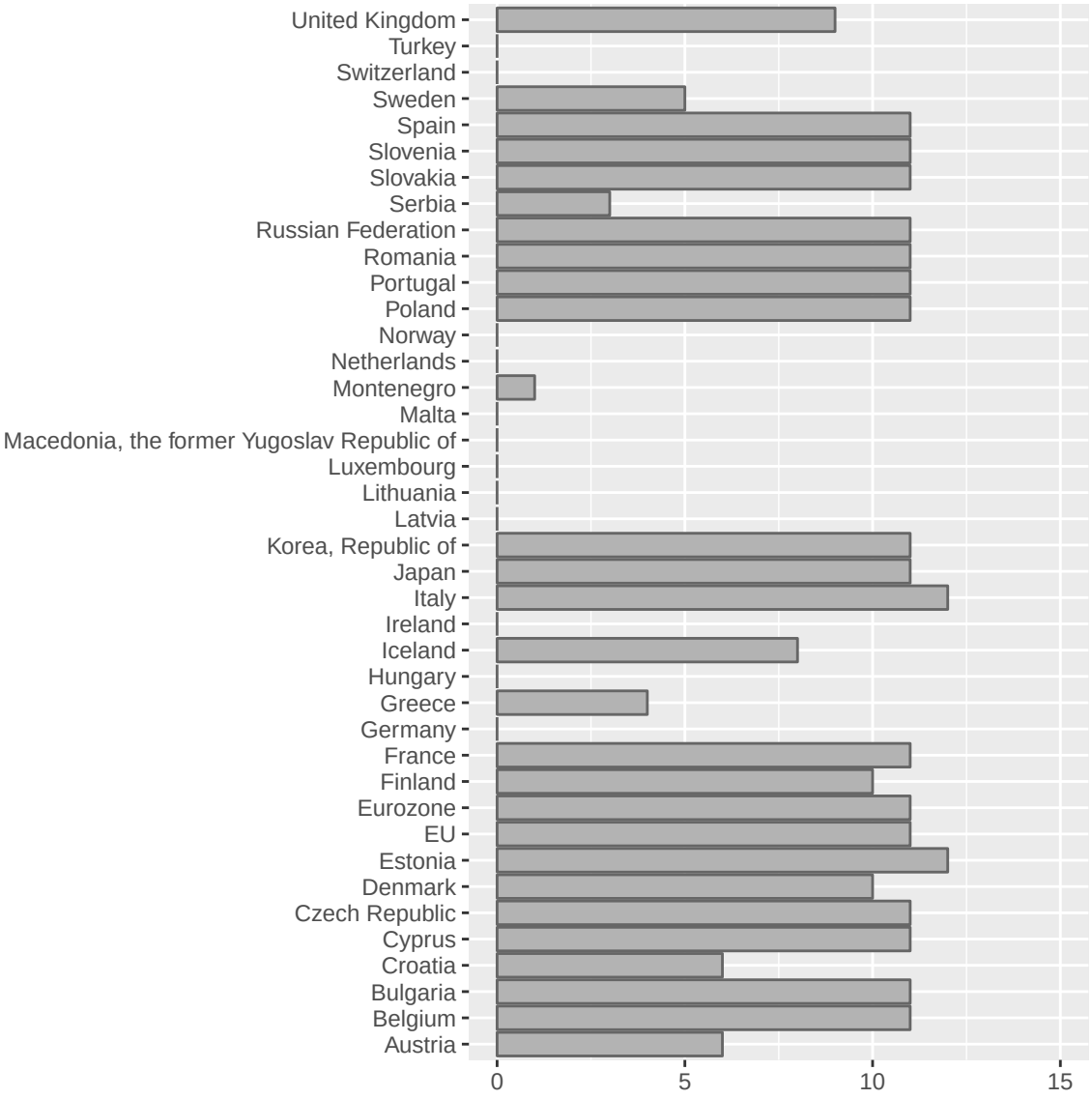

Fig 5: Numbers of data points for PNP shares by country. There are no data for a number of small countries and, perhaps surprisingly, no data for Germany. According to the OECD report 'OECD Science, Technology and Industry Scoreboard 2015 Innovation for growth and society', p176, the PNP sector is included in the Government sector for Germany, Luxembourg, Netherlands, and Norway, and PNP data are not available for Ireland and Turkey.

