The Oaxaca-Blinder decomposition in Stata: an update

Ben Jann

University of Bern

2025 German Stata Conference
Hamburg, March 28, 2025

Outline

Introduction

- In 2008, I published Stata command `oaxaca`, which implements the Oaxaca-Blinder (OB) decomposition technique (Jann 2008).
- The OB decomposition (Blinder 1973, Oaxaca 1973) is used to analyze differences in outcomes between groups, such as the wage gap by gender or race (for a general overview of counterfactual decomposition methods see Fortin et al. 2011).
- The technique is highly popular in applied research (over 10'000 citations of both Oaxaca 1973 and Blinder 1973 on Google Scholar; about 3000 citations of Jann 2008).
- Over the years, both the functionality of Stata and the literature on decomposition methods have evolved, so that an update of the `oaxaca` command is long overdue.

Desired features

- 👍 Overall and detailed decompositions supporting different solutions to the index problem (see, e.g., Jann 2008).
- 👍 Variance estimation (Jann 2008).
  - Support for survey estimation (`pweights`, clustered standard errors, general support for `svy`).
  - Provided by existing `oaxaca`, but there is scope for improvement.
- 👍 Support for binary dependent variables (Yun 2004)
- 👍 „Normalization“ for categorical predictors (Yun 2008)

(👍 = supported by current version of `oaxaca`; 👎 = currently not supported)

Desired features

- Support for factor variables.
- Support for more than two groups (series of decompositions against a reference group or an overall average).
- Alternative “normalization” approaches (Kim 2013, Horrace and Oaxaca 2001).
- Decompositions based on reweighted techniques (DiNardo et al. 1996) such as IPW or entropy balancing (Hainmueller 2012).
- Decompositions for arbitrary statistics (rather than just the mean) based on recentered influence functions (RIF) (Firpo et al 2009, 2018, Rios-Avila 2020).
- Support for difference-in-differences decompositions (Smith and Welch 1987, Kröger and Hartmann 2021).

Desired features

- There are further decomposition approaches for which an integration into `oaxaca` appears less obvious. For example:
  - Fairlie (2005) decomposition for binary dependent variables (see `fairlie` by Jann 2006 for an implementation).
  - Juhn et al. (1991, 1993) decompositions based on residual distributions (see `jmpierce` and `jmpierce2` by Jann 2005a,b for implementations).
  - Distributions based on quantile regression process or distribution regression (Chernozhukov et al. 2013; see `cdist` by Jann 2023a for an implementation).

Methods

- The general idea of counterfactual decomposition methods is to decompose a group difference in a distributional statistic (Δ^ν) into a part that is related to compositional differences between the groups (Δ^ν_X) and a part that is related to group-specific "mechanisms" (structural functions) (Δ^ν_S).

$$\Delta^\nu = \Delta^\nu_X + \Delta^\nu_S$$

- The classical Oaxaca-Blinder decomposition (a) focuses on the mean and (b) uses linear regression for the structural function. In its simplest form, it can be written as

$$\underbrace{\bar{Y}^1 - \bar{Y}^2}_{\widehat{\Delta}^\mu} = \underbrace{(\bar{X}^1 - \bar{X}^2)\widehat{\beta}^1}_{\widehat{\Delta}^\mu_X} + \underbrace{\bar{X}^2(\widehat{\beta}^1 - \widehat{\beta}^2)}_{\widehat{\Delta}^\mu_S}$$

where $\bar{Y}^g$ is the mean of the outcome, $\bar{X}^g$ is the mean vector of characteristics, and $\widehat{\beta}^g$ is the coefficient vector of a regression of Y on X in group g.

Methods

- Variants of the classical decomposition differ in how exactly the group means and coefficients are combined to form the two terms (and some variants also have a third term), but the basic principle is the same.
- In case of reweighting, weights are computed that balance the distribution of characteristics between groups, and a (four-term) decomposition is obtained by comparing weighted and unweighted results.
- In case of RIF decomposition, Y is replaced by the (group-specific) recentered influence function of statistic $\nu(F_Y)$ (e.g. the RIF of the Gini coefficient of Y). All else stays the same.
- In case of a difference-in-differences decomposition, an additional group layer (e.g. two time points) is added and additional terms are defined, but the logic stays the same.

