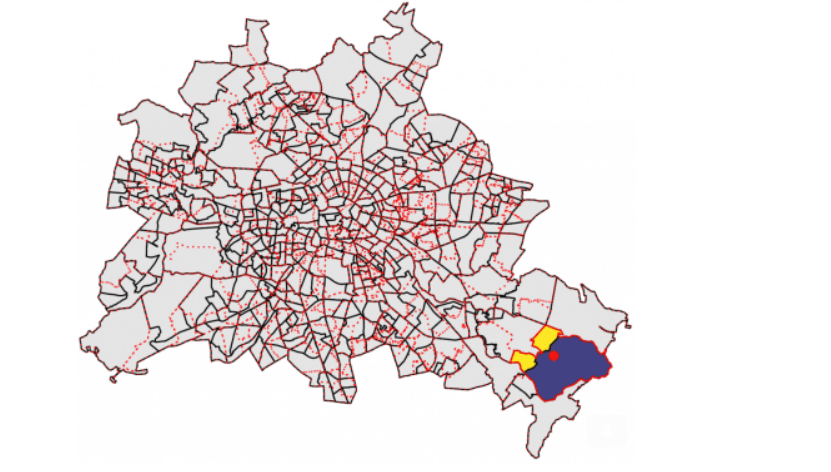
**Data**



For this example, I’d like to compare the percentage of children whose parents obtain social welfare in the neighborhood regions around public and private primary schools in Berlin. This blog post concentrates on how to join the point samples (the schools) with the surrounding statistical regions and calculate a spatially weighted average the welfare rate, so I will present only a few descriptive results in the end.

We will work with several datasets: The first spatial dataset contains the shape of the statistical regions in Berlin, the second dataset contains the socioeconomic data for these regions, the third and fourth datasets contain the locations and other attributes of public and private primary schools in Berlin, respectively.

All data and the code are available in a [GitHub repository](https://github.com/WZBSocialScienceCenter/spatially_weighted_avg). We will use the [*sf* package](https://cran.r-project.org/web/packages/sf/index.html) for working with spatial data in R, [*dplyr*](https://dplyr.tidyverse.org/) for data management and [*ggplot2*](https://ggplot2.tidyverse.org/) for a few more advanced visualizations, i.e. when base plot() is not sufficient.

library(sf)

library(dplyr)

library(ggplot2)

**Socioeconomic data for statistical regions**

We will at first load a dataset with the most granular official statistical regions for Berlin, called [*Planungsräume* (planning areas)](https://www.stadtentwicklung.berlin.de/planen/basisdaten_stadtentwicklung/lor/). We select the area ID and name as spatial attributes. The result is a spatial dataframe (a *simple feature (sf)* collection).

bln\_plan <- read\_sf('data/berlin\_plr.shp') %>%

mutate(areaid = as.integer(SCHLUESSEL)) %>%

select(areaid, name = PLR\_NAME)

head(bln\_plan)

Simple feature collection with 6 features and 2 fields

[...]

Projected CRS: ETRS89 / UTM zone 33N

areaid name geometry

1 1011101 Stülerstr. (((387256.6 5818552, 387323.1 5818572, 387418.9 5…

2 1011102 Großer Tiergar… (((386767.5 5819393, 386768.3 5819389, 386769.6 5…

3 1011103 Lützowstr. (((387952.6 5818275, 387986.7 5818313, 387994.6 5…

4 1011104 Körnerstr. (((388847.1 5817875, 388855.5 5817899, 388865.1 5…

5 1011105 Nördlicher Lan… (((388129.5 5819015, 388157.1 5819017, 388170.8 5…

6 1011201 Wilhelmstr. (((389845.7 5819286, 389840.9 5819311, 389846.1 5…

When printing this dataframe, the header reveals another important information: The coordinate reference system (CRS) of this dataset is [ETRS89 / UTM zone 33N](http://epsg.io/25833). We will later need to make sure that the coordinates of the school locations and the coordinates of the planning areas use the same coordinate system.

