Methodological Considerations for Disaggregation

Panel 2: Counting the Uncounted

Global Preparatory Seminar for the
United Nations World Data Forum

September 7, 2016

My perspective

- Academic demographer
- Work on questions in demography and epidemiology in Africa
- Collaborate closely with academic statisticians developing new statistical methods for population and health data and research

Disaggregation

- Quick review of why disaggregation is hard
- The number of unique 'cells' in a dataset is the product of the number of effective categories defined by each variable:

$$\text{cells} = \textit{categories}_{v1} \times \textit{categories}_{v2} \times \cdots \times \textit{categories}_{vn}$$

- Adding either new variables or new categories for existing variables can *greatly* increases the number of cells
- To produce useful population-level measures, *each cell* must have a reasonable number of observations
- Consequently, the number of observations required is (very) large and must include a wide variety of subjects

Possible ways to disaggregate

There are at least four general approaches to address the challenges of disaggregation

1. **Collect lots of data** that includes many subjects in each cell
   - This is logistically difficult and complex
   - *Very expensive*
   - Produces measures rather than estimates
2. **Smoothing and interpolation**
   - Can incorporate data from many different sources to infer reasonable values for cells with missing data
   - Can operate on many dimensions of the data simultaneously and account for uncertainty well
   - Produces estimated values, not raw measures

Possible ways to disaggregate

3. **Use models** to generate/estimate values
   - These incorporate independent, external knowledge of the processes generating the data
   - The resulting estimates are a hybrid of data and our understanding of how the data are generated
4. **Borrow data from a similar setting**
   - When direct measures are not possible, use what we know about a similar setting
   - The result is a contextualized version of information from elsewhere
5. **All of the above**

Thoughts on the way forward

1. **Short term: use what we have better**
   - Utilize all alternative approaches to disaggregation that do not require large amounts of new data
   - Key requirements: more people trained in required methods and better availability of existing data
2. **Medium term: selectively invest in collecting new data** where it is most effective in improving estimates
   - Key requirements: understanding the importance/influence of different data and data sources and focused investment in the most useful of those
3. **Long term: collect a lot more data**

Example: child mortality in Tanzania

- **Goal:** national and small-area estimates of child mortality (U5MR) for past several decades using as much of available data as possible
- **Data:**
  1. All demographic and health surveys for Tanzania: DHS - household sample surveys
  2. Two demographic and health surveillance system sites: HDSS - intensive surveillance of small, geographically limited populations
     - These are very different data sources with completely different designs

Example: child mortality in Tanzania

- **Methods:**
  - Survival analysis of child mortality in all possible times and places using all data sources accounting for data design - sampling, etc.
  - Space-time smoothing model to integrate/interpolate/smooth all estimates so they are consistent with one another
- **Results:**
  - Consistent time trends in child mortality at both national and subnational levels
  - Consistent uncertainty/precision in all estimates
  - Possible to disaggregate further, e.g. by age and socioeconomic status, within the same framework, possibly using models and limited information from similar populations