Methods

- The basic message is that we can put all of the above into a common framework without much conceptual complication.
- Variance estimation (taking account of reweighting and including support for `svy`) can easily be implemented using influence functions (see Jann 2019, 2020b, 2021).
- The basic elements we need are:
  - Mean estimates (influence function = demeaned variable).
  - Coefficients from regression models (influence functions for linear regression and maximum likelihood estimators are very easy to obtain; just need the scores and the information matrix).
  - Recentered influence function for the statistic of interest (a wide variety of RIFs is provided by command `dstat` by Jann 2020a).
- However, as usual, there are many little details to take care of.

Syntax

New `kob` command:[1]

> `kob` *statistic* *depvar* [*indepvars*] [*if*] [*in*] [*weight*] ,
> `by(`*groupvar* [*groupvar2*]`)`
> [`reweight`[`(`*varlist*`)`] `vce(`*vcetype*`)` *options*]

- *statistic*: any statistic allowed by `dstat`
- *groupvar2*: for DID decomposition
- `reweight()`: apply reweighting
- *vcetype*: `robust`, `cluster`, `svy`, `bootstrap`, `jackknife`
- *options*: type of decomposition, reporting, etc.

[1] `kob` = Kitagawa-Oaxaca-Blinder (see Kitagawa 1955); the name of the command may still change.

Example: Private–public gap in wage inequality

Data from the German Socio-Economic Panel (GSOEP), wave 2015.

```
. use gsoep-extract, clear
(Example data based on the German Socio-Economic Panel)

. keep if wave==2015
(29,970 observations deleted)

. keep if inrange(age, 25, 55)
(5,671 observations deleted)

. generate lnwage = ln(wage)
(1,709 missing values generated)

. summarize public wage lnwage yeduc expft weight psu
```

Variable	Obs	Mean	Std. dev.	Min	Max
public	5,770	.2353553	.4242574	0	1
wage	5,600	17.57278	9.858855	3.03	121.42
lnwage	5,600	2.736721	.5062968	1.108563	4.799255
yeduc	7,121	12.28823	2.783974	7	18
expft	7,274	11.63359	9.556508	0	39.5
weight	7,309	2204.229	3025.122	3.3	32681.6
psu	7,309	2437.243	1413.001	1	4893

Private–public wage gap

Current oaxaca implementation:

```
. generate expft2 = expft^2
(35 missing values generated)

. oaxaca lnwage yeduc expft expft2 [pw=weight], by(public) weight(1) ///
>     nodetail vce(cluster psu)

Blinder-Oaxaca decomposition                    Number of obs     =      5,458
                                                Model             =     linear
Group 1: public = 0                             N of obs 1        =      4,184
Group 2: public = 1                             N of obs 2        =      1,274

    explained: (X1 - X2) * b1
  unexplained: X2 * (b1 - b2)

                              (Std. err. adjusted for 2,036 clusters in psu)
------------------------------------------------------------------------------
             |               Robust
      lnwage | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   2.732109   .0139572   195.75   0.000     2.704754    2.759465
     group_2 |   2.866068   .0213964   133.95   0.000     2.824132    2.908005
  difference |  -.1339592   .0249932    -5.36   0.000     -.182945   -.0849735
   explained |  -.1262644   .0170697    -7.40   0.000    -.1597204   -.0928084
 unexplained |  -.0076948   .0226291    -0.34   0.734    -.0520471    .0366575
------------------------------------------------------------------------------
```