The time series without the gaps due to missing values and without the few data points for 2014 are displayed in Figure 6.

```
d7 <- datw %>% filter(!is.na(values), Year < 2014)
ggplot(d7, aes(Year, values)) + geom_path(aes(group=sectperf, colour=sectperf), size=0.5) +
  geom_point(aes(group=sectperf, colour=sectperf), size=0.5) + facet_wrap(~c1) +
  theme(axis.text.x=element_blank(), legend.position="bottom",
  legend.title=element_blank()) + labs(x=NULL, y=NULL)
```

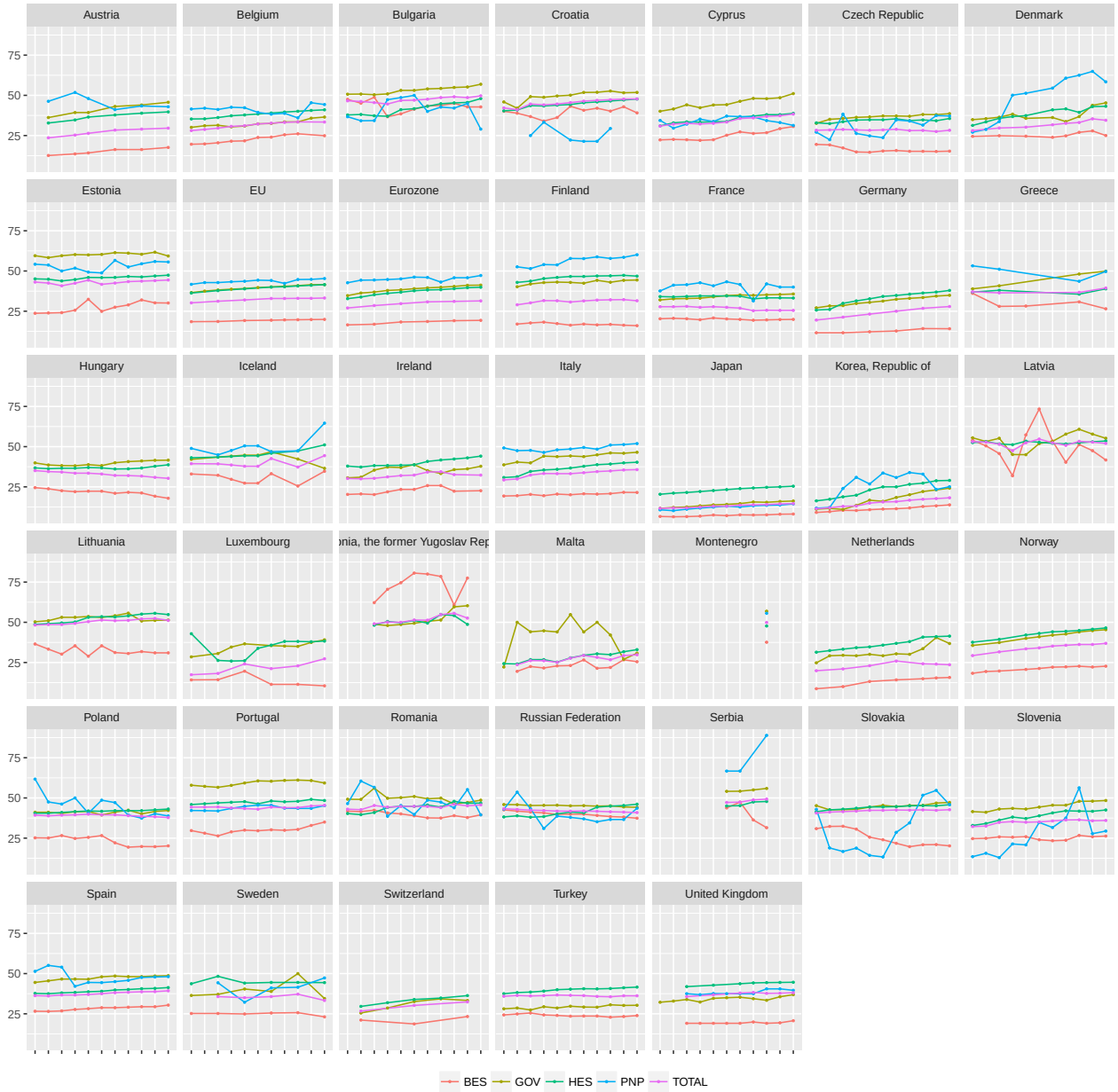

Fig 6: As in Figure 2, these time series are for shares of women researchers. A number of countries (e.g., Latvia, Malta, Slovakia, Slovenia) have individual series with high variability, probably due to small populations. The bigger countries have more stable patterns across the sectors, flat or slightly rising.