This data can be joined with socioeconomic information provided from official sources. Luckily, [Helbig/Salomo 2021](https://shiny2.wzb.eu/konrad/salomo_helbig_dashboard/) compiled these information for some cities in Germany (available for [download](https://shiny2.wzb.eu/konrad/salomo_helbig_dashboard/download/Helbig,%20Salomo%20-%20Sozialraeumliche%20Ungleichheiten%20-%20Daten%20Stand%202021-01-26.xlsx)) among which is data for Berlin from 2020. I’ve created an excerpt with percentages of residents receiving social welfare (welfare) and percentage of children under 15 years whose parents receive social welfare (welfare\_chld):

bln\_welfare <- read.csv('data/berlin\_welfare.csv', stringsAsFactors = FALSE)

head(bln\_welfare)

areaid areaname welfare welfare\_chld

1 1011101 Stülerstraße 10.09 15.44

2 1011102 Großer Tiergarten 4.76 0.00

3 1011103 Lützowstraße 22.21 36.80

4 1011104 Körnerstraße 24.81 42.14

5 1011105 Nördlicher Landwehrkanal 2.82 3.53

6 1011201 Wilhelmstraße 12.13 19.03

We can use the area ID for augmenting the planning areas with the welfare statistics. We’re joining a spatial with an ordinary dataframe, so we can use dplyr’s inner\_join. Before that we can check that for each planning area we have welfare statistics information and vice versa:[1]Note that when joining spatial and ordinary dataframes, the order of arguments in the join function matters. If you have a spatial dataframe on the “left side” (x argument), the result … Continue reading

setequal(bln\_plan$areaid, bln\_welfare$areaid)

[1] TRUE

bln <- inner\_join(bln\_plan, bln\_welfare, by = 'areaid') %>%

select(-name)

head(bln)

Simple feature collection with 6 features and 4 fields

[...]

Projected CRS: ETRS89 / UTM zone 33N

areaid geometry areaname welfare welfare\_chld

1 1011101 (((387256.6 5818552, 387323.1 581… Stülerstr… 10.1 15.4

2 1011102 (((386767.5 5819393, 386768.3 581… Großer Ti… 4.76 0

3 1011103 (((387952.6 5818275, 387986.7 581… Lützowstr… 22.2 36.8

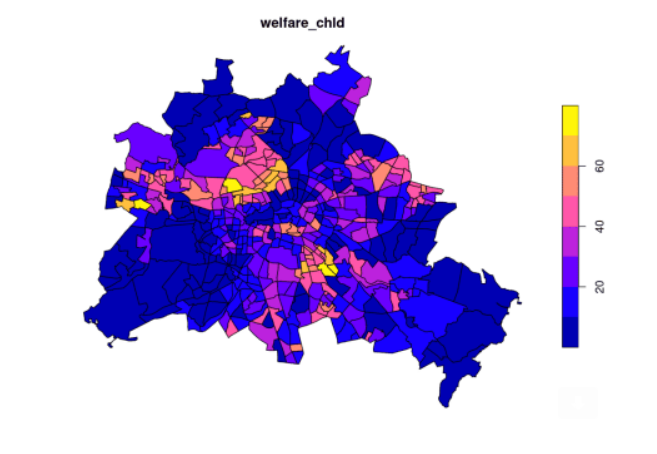
4 1011104 (((388847.1 5817875, 388855.5 581… Körnerstr… 24.8 42.1