Private–public wage gap

New kob command:

```
. kob mean lnwage yeduc c.expft##c.expft [pw=weight], by(public) vce(cluster psu)
Kitagawa-Oaxaca-Blinder decomposition               Number of obs     =      5,458
                                                    Statistic         =       mean
                                                    Model             =     linear
Group 1: public = 0                                 N of obs 1        =      4,184
Group 2: public = 1                                 N of obs 2        =      1,274
delta_X: (X1 - X2) * b1
delta_S: X2 * (b1 - b2)
                                (Std. err. adjusted for 2,036 clusters in psu)
------------------------------------------------------------------------------
             |               Robust
      lnwage | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
levels       |
     group_1 |   2.732109   .0141087   193.65   0.000     2.704457    2.759762
     group_2 |   2.866068   .0221403   129.45   0.000     2.822674    2.909463
-------------+----------------------------------------------------------------
g1_vs_g2     |
         gap |  -.1339592   .0256495    -5.22   0.000    -.1842314   -.0836871
     delta_X |  -.1262644   .0171534    -7.36   0.000    -.1598845   -.0926443
     delta_S |  -.0076948   .0226074    -0.34   0.734    -.0520046    .0366149
------------------------------------------------------------------------------
```

Private–public gap in wage inequality

Gini coefficient:

```
. kob gini wage yeduc c.expft##c.expft [pw=weight], by(public) vce(cluster psu)
Kitagawa-Oaxaca-Blinder decomposition            Number of obs     =      5,458
                                                 Statistic         =       gini
                                                 Model             =     linear
Group 1: public = 0                              N of obs 1        =      4,184
Group 2: public = 1                              N of obs 2        =      1,274
delta_X: (X1 - X2) * b1
delta_S: X2 * (b1 - b2)
                                (Std. err. adjusted for 2,036 clusters in psu)
------------------------------------------------------------------------------
             |               Robust
        wage | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
levels       |
     group_1 |   .2783233   .0056676    49.11   0.000      .267215    .2894316
     group_2 |   .2213006   .0081333    27.21   0.000     .2053596    .2372415
-------------+----------------------------------------------------------------
g1_vs_g2     |
         gap |   .0570227   .0098305     5.80   0.000     .0377553    .0762901
     delta_X |  -.0093274   .0048026    -1.94   0.052    -.0187404    .0000856
     delta_S |   .0663501   .0109198     6.08   0.000     .0449477    .0877525
------------------------------------------------------------------------------
```

Private–public gap in wage inequality

Variance of logarithm:

```
. kob vlog wage yeduc c.expft##c.expft [pw=weight], by(public) vce(cluster psu)
Kitagawa-Oaxaca-Blinder decomposition           Number of obs     =      5,458
                                                Statistic         =       vlog
                                                Model             =     linear
Group 1: public = 0                             N of obs 1        =      4,184
Group 2: public = 1                             N of obs 2        =      1,274
delta_X: (X1 - X2) * b1
delta_S: X2 * (b1 - b2)
                                (Std. err. adjusted for 2,036 clusters in psu)
------------------------------------------------------------------------------
             |               Robust
        wage | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
levels       |
     group_1 |   .2508589   .0098729    25.41   0.000     .2315083    .2702095
     group_2 |   .1970238   .0178798    11.02   0.000     .1619801    .2320676
-------------+----------------------------------------------------------------
g1_vs_g2     |
         gap |   .0538351   .0203442     2.65   0.008     .0139613    .0937089
     delta_X |  -.0207097   .0080783    -2.56   0.010    -.0365429   -.0048765
     delta_S |   .0745448   .0206431     3.61   0.000     .0340851    .1150045
------------------------------------------------------------------------------
```

