Popul Res Policy Rev (2007) 26:601–618
DOI 10.1007/s11113-007-9046-5

Geographic Information Systems and Spatial Data Processing in Demography: a Review

Michael Reibel

Published online: 6 September 2007

Abstract This paper reviews the use of geographic information systems (GIS) software for spatial data processing in demography. The review begins with an introduction to GIS. Next, it traces the three major types of spatial data problems confronting demographers: the geocoding and geoprocessing of microdata, estimation of detailed population surfaces, and combining data aggregated to incompatible zone systems. GIS and non-GIS solutions to these problems are contrasted, with examples from published research. Spatially pre-processed datasets available to demographers are then discussed. The author concludes by noting that the solutions GIS provides to previously intractable data problems in spatial demography might encourage a focus on dynamic processes of population change in local areas.

Keywords Areal interpolation · Geocoding · GIS · Small area demography · Spatial data processing

Introduction

Many of the important theoretical problems in demography and nearly all the methodological approaches in applied demography are inherently spatial. Even when locations are not themselves units of analysis, or key characteristics of units of analysis, ecological context as a vector of causes is critical to demographic phenomena as diverse as residential mobility and teenage pregnancy. Spatial data is thus essential to the investigation of many demographic and particularly social demographic phenomena.

M. Reibel (✉)
Department of Geography and Anthropology, California State University, 3801 W. Temple Place, Pomona, CA 91768, USA
e-mail: mreibel@csupomona.edu

Demographers and allied researchers in geography, urban planning, real estate, public health, and other related disciplines therefore use spatial data regularly. For the most part, however, this does not require explicitly spatial data processing because the data used take the form of spatially aggregated social, housing, and demographic data from the U.S. Census Bureau and other secondary sources. In other words, spatial data processing, in the sense of attaching geographic codes to data and reassigning, reaggregating or disaggregating data between geographic schemes, has been performed by the data provider. Further data processing and preparation by investigators may involve the reaggregation of data to larger areas, but these steps are typically performed using the existing set of geocodes and are thus no different procedurally than nonspatial recodes and aggregations using existing data values.

There are three major situations in demography that typically require spatial data processing by investigators: geocoding and geoprocessing of microdata, estimating detailed population surfaces, and combining data aggregated to incompatible zone systems (we shall see that the last two are often, but not always, linked in practice). In the first situation, individual data points such as persons, households, crime scenes, or buildings must be *georeferenced*, i.e., assigned location coordinates such as latitude/longitude, so that they can be digitally mapped. Georeferencing is also the first step in the process of *geocoding*, i.e., automatically assigning geocodes (codes for voter precincts, census tracts, fire department response areas, etc.) based on the coordinates of the data point's location. Detailed population surface estimates seek to code, or map, the true geographic variation in population or population density, rather than the aggregated populations of a set of geographic zones. This is often a necessary task for risk analysis and emergency management planning, e.g., to predict the local impact of flooding. It is also a very common intermediate processing step in combining data aggregated to incompatible zone systems. Incompatible aggregation problems are very common in spatial and applied demography; examples are the need to combine tract or block group level data across two decennial censuses in order to compute trends, and the need (very common in market research) to combine census tract level data with zip code data or zip coded customer microdata.

Solutions to these problems have evolved over time as researchers have applied richer data sources and more powerful analytical tools. Notably, progress in this area has accelerated in recent years due to advances in geographic information systems (GIS) computer applications and the increasing availability of GIS compatible data. Accordingly, this review will begin with a brief general discussion of GIS as it applies to demographic and related social research.

The review will continue by addressing each of the three major spatial data processing problems in demography, highlighting relevant methodological approaches in past population studies, and work in related social sciences as well as contemporary GIS approaches to spatial data processing. The article is intended for demographers who are not GIS specialists. It specifically addresses practical considerations for solving data processing problems in basic and applied research, and particular attention is given to the trade-offs between the accuracy of various techniques and their ease of use. It should be noted this review is restricted to a

discussion of spatial data processing and does not address the many recent advances, primarily by geographers, in applying spatial statistics analytically to demography. This review also is restricted to the U.S. literature, with the exception of the treatment of areal interpolation research. Much progress in areal interpolation has been achieved by British geographers, and whether British or American, areal interpolation studies tend to appear in international journals.

