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 $\boldsymbol{X}_{\mathbf{ST}}$ ...

- From source zones $\mathbf{S}$ (geographic level $\boldsymbol{G}_{\mathbf{S}}$ in year $\boldsymbol{Y}_{\mathbf{S}}$)
- To target zones $\mathbf{T}$ (geographic level $\boldsymbol{G}_{\mathbf{T}}$ in year $\boldsymbol{Y}_{\mathbf{T}}$)
- Including exactly one record per atom $\boldsymbol{st}$ (an intersection between source zone $\boldsymbol{s}$ and target zone $\boldsymbol{t}$)
- With interpolation weights $\boldsymbol{w}_{\mathbf{ST}}$
  - A single weight $\boldsymbol{w}_{\mathbf{cst}}$ for each count variable $\boldsymbol{c}$ in $\mathbf{C}$, for each atom $\boldsymbol{st}$
    - $\boldsymbol{w}_{\mathbf{cst}}$ = proportion of $\boldsymbol{c}$ in $\boldsymbol{s}$ (denominator) that is also in $\boldsymbol{st}$ (numerator) = $\dfrac{\boldsymbol{c}_{\boldsymbol{st}}}{\boldsymbol{c}_{\boldsymbol{s}}}$
    - 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 $\boldsymbol{X}_{\mathbf{S}'\mathbf{T}'}$ ...
  - 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:
    - $\boldsymbol{G}_{\mathbf{S}'} = \boldsymbol{G}_{\mathbf{T}'}$ = blocks
    - $\boldsymbol{Y}_{\mathbf{S}'} = \boldsymbol{Y}_{\mathbf{S}}$ and $\boldsymbol{Y}_{\mathbf{T}'} = \boldsymbol{Y}_{\mathbf{T}}$
  - Includes weights $\boldsymbol{w}_{\mathbf{S}'\mathbf{T}'}$ indicating proportion of each source sub-zone's features (population & housing) in each sub-zone atom $\boldsymbol{s't'}$.
- Include every $\boldsymbol{s}$ in $\mathbf{S}$ and every $\boldsymbol{t}$ in $\mathbf{T}$.
  - Atom records may have *null* $\boldsymbol{s}$ where $\boldsymbol{t}$ where a zone in one set lies beyond the spatial extent of the other set, or the intersection is outside the extent of $\boldsymbol{X}_{\mathbf{S}'\mathbf{T}'}$.
    - In our case, $\boldsymbol{X}_{\mathbf{S}'\mathbf{T}'}$ 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* $\boldsymbol{t}$ for source zones that lie entirely off-coast.

3 Summary of key input parameters

- $\boldsymbol{G}_{\mathbf{S}}$ = source geographic level
- $\boldsymbol{Y}_{\mathbf{S}}$ = source year
- $\boldsymbol{G}_{\mathbf{T}}$ = target geographic level
- $\boldsymbol{Y}_{\mathbf{T}}$ = target year
- $\mathbf{C}$ = set of count variables for which to derive separate weights

4 General steps

1. Obtain & load sub-zone crosswalk (blocks-to-blocks) $\boldsymbol{X}_{\mathbf{S'T'}}$.
2. Obtain & load data for source sub-zone counts (source-year block data) $\boldsymbol{C}_{\mathbf{S'}}$.
   (a) Include any identifiers needed to associate $\mathbf{S'}$ with $\mathbf{S}$.
3. Join base crosswalk $\boldsymbol{X}_{\mathbf{S'T'}}$ to source sub-zone data $\boldsymbol{C}_{\mathbf{S'}}$ on $\mathbf{S'}$ 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 $\boldsymbol{s't'}$, identify encompassing zones $\boldsymbol{s}$ and $\boldsymbol{t}$:
   (a) If possible, derive $\mathbf{S}$ and $\mathbf{T}$ identifiers from $\mathbf{S'}$ and $\mathbf{T'}$ identifiers (e.g., tract ID is in block ID).
   (b) Else if possible, derive $\mathbf{S}$ 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 $\boldsymbol{s'}$ is *null* (= ""), omit these dummy sub-zone atoms from subsequent computations.
      i. This may drop some valid $\boldsymbol{t}$ from the computations, but step 9 will re-add them if needed.
5. Compute counts for all weighting variables in each sub-zone atom: $\mathbf{c}_{\mathbf{S'T'}} = \mathbf{w}_{\mathbf{S'T'}} * \mathbf{C}_{\mathbf{S'}}$.
6. Compute counts for all weighting variables in each atom of interest: $\mathbf{c}_{\mathbf{ST}} = \sum \mathbf{c}_{\mathbf{S'T'}}$ group by $\mathbf{S}$, $\mathbf{T}$.
   (a) Steps 5 & 6 can be combined into single formula by substituting $\mathbf{w}_{\mathbf{S'T'}} * \mathbf{C}_{\mathbf{S'}}$ for $\mathbf{c}_{\mathbf{S'T'}}$ in step 6.
7. Compute counts for all weighting variables in each source zone: $\mathbf{c}_{\mathbf{S}} = \sum \mathbf{c}_{\mathbf{ST}}$ group by $\mathbf{S}$.
8. Compute all weights for all atoms of interest: $\mathbf{w}_{\mathbf{CST}} = \dfrac{\mathbf{c}_{\mathbf{st}}}{\mathbf{c}_{\mathbf{s}}}$.
   (a) Where $\boldsymbol{c_s} = 0$, set all $\boldsymbol{w_{cst}} = 0$.
9. If $\mathbf{w}_{\mathbf{CST}}$ is missing data for any $\boldsymbol{s}$ in $\mathbf{S}$ or $\boldsymbol{t}$ in $\mathbf{T}$, add dummy atoms with *null* $\boldsymbol{t}$ for non-*null* $\boldsymbol{s}$ or *null* $\boldsymbol{s}$ for non-*null* $\boldsymbol{t}$, and for these atoms set all $\boldsymbol{w_{cst}} = 0$.
   (a) As in step 4, it may not be possible to identify all $\boldsymbol{s}$ in $\mathbf{S}$ or $\boldsymbol{t}$ in $\mathbf{T}$ from the base sub-zone crosswalk.
      i. e.g., for 1990 BGPs, obtain complete identifiers from 1990 STF1 BGP-level data.
10. Export clean, complete file for distribution.
   (a) Exact specifications TBD.