Future Directions in Spatial Demography

ANDREW A. BEVERIDGE
Institute for Demographic Research
Queen’s College, CUNY
Email: andrew.beveridge@qc.cuny.edu

Although spatial thinking and the use of new forms of geospatial data have grown rapidly in the social sciences, their implementation in the demographic sciences has lagged. This statement in the call for potential participants for this specialist workshop is particularly telling. In short, though there have been major advances and research commitments to the development of spatial data resources that can be used to address important issues in demography and health research, as well as both the development and dissemination of spatial statistical and modeling techniques that can used to analyze spatial data, such data and methods are only slowly permeating demography and related fields.

It seems to me that there are several approaches and opportunities for applying a spatial dimension to demographic work that have the potential to help in the formulation and solution of various demographic problems. Such approaches some of which are analytically contradictory, in fact share a great deal. In the cases below, with which I have had some experience, improved data resources and improved modeling approaches would certainly be of great value.

Analyses of Data from Individuals in their Geo-Demographic Context
Demographers and other social scientists now routinely add data from the neighborhood, community or other context to surveys or other information on individual respondents or events to both large scale (e.g., Census, data on hospital admission, mortality, etc.) and smaller scale (e.g., surveys or health, education or other activities) data sets. Analyses can then proceed to relate the neighborhood or community data to the information regarding the individual respondent. The default method for such analysis, of course, is hierarchical or mixed modeling, where information about the higher level units (e.g., neighborhood and community) is used to condition the modeling of the individual level information. These methods though they do link individual levels to spatial units, generally do not make explicit use of spatial information (e.g., the spatial relationship of contextual units one to another.) In effect, the information related to the contextual unit is reduced to a series of measures of that unit, but not of its location. So, if there is a spatial effect, (for instance, the effect of poverty is exacerbated by it being spatially concentrated) remains unmeasured. Standardized methods to integrate spatial information into

---

1 The general approach I am discussion can also use aggregated data at more than one level. For instance, census tract or block group characteristics can nest within county or place, however defined.
such modeling, along with characteristics of the so-called higher level units, is particularly important for the impact of spatial effects not to be overlooked.\textsuperscript{2}

**The Use of Spatial Location to Impute Social and Economic Status (SES) Characteristics**

An issue related to, but seemingly in contradiction to the first issue, is the use of residential or other location to impute SES measures, which are then used for other modeling. Generally, when this is done, for reasons of endogeneity, it is not appropriate to use spatial variables in the models. Instead, the spatial term is used, in effect, to impute SES characteristics. Such methods are particularly useful in situations where it is infeasible to directly collect SES data. Two examples with which I have had experience include using such methods to add some information on SES to educational records and hospital discharge data. Obviously, in such an instance, it would be especially useful to be able to estimate the extent to which such an imputation was precise. This precision certainly can have a spatial distribution. Having geocoded the Early Childhood Longitudinal Study from Kindergarten (ECLS-K) for a number of waves, the National Center for Education Statistics invited me to analyze the extent to which using residential location was a reasonable proxy for parents’ education, occupation and economic status for the National Assessment of Educational Progress (NAEP). NAEP is a huge study of the performance of 4\textsuperscript{th}, 8\textsuperscript{th} and 12\textsuperscript{th} graders, along with a longitudinal study of a sample of 9, 13 and 17 year olds. For the major study, which includes testing and a student instrument, it is not feasible to collect from children information about their parents. However, to stratify the results beyond race and Hispanic status, free or reduced lunch status, and general geographic location, some sort of SES variable was sought. Since ECLS-K has a parent survey, it was used to assess the potential of using information about the residential location (most generally the student’s address is available from administrative records) to create proxies for economic and social status of the parents. Having more information on how such approaches works in practice and where such proxies provide useful estimates of SES and where they do not, would be very useful. Then it would be feasible to add such information to a wide range of health, education and other types of records where address is routinely collected.\textsuperscript{3} To understand the spatial distribution of the precision of such methods based upon the variables chosen for imputation and the type of location would be very useful.

