The German NGO [Open Knowledge Foundation Deutschland e.V.](https://okfn.de/) has made German Trade Resister data available via the project [OffeneRegister.de](https://www.offeneregister.de/), together with the British NGO [opencorporates](https://opencorporates.com/). I checked the general accessibility of the data. In this quick follow-up post I follow an idea to map the gender balance of German corporate officers.

Here is the code for generating the necessary data. I focus on non-dismissed officers and use a first name list which originates from a list provided by a German computer magazine (c’t) This is not perfect but the best that I can do, given the data. I eliminate names that have duplicate gender classifications.

library(DBI)

library(tidyverse)

library(ggmap)

library(rgdal)

library(rgeos)

tmp <- tempdir()

db <- "~/Downloads/handelsregister.db"

con <- dbConnect(RSQLite::SQLite(), db)

sql <- "select id, company\_number as company\_id, current\_status, retrieved\_at, registered\_address from company"

res <- dbSendQuery(con, sql)

company <- dbFetch(res)

dbClearResult(res)

sql <- "select id as officer\_id, company\_id, firstname, start\_date, end\_date, dismissed from officer"

res <- dbSendQuery(con, sql)

officer <- dbFetch(res)

dbClearResult(res)

dbDisconnect(con)

company %>%

filter(current\_status == "currently registered",

!is.na(registered\_address)) %>%

mutate(plz = str\_extract(registered\_address, "\\d{5}")) %>%

filter(plz != "") %>%

select(company\_id, plz) -> company\_plz

officer %>%

filter(firstname != "",

!is.na(firstname),

is.na(dismissed)) -> officer\_firstname

company\_plz %>%

left\_join(officer\_firstname) %>%

filter(!is.na(firstname)) -> company\_firstname

nl <- read\_csv2("https://raw.githubusercontent.com/MatthiasWinkelmann/firstname-database/master/firstnames.csv") %>%

rename(firstname = name) %>%

select(firstname, gender) %>%

group\_by(firstname) %>%

filter(n() == 1) %>%

ungroup()

company\_firstname %>%

left\_join(nl) %>%

filter(gender == 'M' | gender == 'F') -> plz\_name\_gender

The overall share of female officers in Germany based on this classification is 17.0%. How does this share vary across German regions? Let’s get some OSM shape data.

# Shape file is based on OSM data.

# Source: https://www.suche-postleitzahl.org/downloads

download.file("https://www.suche-postleitzahl.org/download\_files/public/plz-gebiete.shp.zip",

file.path(tmp, "shape5.zip"))

unzip(file.path(tmp, "shape5.zip"), exdir = tmp)

plz5\_polys <- readOGR(file.path(tmp, "plz-gebiete.shp"))

download.file("https://www.suche-postleitzahl.org/download\_files/public/plz-2stellig.shp.zip",

file.path(tmp, "shape2.zip"))

unzip(file.path(tmp, "shape2.zip"), exdir = tmp)

plz2\_polys <- readOGR(file.path(tmp, "plz-2stellig.shp"))

Now we can create maps. First at the post area level (identified by the first to digits of the PLZ).

plz\_name\_gender %>%

mutate(plz2 = substr(plz, 1, 2)) %>%

group\_by(plz2) %>%

summarise(nobs = n(),

female = sum(gender == 'F')/n()) -> plz2\_female\_share

plz\_map <-

fortify(plz2\_polys, region = "plz") %>%

left\_join(plz2\_female\_share, by = c("id" = "plz2"))

plz2\_map <- ggplot(plz\_map, aes(x = long, y = lat, group = group, fill = female)) +

geom\_polygon(colour = NA, lwd=0, aes(group = group)) +

scale\_fill\_gradient2(name = "Share of female officers",

low = "orange", mid = "gray90", high = "purple",

midpoint = median(plz2\_female\_share$female, na.rm = TRUE),

breaks = 0.05\*(2:4), labels = scales::percent(0.05\*(2:4),

accuracy = 1)) +

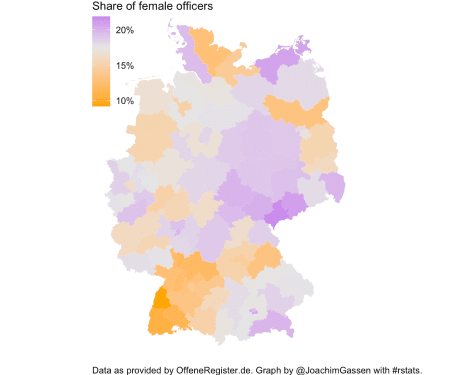
coord\_map() +

theme\_void() +

theme(legend.justification=c(0,1), legend.position=c(0,1), plot.caption = element\_text(hjust = 0)) +

labs(caption = "Data as provided by OffeneRegister.de. Graph by @JoachimGassen with #rstats.")

plz2\_map



This does not look good for Baden-Wuerttemberg. Also, it seems as if the female officer share is larger in former East Germany. Without further analysis I can only guess that regions with significant tourist industry (Northern Mecklenburg-Vorpommern, Southern Bavaria) tend to have a higher share of female officers while areas with a significant agricultural sector (e.g., Northern Brandenburg) have lower female representation.