Geographic Information Systems

Geographic information systems (GIS) are not a new type of computer application. The first commercial GIS software products date to the late 1960s. In essence, GIS are nothing more or less than relational database management systems (DBMS) in which records are *georeferenced*, meaning associated with a particular geographic location, and can thus be subjected to spatial queries based on their location. In practice all contemporary GIS software packages have far more spatial data processing and spatial analysis capabilities. In recent years, greater data and file format compatibility, relatively user friendly interfaces, and vastly greater processing power have combined to make GIS useful for specific purposes (such as small area demographic data processing) by computer-literate persons with little or no formal training, following a reasonable learning curve.

The two principle data architectures employed by the various GIS applications are *raster* and *vector* architectures (see Fig. 1). In raster mode, a GIS treats a continuous surface as a fine-grained grid composed of pixel cells, or rasters. Objects are georeferenced to a single raster (in the case of point locations), a sequence of rasters (line objects such as roads or streams), or a set of rasters corresponding to an area. The raster resolution is user-defined, and can be rectified to coincide with satellite image resolutions and coverages (e.g., the 30×30 m resolution of LANDSAT images). Raster analysis is therefore ideal for the analysis of continuous phenomena such as elevation, rainfall, and land cover. A vector GIS builds objects by connecting point locations using vectors (because they account for the curvature of the earth, these objects are strictly speaking not vectors but arcs—hence ArcGIS,

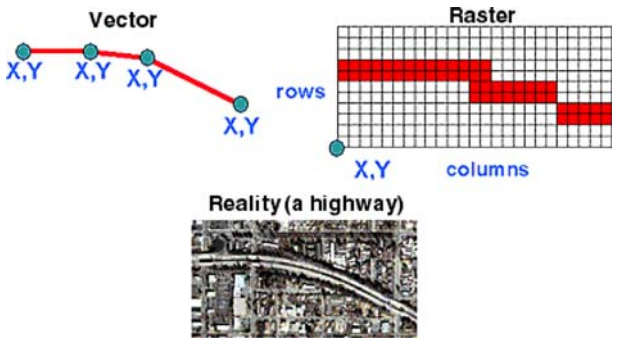

Fig. 1 Raster and vector data and representations in GIS

the name of the largest selling GIS application). A vector GIS builds geographic features by linking sequences of vectors to create line objects, and by closing line objects (making their final end points identical to their initial starting points) to enclose areas in polygons. Despite the need for complex topological definitions in the data structure, vector GIS are more parsimonious than raster structures for working with zone systems such as census geography. Complex GIS data processing and analysis often require the combination of data in raster and vector formats, and most GIS packages have tools for converting geographic information between the two formats.

There are two major advantages of GIS over other DBMS. First, multiple sets of georeferenced records called layers, often including linked information stored in external tables, can be combined in a given study area for joint geoprocessing and analysis, a process called map overlay (Longley et al. 2005; McHarg 1969). Second, georeferenced objects can be subjected to spatial queries and geoprocessing, both within a single layer or across multiple layers. Spatial queries are performed in GIS using spatial algorithms that permit the user to select features in one layer based on their topological relationship to other features in the same or other data layers (inside, outside, intersecting, tangent, etc.), or by creating buffers of given radius around objects and selecting records based on topological relationship to the buffer objects. For example, one could subset all the houses that are more than half a mile from any road, or a selected subset of roads (see Fig. 2).

Geoprocessing tools in GIS include, but are not limited to, topological operations such as selection, performing spatial intersections and merges, and other operations

