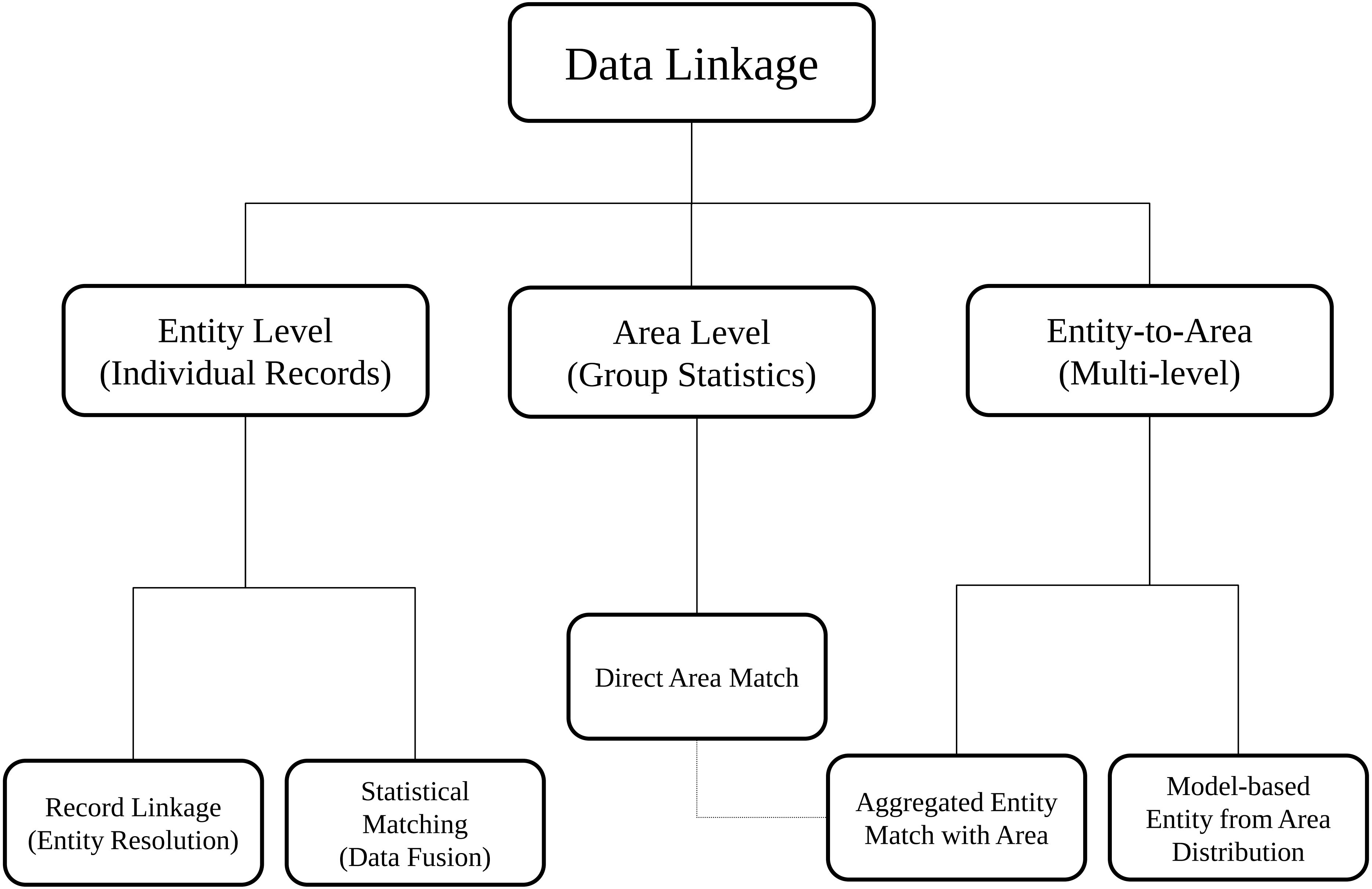
Combining Data by Statistical Matching, Imputation and Modeling

Purpose for combining data

- Improve coverage
    - Survey data from different frames (e.g. landline and cell phone)
- Increase sample size
    - Meta analysis
    - Combining probability sample with nonprobabilty sample (improves coverage as well)
- Bring together variables from different files
    - Neighborhood Air quality measurements


Data Linkage

Entity Level (Individual Records)

Area Level (Group Statistics)

Entity-to-Area (Multi-level)

Record Linkage (Entity Resolution)

Statistical Matching (Data Fusion)