Can we say something when we look at a more granular level (PLZs instead of two digit PLZ regions)? Here I limit the data to PLZs that list at least 30 officers.

plz\_name\_gender %>%

group\_by(plz) %>%

summarise(nobs = n(),

female = sum(gender == 'F')/n()) %>%

filter(nobs >= 30) -> plz\_female\_share

plz\_map <-

fortify(plz5\_polys, region = "plz") %>%

left\_join(plz\_female\_share, by = c("id" = "plz"))

plz5\_map <- ggplot(plz\_map, aes(x = long, y = lat, group = group, fill = female)) +

geom\_polygon(colour = NA, lwd=0, aes(group = group)) +

scale\_fill\_gradient2(name = "Share of female officers",

low = "orange", mid = "gray90", high = "purple",

midpoint = median(plz\_female\_share$female, na.rm = TRUE),

breaks = 0.1\*(0:5), labels = scales::percent(0.1\*(0:5),

accuracy = 1)) +

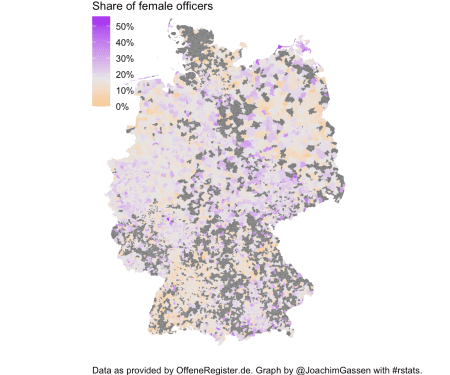
coord\_map() +

theme\_void() +

theme(legend.justification=c(0,1), legend.position=c(0,1), plot.caption = element\_text(hjust = 0)) +

labs(caption = "Data as provided by OffeneRegister.de. Graph by @JoachimGassen with #rstats.")

plz5\_map



A similar but significantly more patchy picture surfaces. How does this look like when we focus on cities (e.g., Berlin)?

plz\_name\_gender %>%

group\_by(plz) %>%

summarise(nobs = n(),

female = sum(gender == 'F')/n()) -> plz\_female\_share

berlin\_polys <- plz5\_polys[str\_detect(plz5\_polys$note, "\\d{5} Berlin"), ]

plz\_map <-

fortify(berlin\_polys, region = "plz") %>%

left\_join(plz\_female\_share, by = c("id" = "plz"))

berlin\_map <- ggplot(plz\_map, aes(x = long, y = lat, group = group, fill = female)) +

geom\_polygon(colour = NA, lwd=0, aes(group = group)) +

scale\_fill\_gradient2(name = "Share of female officers",

low = "orange", mid = "gray90", high = "purple",

midpoint = median(plz\_female\_share$female, na.rm = TRUE),

breaks = 0.1\*(0:5), labels = scales::percent(0.1\*(0:5),

accuracy = 1)) +

coord\_map() +

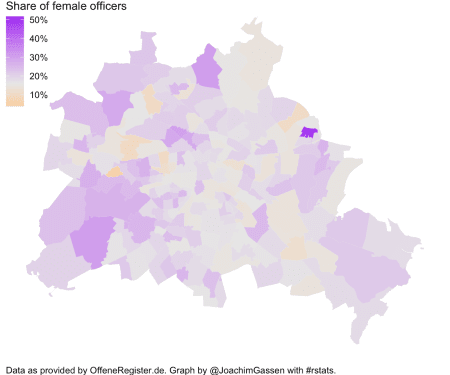
theme\_void() +

theme(legend.justification=c(0,1), legend.position=c(0,1),

plot.caption = element\_text(hjust = 0)) +

labs(caption = "Data as provided by OffeneRegister.de. Graph by @JoachimGassen with #rstats.")

berlin\_map



Well… Not much interesting here (besides the ladies camp in Marzahn!) but maybe other cities/areas reveal more insightful patterns. I will leave this question to the interested reader.