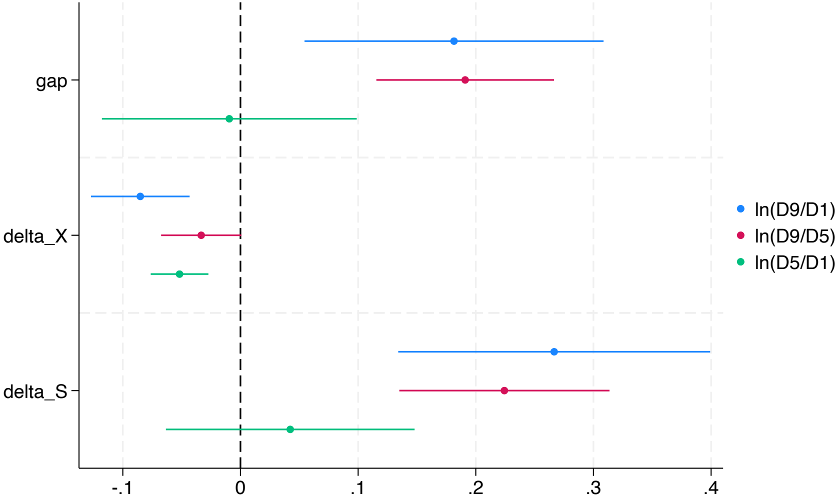
Could also type:

```
. kob variance lnwage yeduc c.expft##c.expft [pw=weight], by(public) vce(cluster psu)
Kitagawa-Oaxaca-Blinder decomposition             Number of obs     =      5,458
                                                  Statistic         =   variance
                                                  Model             =     linear
Group 1: public = 0                               N of obs 1        =      4,184
Group 2: public = 1                               N of obs 2        =      1,274
delta_X: (X1 - X2) * b1
delta_S: X2 * (b1 - b2)
                                   (Std. err. adjusted for 2,036 clusters in psu)
```

lnwage	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
levels						
group_1	.2508589	.0098729	25.41	0.000	.2315083	.2702095
group_2	.1970238	.0178798	11.02	0.000	.1619801	.2320676
g1_vs_g2						
gap	.0538351	.0203442	2.65	0.008	.0139613	.0937089
delta_X	-.0207097	.0080783	-2.56	0.010	-.0365428	-.0048765
delta_S	.0745448	.0206431	3.61	0.000	.0340851	.1150045

Private–public gap in wage inequality

Quantile ratios:



Private–public gap in wage inequality


```
kob iqr(10,90) lnwage yeduc c.expft##c.expft [pw=weight], by(public) vce(cluster psu)
est sto d9d1
kob iqr(50,90) lnwage yeduc c.expft##c.expft [pw=weight], by(public) vce(cluster psu)
est sto d9d5
kob iqr(10,50) lnwage yeduc c.expft##c.expft [pw=weight], by(public) vce(cluster psu)
est sto d5d1
coefplot d9d1 d9d5 d5d1, keep(g1_vs_g2:) xline(0) plotl(ln(D9/D1) ln(D9/D5) ln(D5/D1))
```

Conclusions

- A general and flexible command for Oaxaca-Blinder decompositions, including RIFs and reweighting as well as support for survey estimation, is straightforward to implement (at least conceptually).
- First steps have been taken ...
- ... but I am not quite done yet.
- I was too busy working on other stuff such as, e.g., `geoplot` (Jann 2023b).
- Also check out the new `crosswalk` command for bulk recoding (Jann 2025).

References I

- Blinder A.S. 1973. Wage discimination: Reduced form and structural estimates. *Journal of Human Resources* 8: 436–455.
- Chernozhukov, V., I. Fernández-Val, B. Melly (2013). Inference on Counterfactual Distributions. *Econometrica* 81:2205–2268.
- DiNardo, J.E., N. Fortin, T. Lemieux. 1996. Labour Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica* 64:1001–1046.
- Fairlie, R.W. 2005. An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. *Journal of Economic and Social Measurement* 30:305–316.
- Firpo, S., N. Fortin, T. Lemieux (2009). Unconditional Quantile Regressions. *Econometrica* 77:953–973.
- Firpo, S., N. Fortin, T. Lemieux. 2018. Decomposing Wage Distributions Using Recentered Influence Function Regressions. *Econometrics* 6: 28.
- Fortin, N., T. Lemieux, S. Firpo. 2011. Decomposition Methods in Economics. P. 1–102 in: O. Ashenfelter and D. Card (eds.). *Handbook of Labor Economics*. Amsterdam: Elsevier.
- Hainmueller, J. 2012. Entropy Balancing: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis* 20:25–46.

