

General Crosswalk Construction Framework

Jonathon Schroeder and James Gaboardi

May 2020

1 Notation formatting

- **Bold** = variable or set
- *Italic* = a single instance (= item in set)
- Non-italic = a set
- UPPER CASE = input parameter
- lower case = derived from input parameters

2 Goal specification

Generate a crosswalk $\mathbf{X}_{\mathbf{ST}}$...

- From source zones \mathbf{S} (geographic level $\mathbf{G}_{\mathbf{S}}$ in year $\mathbf{Y}_{\mathbf{S}}$)
- To target zones \mathbf{T} (geographic level $\mathbf{G}_{\mathbf{T}}$ in year $\mathbf{Y}_{\mathbf{T}}$)
- Including exactly one record per atom \mathbf{st} (an intersection between source zone \mathbf{s} and target zone \mathbf{t})
- With interpolation weights $\mathbf{w}_{\mathbf{ST}}$
 - A single weight $\mathbf{w}_{\mathbf{cst}}$ for each count variable \mathbf{c} in \mathbf{C} , for each atom \mathbf{st}
 - * $\mathbf{w}_{\mathbf{cst}}$ = proportion of \mathbf{c} in \mathbf{s} (denominator) that is also in \mathbf{st} (numerator) = $\frac{\mathbf{c}_{\mathbf{st}}}{\mathbf{c}_{\mathbf{s}}}$
 - * All \mathbf{C} are count variables (e.g., population, housing units, etc.) that have been reported for a set of sub-zones \mathbf{S}' (blocks)
- Build from an existing crosswalk $\mathbf{X}_{\mathbf{S}'\mathbf{T}'}$...
 - From source sub-zones \mathbf{S}' , which nest within \mathbf{S}
 - To source sub-zones \mathbf{T}' , which nest within \mathbf{T}
 - In our setting, we can assume:
 - * $\mathbf{G}_{\mathbf{S}'} = \mathbf{G}_{\mathbf{T}'} = \text{blocks}$
 - * $\mathbf{Y}_{\mathbf{S}'} = \mathbf{Y}_{\mathbf{S}}$ and $\mathbf{Y}_{\mathbf{T}'} = \mathbf{Y}_{\mathbf{T}}$
 - Includes weights $\mathbf{w}_{\mathbf{S}'\mathbf{T}'}$ indicating proportion of each source sub-zone's features (population & housing) in each sub-zone atom $\mathbf{s}'\mathbf{t}'$.
- Include every \mathbf{s} in \mathbf{S} and every \mathbf{t} in \mathbf{T} .
 - Atom records may have *null* \mathbf{s} where \mathbf{t} where a zone in one set lies beyond the spatial extent of the other set, or the intersection is outside the extent of $\mathbf{X}_{\mathbf{S}'\mathbf{T}'}$.
 - * In our case, $\mathbf{X}_{\mathbf{S}'\mathbf{T}'}$ is a block-to-block crosswalk based on NHGIS shapefiles, which are clipped at the coast. The 1990-2010 crosswalk omits “off-coast” 1990 blocks that are not in the shapefile. For crosswalks with 1990 source zones, we may have *null* \mathbf{t} for source zones that lie entirely off-coast.

3 Summary of key input parameters

- G_S = source geographic level
- Y_S = source year
- G_T = target geographic level
- Y_T = target year
- C = set of count variables for which to derive separate weights

4 General steps

1. Obtain & load sub-zone crosswalk (blocks-to-blocks) $X_{S'T'}$.
2. Obtain & load data for source sub-zone counts (source-year block data) $C_{S'}$.
 - (a) Include any identifiers needed to associate S' with S .
3. Join base crosswalk $X_{S'T'}$ to source sub-zone data $C_{S'}$ on S' identifiers.
 - (a) Use a “left join” to ensure that all sub-zone atoms are included, even those without a matching record in the sub-zone data file (especially important for 1990 blocks).
4. For each sub-zone atom $s't'$, identify encompassing zones s and t :
 - (a) If possible, derive S and T identifiers from S' and T' identifiers (e.g., tract ID is in block ID).
 - (b) Else if possible, derive S identifiers from source sub-zone data from step 2.
 - (c) Else, obtain identifiers through other means...
 - i. 1990 block-group parts require some [special handling](#) because neither 4a nor 4b pertain to all BGPs.
 - ii. If we generate crosswalks for target zones that cannot be identified from block IDs (e.g., places, county subdivisions, etc.), we'll need to add a step to join block crosswalk to target-year block data that includes identifiers for the target zones.
 - (d) Where s' is *null* (= “”), omit these dummy sub-zone atoms from subsequent computations.
 - i. This may drop some valid t from the computations, but step 9 will re-add them if needed.
5. Compute counts for all weighting variables in each sub-zone atom: $c_{S'T'} = w_{S'T'} * C_{S'}$.
6. Compute counts for all weighting variables in each atom of interest: $c_{ST} = \sum c_{S'T'}$ group by S, T .
 - (a) Steps 5 & 6 can be combined into single formula by substituting $w_{S'T'} * C_{S'}$ for $c_{S'T'}$ in step 6.
7. Compute counts for all weighting variables in each source zone: $c_S = \sum c_{ST}$ group by S .
8. Compute all weights for all atoms of interest: $w_{CST} = \frac{c_{st}}{c_s}$.
 - (a) If $c_s = 0$, set $w_{cst} = 0$.
9. If w_{CST} is missing data for any s in S or t in T , add dummy atoms with *null* t for non-*null* s or *null* s for non-*null* t , and set $w_{cst} = 0$.
10. Export clean, complete file for distribution.
 - (a) Exact specifications TBD.