5 1011105 (((388129.5 5819015, 388157.1 581… Nördliche… 2.82 3.53

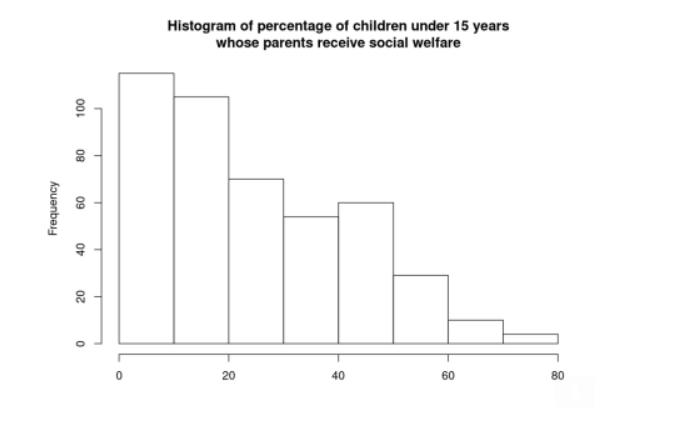
6 1011201 (((389845.7 5819286, 389840.9 581… Wilhelmst… 12.1 19.0

A quick plot confirms that it is similar to the figure from the [dashboard](https://shiny2.wzb.eu/konrad/salomo_helbig_dashboard/?_state_id_=49379e5c63e0e14c).[2]I prefer using the base plot function for quick exploration of spatial data and usually only turn to ggplot2 for more advanced or “publication ready” plots. The help page for plot.sf … Continue reading

plot(bln['welfare\_chld'])



The median percentage of children whose parents receive social welfare is ~20% with an interquartile range of about 29%. The following shows the distribution of this welfare rate:



**Public and private primary schools**

The [Berlin geodata catalog “FIS Broker”](https://stadtentwicklung.berlin.de/geoinformation/fis-broker/) provides the [locations of public schools in Berlin](https://fbinter.stadt-berlin.de/fb/index.jsp?loginkey=zoomStart&mapId=schulen@senstadt&bbox=362719,5798847,423243,5838687).[3]The catalog is a bit clumsy to use, but actually works quite well: You search for the data, get an URL to the WFS endpoint from the data’s metainformation panel and use that URL to obtain the … Continue reading I obtained the data and converted it to GeoJSON, which we can now load. We’ll only retain primary schools and add a variable denoting that these are public schools. We also see that the CRS of the school locations matches the CRS of the Berlin statistical regions data.

pubschools <- read\_sf('data/berlin\_pubschools.geojson') %>%

filter(SCHULART == 'Grundschule') %>%

select(name = NAME) %>%

mutate(ownership = 'pub', .before = 1)

head(pubschools)

Simple feature collection with 6 features and 2 fields

[...]

Projected CRS: ETRS89 / UTM zone 33N

ownership name geometry

1 pub Grundschule am Arkonaplatz (391497.3 5821994)

2 pub Papageno-Grundschule (390876.3 5821514)

3 pub Kastanienbaum-Grundschule (391579.6 5820819)

4 pub Grundschule Neues Tor (390139.2 5820930)

5 pub GutsMuths-Grundschule (393086.4 5819617)

6 pub Grundschule am Brandenburger Tor (390255 5819341)

Now to the private schools’ locations. [Marcel Helbig](https://wzb.eu/de/personen/marcel-helbig), [Rita Nikolai](https://www.erziehungswissenschaften.hu-berlin.de/de/institut/abteilungen/didaktik/As%20Kol/nikolai) and me collected data on school locations in East Germany from 1992 to 2015 in order to analyze the [development of the network of schools in East Germany and which role private schools play in it](https://bibliothek.wzb.eu/wzbrief-bildung/WZBriefBildung382018_helbig_konrad_nikolai.pdf) (Helbig/Konrad/Nikolai 2018). Besides creating an [interactive map](https://schulenkarte.wzb.eu/), we also [published the data](https://schulenkarte.wzb.eu/#daten) and are planning an update with newer data (until 2020) from which will we now use an excerpt. This dataset provides school locations from 2019 as [longitude/latitude WGS84 coordinates](https://gisgeography.com/wgs84-world-geodetic-system/) which we can load and convert to a spatial dataset using st\_as\_sf. We also transform these locations to the ETRS89 CRS used in all prior spatial datasets.