References II

- Horrace, W.C., R.L. Oaxaca. 2001. Inter-Industry Wage Differentials and the Gender Wage Gap: An Identification Problem. *Industrial and Labor Relations Review* 54(3):611–618.
- Jann, B. 2005a. jmpierce: Stata module to perform Juhn-Murphy-Pierce decomposition. Available from https://ideas.repec.org/c/boc/bocode/s449301.html.
- Jann, B. 2005b. jmpierce2: Stata module to compute trend decomposition of outcome differentials. Available from https://ideas.repec.org/c/boc/bocode/s448804.html.
- Jann, B. 2006. fairlie: Stata module to generate nonlinear decomposition of binary outcome differentials. Available from https://ideas.repec.org/c/boc/bocode/s456727.html.
- Jann, B. 2008. The Blinder–Oaxaca Decomposition for Linear Regression Models. *The Stata Journal* 8: 453–479.
- Jann, B. 2019. Influence functions for linear regression (with an application to regression adjustment). University of Bern Social Sciences Working Paper No. 32 (https://ideas.repec.org/p/bss/wpaper/32.html).
- Jann, B. 2020a. dstat: Stata module to compute summary statistics and distribution functions including standard errors and optional covariate balancing. Available from https://ideas.repec.org/c/boc/bocode/s458874.html.

References III

- Jann, B. 2020b. Influence functions continued. A framework for estimating standard errors in reweighting, matching, and regression adjustment. University of Bern Social Sciences Working Paper No. 35 (https://ideas.repec.org/p/bss/wpaper/35.html).
- Jann, B. 2021. Entropy balancing as an estimation command. University of Bern Social Sciences Working Paper No. 39 (https://ideas.repec.org/p/bss/wpaper/39.html).
- Jann, B. 2023a. cdist: Stata module for counterfactual distribution estimation and decomposition of group differences.Available from https://ideas.repec.org/c/boc/bocode/s4459187.html.
- Jann, B. 2023b. geoplot: Stata module to draw maps. Available from https://ideas.repec.org/c/boc/bocode/s459211.html.
- Jann, B. 2025. crosswalk: Stata module to recode variable based on crosswalk table (bulk recoding). Available from https://ideas.repec.org/c/boc/bocode/s459420.html.
- Juhn, C., K.M. Murphy, B. Pierce. 1991. Accounting for the Slowdown in Black-White Wage Convergence. P. 107–143 in: M. Kosters (ed.). *Workers and Their Wages*. Washington, DC: AEI Press.
- Juhn, C., K.M. Murphy, B. Pierce. 1993. Wage Inequality and the Rise in Returns to Skill. *Journal of Political Economy* 101:410–442.
- Kim, C. 2013. Detailed Wage Decompositions. Revisiting the Identification Problem. *Sociological Methodology* 43:346–363.

References IV

- Kitagawa, E.M. 1955. Components of a Difference Between Two Rates. *Journal of the American Statistical Association* 50: 1168–1194.
- Oaxaca R. 1973. Male–female wage differentials in urban labor markets. *International Economic Review* 14: 693–709.
- Rios-Avila, F. 2020. Recentered influence functions (RIFs) in Stata: RIF regression and RIF decomposition. *The Stata Journal* 20:51–94
- Yun, M. 2004. Decomposing differences in the first moment. *Economics Letters* 82:275–280.
- Yun, M. 2008. Identification problem and detailed Oaxaca decomposition: A general solution and statistical inference. *Journal of Economic and Social Measurement* 33:27–38.

