Spatial Statistical Methods

Paul Voss
Carolina Population Center
Odum Institute for Research in Social Science
University of North Carolina, Chapel Hill

Santa Barbara Specialist Meeting:
“Future Directions in Spatial Demography”
December 12-13, 2011
“I’ve tried them all”

Probably not!
Huge body of “stuff”

• Much of what needs to be said has already been said
Huge body of “stuff”

• Much of what needs to be said has already been said
  – Fischer & Getis, 2010
  • 600+ pp.
  • Seven major sections
  • 35 chapters
Huge body of “stuff”

• Much of what needs to be said has already been said
  – Fischer & Getis, 2010
  – Anselin, 2011
    • Highly personal
      & focused account
    • Richly documented
Huge body of “stuff”

- Much of what needs to be said has already been said
  - Fischer & Getis, 2010
  - Anselin, 2011
  - de Smith, Goodchild & Longley (v. 3.15, 2011)

- Visualization examples are wonderful
- Coverage encyclopedic e.g., GIS Software: 188 products
Huge body of “stuff”

- Much of what needs to be said has already been said
  - Fischer & Getis, 2010
  - Anselin, 2011
  - de Smith, Goodchild & Longley (v. 3.15, 2011)
  - Journals
    - Many dozens
So... what to do(?)
Focus on just one small topic
Small-area population estimates
Two areas where most (applied) demographers need to learn from their statistical colleagues:

- Producing small-area population estimates
- Using small-area population estimates
Prefatory comments…

• I’m going to be critical, but it’s largely self-criticism; I spent the majority of my early career doing precisely what I here criticize

• Define “small area”
  – …areas with populations for which reliable estimates simply cannot be produced due to limitations of the available data (Jiang & Lahiri, 2006)
  – these need not always refer to geographic regions; “small-domain” is a better term, referring to estimates of attributes for some demographic group (spatial or not)
Claim 1: Most demographers who make small-area population estimates are woefully behind the state-of-the-art

- Most population estimates are generated using “models” that were introduced 30-50 years ago
  - estimation systems are mostly accounting devices; non-stochastic & non-spatial; interest is in point estimation; little concern for reliability
- The relatively large literature addressing statistical models for small-area population estimation is, as a factual matter, almost completely ignored
  - standard mixed effects models & Bayesian hierarchical models
Perhaps it’s okay?

- Most such demographers have little formal training in demography or statistics.
- Most population estimation systems are designed as large-scale production engines; not much incentive or capacity to annually produce hundreds of estimates using sophisticated truly model-based methodologies; roll-ups are straightforward.
- Consumers of the estimates don’t much care. They want point estimates and don’t wish to be bothered by considerations of uncertainty.
- Tests of simple estimation systems generally reveal that they produce tolerably good point estimates.
- Additional evidence reveals that spatial niceities don’t much improve such estimates; viewed largely as impractical academic exercises.
Perhaps not okay?

• A great deal of public money is allocated each year based on such estimates; shouldn’t they be as good as they possibly can be?

• A large statistical literature presents alternative, much better ways of producing small-area population estimates; why continue to ignore this?

• What happens if, say, a state demography office or an independent demographic consultant is sued over estimates that are not produced by the best possible methodologies? Not a pretty picture

• Consumers should demand better
Claim 2: (Specifically regarding the American Community Survey) it appears that most of us would rather complain about the estimates than figure out how to extract better information from them

• For most small geographic areas, ACS estimates have unacceptable, intolerable MOEs

• There exist established statistical methodologies of “borrowing strength” across space and time to adjust ACS estimates to useful estimates that enable monitoring change over time or assessing a more realistic extent of spatial heterogeneity

• These can be fully spatial-temporal methodologies

• But the work is not easy; high price of admission
What are these methodologies?

• Actually there are many
  – “Synthetic estimates” combining direct (sample-based) estimates with regression model-based estimates (e.g., Census Bureau’s SAIPE estimates for counties)
  – Various mixed-effects models
  – Complex spatial Bayesian approaches (e.g., BYM model in which small-area variation not explained by covariates is generally expressed as a spatially unstructured random effects and spatially correlated random effects

• How do we learn about this?
  – Use your web browser; the literature is large
  – Carl Schmertmann
  – New node in NCRN network (Univ. of Missouri) “Improving the Interpretability and Usability of the ACS through Hierarchical Multiscale Spatio-Temporal Statistical Models”
Some examples from ACS…