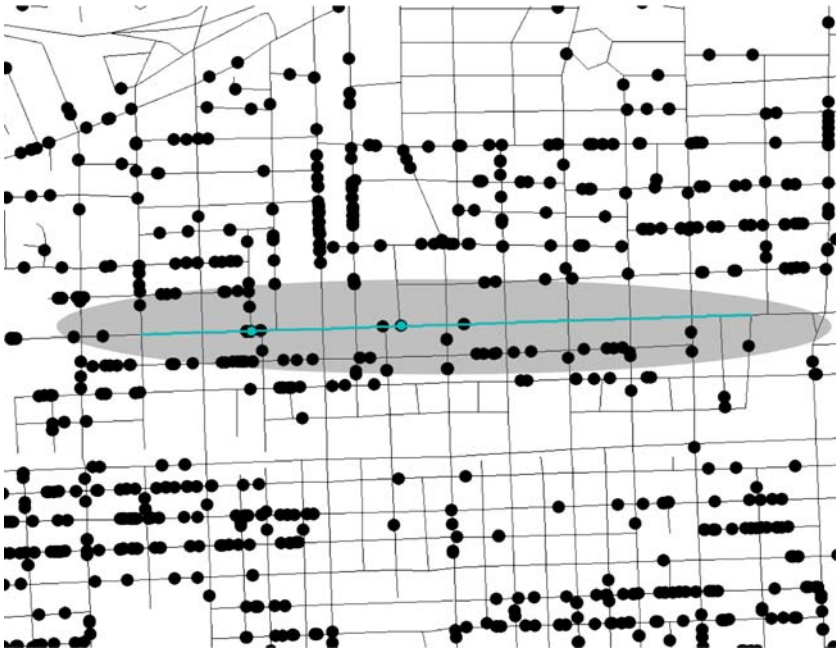

Fig. 2 Spatial selection in GIS by buffer

on georeferenced objects. A typical simple example would be to overlay a map of California county boundaries on a statewide (California) map of elementary schools, select Orange County in the county layer, then subset and extract all Orange County schools by topological reference to the superimposed object in the other (county) layer. In most GIS selection can be done through SQL queries or by point-and-click operations in a map window. In some GIS, point-and-click can also be used to instantly identify and open records. This provides a powerful tool for exploratory data analysis via analytical cartography. For instance, a region's zones could be mapped in a color ramp representing the magnitude of some attribute, such as population density. Outliers can be visually identified and their records opened with a single click.

Geocoding and Geoprocessing of Microdata

The process of placing a single demographic data point (person or household) in spatial context is referred to broadly as geocoding. Geocoding is possible without a GIS, but it is extremely labor intensive: because of the need to process individual records one at a time, GIS analysts sometimes refer to such attempts at geoprocessing of data without automated GIS tools as "manual" computations, even when done with a computer. In a GIS, the process is composed of several steps. First, the observation must be georeferenced, that is, assigned latitude and longitude coordinates (see Table 1). Typically, this automated operation is performed on microdata records containing street addresses. The addresses are matched against available look-up tables of street segments, each of which contains a field for street address ranges, to find the street segment on which each address is located. The address is then georeferenced to a point on that street segment. In more sophisticated automatic street georeferencing, addresses may be georeferenced to one or the other side of the street based on odd/even addresses, and can be placed relatively accurately along the block via linear interpolation of the street address within that street segment's address range.

Following this, the observation is geocoded in the strict sense. This means that overlay geoprocessing is used to superimpose one or more zone systems such as census tract and/or zip code geography on the layer containing observations' point locations, and zone codes are attached to the observations corresponding to the zone in each superimposed system within which the observation is located. Once an observation is geocoded, it is a simple matter requiring no further geoprocessing to assign contextual covariates, including covariates from multiple aggregation scales

Table 1 A table of georeferenced data (schools) with coordinates in decimal degrees

Type	Student population	Testing percentile	Address	Zip	Lat (N)	Lon (W)
HS	1733	51	1849 Maple Street	53727	40.2337	92.0241
E	457	87	787 Oak Street	53711	40.2419	92.0188
E	404	56	22 Beech Avenue	53708	40.2362	92.0167

to which the observations have been geocoded, to observations by simply match-merging the observation data with the data containing contextual covariates based on their geocodes.

The largest geocoder and geoprocessor of demographic microdata is, of course, the U.S. Census Bureau itself. The Bureau maintains a vast collection of geographic information and processing tools called the Topologically Integrated Geographic Encoding and Referencing system, or TIGER. The TIGER system includes TIGER/Line boundary files (georeferenced polygon files corresponding to census geography areas—states, counties, etc., down to tracts and blocks). In addition, TIGER/Line digital geographic layers include centerline data. These are georeferenced line-object files corresponding to the nation's streets and roads. These centerline data contain the address look-up tables that facilitate street address geocoding. All these products are available either directly or indirectly via the Census Bureau and all can be extremely useful for demographic practitioners using GIS.

In terms of research applications, a number of microdata studies of neighborhood change use address codes attached to individual observations to assign corresponding tract codes and match tract level contextual data to the observations; an example is Elliott et al. (1985). The fastest growing application of geocoding for purposes of modeling neighborhood contextual effects and covariates is not, however, in social demography per se but rather in public health research. Beginning in the 1990s, research methods in public health began to incorporate spatial demography in order to investigate such inherently spatial but less traditional public health outcomes such as public safety and teenage reproductive behavior in inner cities (Sucoff and Upchurch 1998; Chen et al. 1998; O'Campo et al. 1997; Krieger 1992). In addition, many specialists in the social aspects of public health have expanded this distinctly geographic turn in their thinking by emphasizing neighborhood scale social inequalities and their consequences for health and wellness overall.