Some new geoplot features

Ben Jann

University of Bern

2025 German Stata Conference
Hamburg, March 28, 2025

Some new features since last year's presentation

- Insets
- Grids and rasters
- Spatial smoothing
- More symbols
- New powerful legend options
- Direct import of ESRI and GeoJSON shape files

Data on Mexico from www.gits.igg.unam.mx/idea/descarga:

```
. geoframe create Estatal "Shapefile - Censo 2010 (Estatal).zip"
(translating Shapefile - Censo 2010 (Estatal).zip/inegi_refcenesta_2010.shp)
(importing shp file) (5 vars, 659,531 obs)
(importing dbf file) (190 vars, 32 obs)
(creating frame Estatal)
(creating frame Estatal_shp)
           Frame name: Estatal [make current]
           Frame type: attribute
         Feature type: <none>
        Number of obs: 32
              Unit ID: _ID
          Coordinates: _CX _CY
    Linked shape frame: Estatal_shp
. frame Estatal: geoframe simplify
(simplification threshold = .0000721)
(simplifying 312 shape items)
(0%...10%...20%...30%...40%...50%...60%...70%...80%...90%...100%)
(refinement threshold = .1827136)
(refining 85 shape items)
(0%...10%...20%...30%...40%...50%...60%...70%...80%...90%...100%)
(dropped 644,157 observations in frame Estatal_shp)
(added 196 observations in frame Estatal_shp)
```

Illustration of `inset()` option (can be repeated):

```
geoplot (area Estatal i._ID), nolegend ///
    inset(area world, lw(.1) color(sand) || area world if _ID==110, color(stc2) || ///
        , nobox size(40) pos(ne) title(Mexico is here) project(orthographic 1 -70) ///
           background(water lc(gray) limits(-180 180 -90 90)))
```

More data on Mexico from www.gits.igg.unam.mx/idea/descarga:

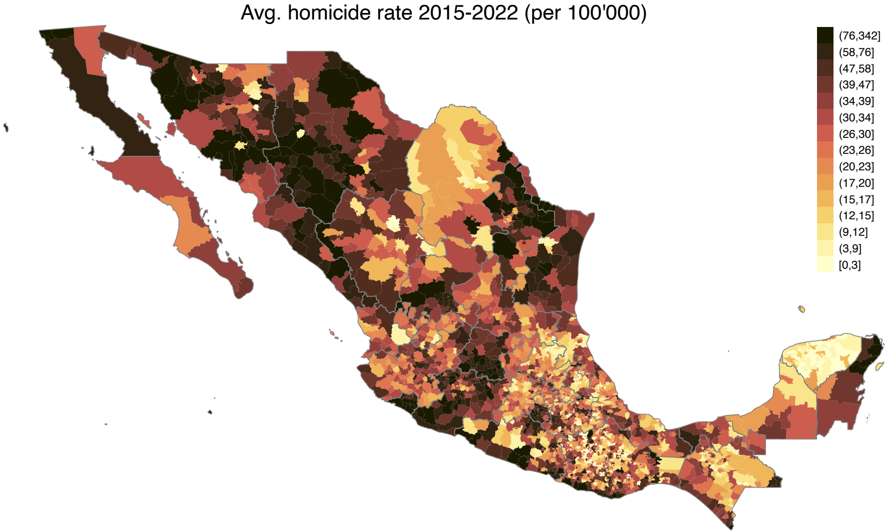
```
. geoframe create Municipal "Shapefile - Censo 2010 (Municipal).zip"
(translating Shapefile - Censo 2010 (Municipal).zip/inegi_refcenmuni_2010.shp)
(importing shp file) (5 vars, 3,283,138 obs)
(importing dbf file) (192 vars, 2,456 obs)
(creating frame Municipal)
(creating frame Municipal_shp)
            Frame name: Municipal [make current]
            Frame type: attribute
          Feature type: <none>
         Number of obs: 2,456
               Unit ID: _ID
           Coordinates: _CX _CY
    Linked shape frame: Municipal_shp
. frame Municipal: geoframe simplify
(simplification threshold = .0000721)
(simplifying 2862 shape items)
(0%...10%...20%...30%...40%...50%...60%...70%...80%...90%...100%)
(refinement threshold = .1827136)
(refining 2567 shape items)
(0%...10%...20%...30%...40%...50%...60%...70%...80%...90%...100%)
(dropped 3178096 observations in frame Municipal_shp)
(added 341 observations in frame Municipal_shp)
```