Concentrating on some of the biggest countries gives Figure 7.

```
d8 <- d7 %>% filter(c1 %in% c("Eurozone", "EU", "France", "Germany", "Italy", "Japan",
  "Russian Federation", "Spain", "Turkey", "United Kingdom"))
ggplot(d8, aes(Year, values)) + geom_path(aes(group=sectperf, colour=sectperf), size=0.5) +
  geom_point(aes(group=sectperf, colour=sectperf), size=0.5) + facet_wrap(~c1, nrow=2) +
  theme(axis.text.x=element_blank(), legend.position="bottom",
  legend.title=element_blank()) + labs(x=NULL, y=NULL)
```

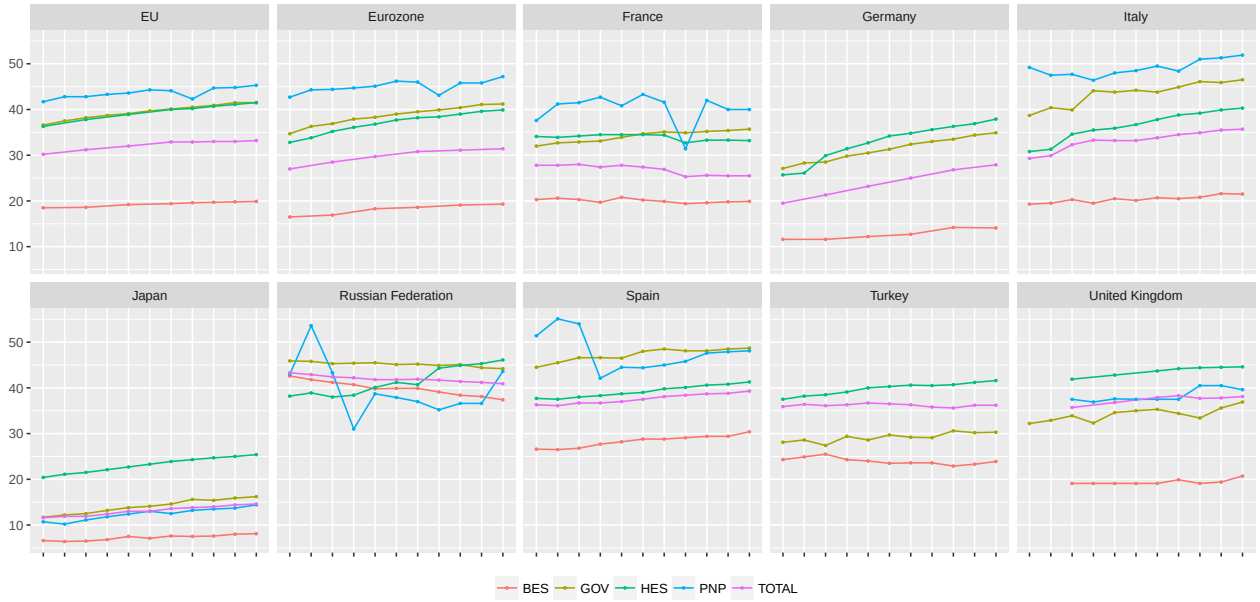

Fig 7: Shares of women researchers by sector for eight large countries, the Eurozone, and the EU as a whole. The TOTAL values for the EU are increasing a little. Japan has very low rates and Germany relatively low rates, but both are rising. The Russian Federation has high rates, as does Spain. The Business sector has almost always the lowest rates. PNP data for France in 2010 look odd. (Note: the vertical scale is not the same as in Figure 3.)

Note: Comparisons of statistics for different countries are fraught with difficulties. Definitions may vary and how the data are collected and reported will not be the same in each country.