Direct Area Match

Aggregated Entity Match with Area

Model-based Entity from Area Distribution

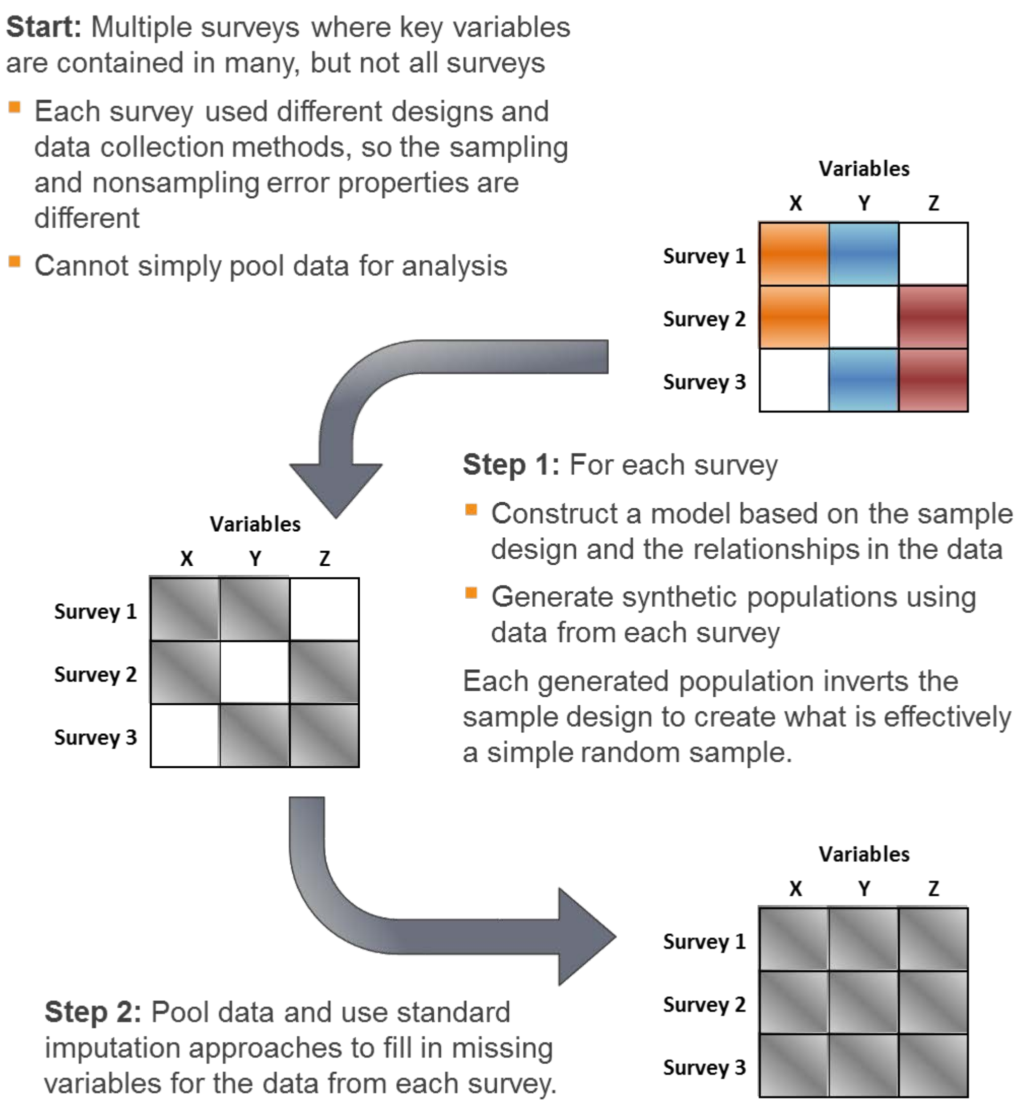
Statistical Matching

- Record’s measurements are at the same level
- Little-to-no overlap of records across samples

	Y_1	Y_2	…	Y_q	X_1	X_2	…	X_p	Z_1	Z_2	…	Z_r
Sample 1	y_{111}	y_{112}	…	y_{11q}	x_{111}	x_{121}	…	x_{11p}				
	y_{121}	y_{122}	…	y_{12q}	x_{121}	x_{122}	…	x_{12p}				
	⋮				⋮							
	$y_{1n_1 1}$	$y_{1n_1 2}$	…	$y_{1n_1 q}$	$x_{1n_1 1}$	$x_{1n_1 2}$	…	$x_{1n_1 p}$				
Sample 2					x_{211}	x_{212}	…	x_{21p}	z_{211}	z_{212}	…	z_{21r}
					x_{221}	x_{222}	…	x_{22p}	z_{221}	z_{222}	…	z_{22r}
					⋮				⋮			
					$x_{2n_2 1}$	$x_{2n_2 2}$	…	$x_{2n_2 p}$	$z_{2n_2 1}$	$z_{2n_2 2}$	…	$z_{2n_2 r}$

Combining Multiple Complex Surveys

Elliot, M.R. (2011), “Statistical Analysis Using Combined Data Sources: Discussion,” 2011 JPSM Distinguished Lecture



Start: Multiple surveys where key variables are contained in many, but not all surveys

- Each survey used different designs and data collection methods, so the sampling and nonsampling error properties are different
- Cannot simply pool data for analysis

	Variables		
	X	Y	Z
Survey 1	🟧	🟦	
Survey 2	🟧		🟥
Survey 3		🟦	🟥

Step 1: For each survey

- Construct a model based on the sample design and the relationships in the data
- Generate synthetic populations using data from each survey

Each generated population inverts the sample design to create what is effectively a simple random sample.

	Variables		
	X	Y	Z
Survey 1	▨	▨	
Survey 2	▨		▨
Survey 3		▨	▨

Step 2: Pool data and use standard imputation approaches to fill in missing variables for the data from each survey.

	Variables		
	X	Y	Z
Survey 1	▨	▨	▨
Survey 2	▨	▨	▨
Survey 3	▨	▨	▨