Cities in NC; poverty rate for children <5 in MC families
Poverty Rate for Children under Age 5
Living in Married Couple Families, 2009

Source: 2009 American Community Survey, Table B17006
Poverty Rate for Children under Age 5 Living in Married Couple Families, 2009

Source: 2009 American Community Survey, Table B17006

90% Confidence Interval
Temporal estimates particularly troublesome

Example: City of Fayetteville
Child poverty estimates from 1-year ACS samples, 2005 to 2009
Poverty Rate for Children under Age 5 Living in Married Couple Families, Fayetteville NC, 2005 to 2009

Source: 2009 American Community Survey, Table B17006
Poverty Rate for Children under Age 5 Living in Married Couple Families, Fayetteville NC, 2005 to 2009

Source: 2009 American Community Survey, Table B17006

90% Confidence Interval
So, the ACS estimates are…

• **Noisy!**
  – small(ish) samples are common
  – margins of error are large
  – year-to-year blips
  – occasional odd or unbelievable estimates
  – goal: increase the signal/noise ratio

• **ACS estimates involving income are temporally complex**
  – overlapping time periods for estimates
  – multiple reference periods for a single question (e.g., “income in past 12 months”) within a sample
INCOME IN THE PAST 12 MONTHS

Mark (X) the "Yes" box for each type of income this person received, and give your best estimate of the TOTAL AMOUNT during the PAST 12 MONTHS. (NOTE: The "past 12 months" is the period from today’s date one year ago up through today.)

Mark (X) the "No" box to show types of income NOT received.

If net income was a loss, mark the "Loss" box to the right of the dollar amount.

For income received jointly, report the appropriate share for each person – or, if that’s not possible, report the whole amount for only one person and mark the "No" box for the other person.

a. Wages, salary, commissions, bonuses, or tips from all jobs. Report amount before
So, for example, in terms of income (poverty) reporting…

- **2010 ACS estimates are based on 12 monthly samples taken Jan10 to Dec10**
- But, for example, the poverty estimates are based on retrospectively reported income covering the period 12 months prior to the survey
- **There are 12 overlapping periods for the “2010” income (poverty) data involving income reports covering 23 months:**
  - “Jan10” survey covers income Jan09 to Dec 09
  - “Feb10” survey covers income Feb09 to Jan10
  - etc.
Temporal complexity...

Chart adapted from presentation by Carl Schmertmann, FSU
Imagine monthly "true" rates:

\[ \theta_1, \theta_2, \ldots, \theta_{83} \]

Jan04, Feb04, ..., Nov10

The 2010 ACS produces an estimate of:

\[
Y_{2010} = \frac{1}{12} (\theta_1 + 2\theta_2 + \ldots + 12\theta_{12} + \ldots + 2\theta_{22} + \theta_{23})
\]

\[
= \sum_{j=i}^{23} c_{j,2010} \theta_j
\]
Therefore…

\[ Y_{2005} = \sum_{j=i}^{23} c_{j,2005} \theta_j \]  

Includes monthly income from Jan04 through Nov05

\[ Y_{2010} = \sum_{j=i}^{23} c_{j,2010} \theta_j \]  

Includes monthly income from Jan09 through Nov10

“True” averages over time

\[ Y = \begin{bmatrix} C & \theta \end{bmatrix} \]

\( (6\times1) \quad (6\times83) \quad (83\times1) \)

ACS averages over time

\[ \hat{Y} = C \theta + \epsilon \]

\( (6\times1) \quad (6\times83) \quad (83\times1) \quad (6\times1) \)

Independent sampling errors with known variances
ACS Likelihood (θ | estimates)

With normal errors ε,

\[
\ln L(\theta | \hat{Y}) = -\frac{1}{2} \sum_{i=1}^{6} \left( \frac{\hat{Y}_i - c_i^t \theta}{\sigma_i} \right)^2 + k
\]

83 parameters and 6 observations

Bayesian priors for \( \theta_1, \ldots, \theta_{83} \)

- Wiggly month-to-month patterns less likely than smooth patterns
- We probably can assign a range for Prior(θ)
Poverty Rate for Children under Age 5 Living in Married Couple Families, Fayetteville NC, 2005 to 2009

Source: 2009 American Community Survey, Table B17006

Very unlikely
Extensions…

• Apply jointly to multidimensional time series (e.g., child poverty and unemployment rate)

• Restructure Bayesian priors to borrow strength not only across time, but also across space & space-time when smoothing the ACS data (requires the 5-year ACS estimates)

• Thanks!