privschools <- read.csv('data/grundschulen\_berlin\_2019.csv',

stringsAsFactors = FALSE) %>%

filter(traeger == 'priv') %>%

select(ownership = traeger, name, lng, lat) %>%

st\_as\_sf(coords = c('lng', 'lat'), crs = 4326) %>% # EPSG 4326 is WGS84 lat/long

st\_transform(crs = st\_crs(pubschools)) # transform to same CRS as publ. schools

head(privschools)

Simple feature collection with 6 features and 2 fields

[...]

Projected CRS: ETRS89 / UTM zone 33N

ownership name geometry

1 priv Freie Waldorfschule Berlin Mitte POINT (391783.5 5820737)

2 priv Freie Waldorfschule Kreuzberg POINT (391544 5818222)

3 priv Freie Waldorfschule am Prenzlauer Berg POINT (395177.8 5822721)

4 priv Annie-Heuser-Schule POINT (384902.8 5816817)

5 priv Freie Waldorfschule Havelhöhe - Eugen Kolisko POINT (374724.8 5813820)

6 priv Rudolf-Steiner-Schule Berlin POINT (382728.5 5813179)

The variable ownership encodes whether a given facility is a public (“pub”) or private (“priv”) primary school. We can now append the public and private primary schools datasets to form a single schools dataset. The public school data comes from 2020 and the private school data from 2019, but this shouldn’t be an issue because the number of public and private schools has been quite stable in recent years.

schools <- bind\_rows(pubschools, privschools) %>%

mutate(schoolid = 1:nrow(.), .before = 1)

head(schools)

Simple feature collection with 6 features and 3 fields

[...]

Projected CRS: ETRS89 / UTM zone 33N

schoolid ownership name geometry

1 1 pub Grundschule am Arkonaplatz (391497.3 5821994)

2 2 pub Papageno-Grundschule (390876.3 5821514)

3 3 pub Kastanienbaum-Grundschule (391579.6 5820819)

4 4 pub Grundschule Neues Tor (390139.2 5820930)

5 5 pub GutsMuths-Grundschule (393086.4 5819617)

6 6 pub Grundschule am Brandenburger Tor (390255 5819341)

In our dataset we now have 361 public and 71 private primary schools in Berlin.

**Public / private primary schools and poverty by statistical region**

Both datasets use the same coordinate system now, so we can plot the school locations on top of the planning areas. I will use ggplot2 this time to make a choropleth map of the welfare\_chld variable and overlay that with the public and private primary school locations.

ggplot() +

geom\_sf(aes(fill = welfare\_chld), color = 'black', data = bln) +

geom\_sf(aes(color = ownership), size = 1, alpha = 0.75, data = schools) +

scale\_fill\_binned(type = 'viridis', guide = guide\_bins(title = '% Welfare')) +

scale\_color\_manual(values = c('pub' = '#c767cb', 'priv' = '#cdc566'),

labels = c('public school', 'private school'),

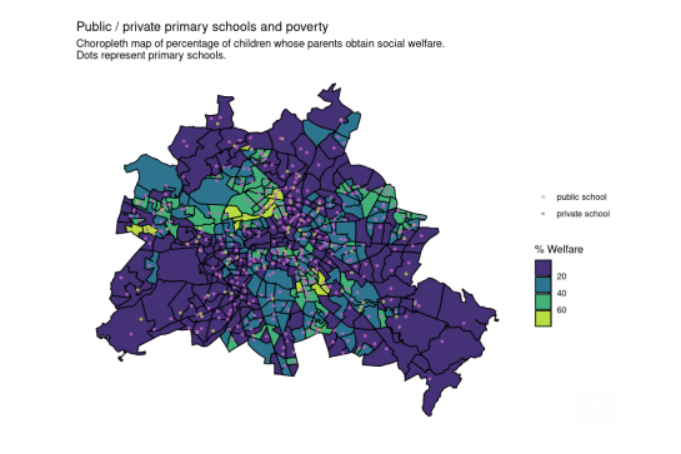
guide = guide\_legend(title = '')) +

coord\_sf(datum = NA) + # disable graticule

labs(title = "Public / private primary schools and poverty",

subtitle = "Choropleth map of percentage of children whose parents obtain social welfare.\nDots represent primary schools.") +

theme\_minimal()