Add homicide data obtained from www.gob.mx:

```
. use Homicides, clear // (number of homicides and femicides in 2015-2022)
. frame Municipal {
.     destring cve_umun, replace
cve_umun: all characters numeric; replaced as int
.     geoframe copy default Homicides, id(cve_umun cvemunicipio)
(all units in frame Municipal matched)
(1 variable copied from frame default)
.     generate double hrate = Homicides/8 / (p_total/100000)
.     format %9.0f hrate
. }
```

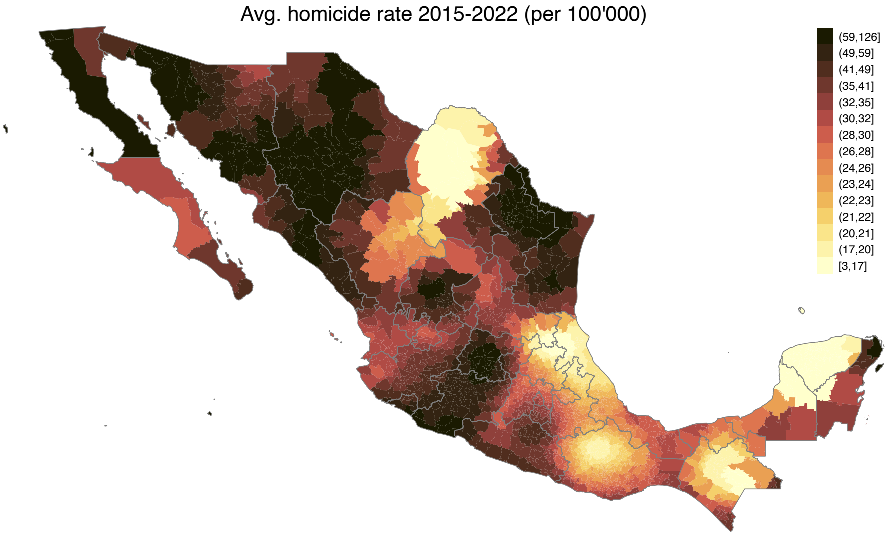
Homicide rate by municipality:

```
geoplot ///
    (area Municipal hrate, levels(15, quantile) color(scico lajolla)) ///
    (area Estatal), subtitle("Avg. homicide rate 2015-2022 (per 100'000)")
```



Apply smoothing:

```
frame Municipal: geoframe spsmooth hrate, generate(shrate)
geoplot ///
    (area Municipal shrate, levels(15, quantile) lab(, format(%9.0f)) color(scico lajolla)) ///
    (area Estatal), subtitle("Avg. homicide rate 2015-2022 (per 100'000)")
```

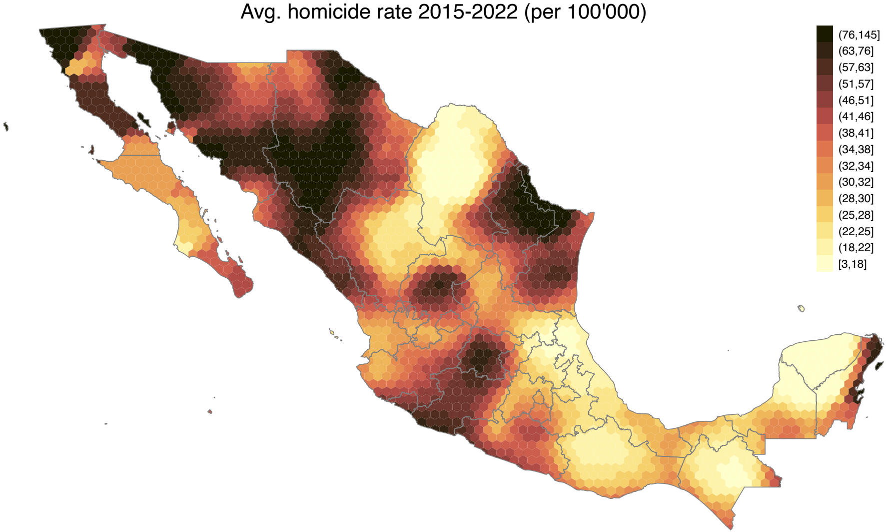


Generate raster:

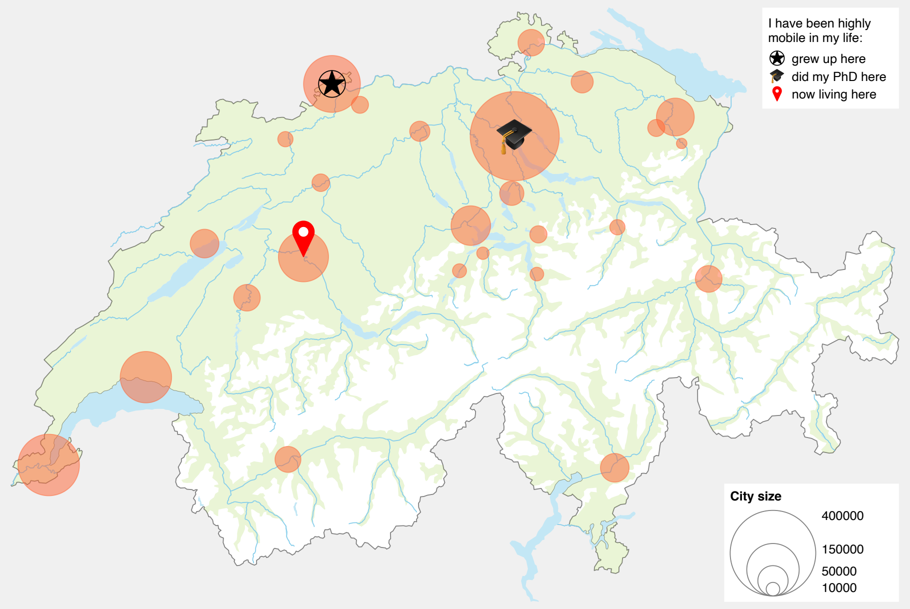
```
frame Estatal: geoframe raster R, n(100) hex
geoplot (area R i.ID, fcolor(*.5)) (area Estatal), nolegend
```

Smooth to raster:

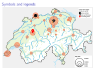
```
frame Municipal: geoframe spsmooth hrate, at(R, fill)
geoplot ///
    (area R hrate, levels(15, quantile) lab(, format(%9.0f)) color(scico lajolla)) ///
    (area Estatal), subtitle("Avg. homicide rate 2015-2022 (per 100'000)")
```



Symbols and legends



Symbols and legends



```
geoplot ///
    (area CH, if(_PLEVEL==0) fcolor(white)) ///
    (area CHvf, color(YellowGreen%20)) ///
    (area lakes) ///
    (line rivers) ///
    (symbol capitals [iw=bbtot], size(*5) color(stc6%50)) ///
    (symbol capitals (circle) if name=="Basel", size(*1.5) lcolor(black)) ///
    (symbol capitals (star)   if name=="Basel", size(*1.5) color(black)) ///
    (symbol capitals ("`=uchar(127891)'") if name=="Zürich", size(*2)) ///
    (symbol capitals (pin2) if name=="Bern", size(*2) color(red)) ///
    , bgcolor(gs15) tight ///
    slegend(1e4 5e4 15e4 4e5, overlay heading("{bf:City size}") ///
        position(se) box(color(white))) ///
    glegend(layout(- "I have been highly" "mobile in my life:" ///
        6&7 "grew up here" 8 "did my PhD here" 9 "now living here") ///
        lineskip(2.5) textwidth(17) box(color(white)))
```