\textsuperscript{2} I should note that the recently released SAS 9.3 does appear to include the ability to integrate spatial terms explicitly into hierarchical or mixed models, as well as enhancing a large number of Bayesian approaches. Some of these techniques have been developed in some procedures in R, including the recently developed spBayes procedure. Such methods are not yet mainstream. However, to expect more rapid dissemination of such techniques into demography, they need to be presented in a platform that people are familiar with and that is robust, so that the models can be reliably computed. The current statistical package of choice for newly trained population and health scientists seems to be STATA, if our five recent hires of demographers and health researchers at Queens are any indication.

\textsuperscript{3} With a collaborator I have geocoded nearly 100 million hospital discharge records in New York State to assess various health outcomes and relate them to neighborhood context. I have used similar methods to impute the race and Hispanic status of prospective jurors and voters. Though perhaps somewhat utilitarian use of spatial information, it is nonetheless extremely important for many issues.
Adding or Recovering Land Use and Facilities Data for Use in Population Science and Health

Currently it is relatively easy to get spatially referenced data from the United States that has very well developed information on the residential characteristics of individual and households living in various census geographies. It is true that the advent of the American Community Survey means that larger samples at the typical census tract or other level now will apply to a longer time horizon of five years, but this will be offset by more frequent releases of data.\(^4\) Though one knows from the typical Census tabulation of aggregate data reporting, who lives in certain areas, there is little or no information regarding the other characteristics of the area, especially when the area is the size of a census tract or even smaller. It is, of course, possible to know where the streets and roads, water features, and some landmarks are located from the standard Census map.

Though collected by the Census Bureau, there is no integrated set of data that relates population characteristics to the existence of business activity broadly defined (the Bureau collects business information including non-profits and public entities, as well as prisons and jails, while the NCES collects data on school and their location), as well as other features of land use. Obviously for the development and change of neighborhoods and communities, as well as the health and well-being of residents, the location of business and other facilities are very important, as are development of transportation, power, water and other infrastructure.

Having these sorts of data readily available would mean that large scale studies of population, population change and health could go beyond simply relating one demographic variable measured at some level to another. It would make it possible to bring to bear information regarding other features. There are a large number of studies that have brought in the location of specific facilities or other information (e.g., fast-food, pollution, transportation networks, business activity), to understand aspects of population distribution and change, as well as health status. If such data were regularly integrated with census and other demographic and health data, relationship that now can only be modeled in one specific location, could examined for much larger areas, perhaps for the United States as a whole.

\(^4\) I should note that the ACS is somewhat misunderstood by many, even demographers. In my role as a demographic consultant for the New York Times and consultant both formally or informally to others, it is obvious to me that few realize that the ACS is now the source for data that used to be collected by the Census long form survey. See my column for a discussion: “Under a New Name Census Data Stands Ready for Perusal.” [http://www.gothamgazette.com/article/demographics/20110811/5/3582](http://www.gothamgazette.com/article/demographics/20110811/5/3582). There are also issues regarding the methods that the ACS used to compute confidence intervals. My exchange with Census Bureau on this topic is available here [http://www.scribd.com/doc/61741043/Memo-Regarding-ACS-With-Response](http://www.scribd.com/doc/61741043/Memo-Regarding-ACS-With-Response). The recent designation of the areas requiring language assistance in voting as required by law was based upon the ACS. This designation was done using model based (i.e., Bayesian) estimating procedures. See various documents at [http://www.census.gov/rdo/data/voting_rights_determination_file.html](http://www.census.gov/rdo/data/voting_rights_determination_file.html) (i.e., Bayesian), thus obviating the problems that I outlined. Though using such estimates to do this designation, at this writing the Census Bureau has not made public any model based estimates for small areas using such estimating procedures.
There are a wide variety of other issues, including: pervasive dissemination and use of population data, the acquiring and use of both micro and aggregated data in a form similar to that available in the US for the rest of the world, as well as others. Plainly there is much to be done to move spatial data, techniques and modeling into a more central role in demography and health research.