From the figure alone, it’s probably hard to assess whether there’s a pattern in the distribution of private and public schools regarding areas with higher welfare rate in the city. In order to compare the social welfare statistics of regions around private schools with those around public schools, we can join the schools’ data with the socioeconomic information of the planning areas they’re located in. This can be done with a [spatial join](https://geocompr.robinlovelace.net/spatial-operations.html#spatial-joining) using [st\_join](https://r-spatial.github.io/sf/reference/st_join.html). By default, this function joins the spatial features of the first argument with features of the second argument **when they intersect** – in our case this means a school is linked with the planning area it’s located in. Note that the order of arguments matters here and that the spatial geometry of the first argument is retained in the resulting dataset.

schools\_plan <- st\_join(schools, bln)

head(schools\_plan)

Simple feature collection with 6 features and 7 fields

[...]

Projected CRS: ETRS89 / UTM zone 33N

schoolid ownership name geometry areaid areaname welfare welfare\_chld

1 1 pub Grundsc... (391497.3 5821994) 1011402 Arkonap... 4.46 3.53

2 2 pub Papagen... (390876.3 5821514) 1011401 Invalid... 6.29 7.7

3 3 pub Kastani... (391579.6 5820819) 1011302 Oranien... 8.16 10.4

4 4 pub Grundsc... (390139.2 5820930) 1011301 Charité... 3.68 3.92

5 5 pub GutsMut... (393086.4 5819617) 1011304 Karl-Ma... 23.5 36.3

6 6 pub Grundsc... (390255 5819341) 1011201 Wilhelm... 12.1 19.0

We can see that the schools’ data was linked with the data from the planning areas. We should also check whether there’s a school that was not located in any planning area (this may for example happen when a school is very close to the Berlin-Brandenburg border):

sum([is.na](http://is.na)(schools\_plan$areaid))

[1] 0

All schools were linked with their planning area, so we can now compare the percentage of children whose parents obtain social welfare between public and private primary schools:

ggplot(schools\_plan) +

geom\_violin(aes(x = ownership, y = welfare\_chld), draw\_quantiles = c(0.5)) +

geom\_jitter(aes(x = ownership, y = welfare\_chld), alpha = 0.25) +

scale\_x\_discrete(labels = c('pub' = 'public primary schools',

'priv' = 'private primary schools')) +

labs(title = 'Percentage of children whose parents obtain social welfare',

x = '', y = '% welfare')

Our descriptive results indicate that the median percentage of children whose parents obtain social welfare is around six percent higher in the statistical regions around public schools than around private schools: [4]I’m using st\_drop\_geometry here, because otherwise a [spatial aggregation](https://geocompr.robinlovelace.net/spatial-operations.html#spatial-aggr) would be performed which takes much longer to compute and is not necessary here.

st\_drop\_geometry(schools\_plan) %>%

group\_by(ownership) %>%

summarise(median\_welfare\_chld = median(welfare\_chld))

ownership median\_welfare\_chld

1 priv 16.8

2 pub 22.8

This is an interesting descriptive result and we may continue with our spatial analysis from here. However, our current approach doesn’t consider the catchment area of a school correctly: Children from nearby planning areas will most likely visit a school, but at the moment we only consider the one planning area in which a school is located. As an example, let’s zoom to school #388 “Evangelische Schule Berlin Buch” in the north of Berlin.