This theoretical and methodological turn toward spatial demography in public health research has received high profile institutional support from Nancy Krieger and her colleagues in the Public Health Disparities Geocoding Project at the Harvard School of Public Health. In addition to conducting a great deal of relevant research (Subramanian et al. 2005; Krieger et al. 2002, 2003a, b), the Geocoding Project team has received funding to train public health and allied investigators in spatial demographic practice including geocoding techniques using GIS.

Population Surface Mapping and Areal Interpolation of Population Surfaces

Another important application of GIS in demography is the generation of highly detailed population maps. Such maps are often called population surface maps because they come close to treating a region's population realistically as a surface of continuous variation. In this sense population surface maps differ markedly from chloropleth population maps, the more familiar thematic maps of population calculated for a set of predefined political or aggregation zones within each of which potentially very large differences in population density are implicitly averaged. Aside from their value for data display and exploratory data analysis, in certain

analytical situations population surface maps provide a partial solution to the modifiable areal unit problem, the property of aggregated spatial data by which a given population surface in a region can generate very different local area count distributions depending on how the aggregation zone boundaries are drawn (Openshaw 1983). Openshaw further showed that the modifiable areal unit problem greatly complicates the analysis of spatially aggregated data and raises potential questions of validity for certain types of analysis (Openshaw 1984). The correction for this type of aggregation bias provided by population surface mapping is only partial because surface estimation ultimately derives counts from a fixed set of aggregation zones.

The most common approach to population surface mapping is dasymetric mapping, which predates GIS and even electronic computers (cf. Wright 1936). In dasymetric mapping, a detailed proxy data layer assumed to be linked to population counts or densities by a known or derivable function is used to generate the population surface from spatially aggregated data. In current practice, dasymetric mapping is typically accomplished in a raster GIS environment using remotely sensed data (satellite images) mapped as grid cells of high resolution, such as the 30×30 m resolution corresponding to Landsat images, or even smaller. Remotely sensed images are classified, meaning that land cover type for each grid cell is inferred from its (typically multispectral) image color values, and the classification is validated. Weights for each land cover type are derived, and the weights are applied to generate a fine-grained grid population surface (for a comparison of a population surface map versus a chloropleth, i.e., tract-area map of the same population and region, see Figs. 3 and 4 in Reibel and Agrawal, 2007).

Some population surface maps derive estimates directly from the proxy data surface. Examples of this type of research include Pozzi et al. (2003), who modeled the earth's population surface using remotely sensed nighttime light imagery. Qiu et al. (2003) used land use change detection and expansion of the system of roads as proxy population measures to model urban expansion. Radeloff et al. (2001) used land cover and land ownership data to model residential penetration as part of a forest management study. Ryznar and Wagner (2001), Ward et al. (2000), and Dragicevic and Marceau (1999) used land cover data to model urban growth and change.

In dasymetric maps, the population surface is derived from a set of aggregation zones with known counts. A detailed proxy data layer is superimposed on the source zones and provides information regarding unobserved count density variations within the source zones. This information is used to derive a set of weights for assigning fractional counts from the source zones to the smaller areas pertaining to the ancillary data layer, often by regressing the set of zone counts for the region on the vector of proxy information variables expected to indicate the local presence of population (Flowerdew and Green 1989, 1992; Langford et al. 1991). The counts interpolated to the more fine-grained geography of the ancillary layer form the approximate population surface (Reibel and Agrawal 2005, 2007; Mennis 2003; Eicher and Brewer 2001; Goodchild et al. 1993; Monmonier and Schnell 1984). This type of areal interpolation is discussed below in the somewhat different context of reconciling data aggregated to incompatible zone systems.

Another approach to population surface generation via areal interpolation of aggregated source zone counts, but one that is geometric rather than dasymetric, is areal interpolation by smoothing. This approach was pioneered by Tobler (1979) and substantially modified by Martin (1989) and Bracken and Martin (1989). In the smoothing approach, source zones are converted to rasters and the source zone counts are interpolated as density gradients along raster paths between each source tract centroid (geometric center point) and the centroids of adjacent tracts. The density gradients are scaled in such a way that the magnitude of the initial source zone count is preserved for the set of counts pertaining to the source zone's associated surface, a property Tobler dubbed "pycnophylactic."

Smoothing techniques for areal interpolation build on simpler interpolation techniques in geographic data processing such as the interpolation of surfaces from data pertaining to a set of sample point locations, a technique long used by geographers to produce contour maps. Like these earlier techniques, they take advantage of the ubiquitous spatial autocorrelation of data to make relatively accurate estimates by assuming an uninterrupted surface and correlation subject to a regular distance decay function similar to the gravity models used by migration demographers.

The major drawbacks to the adoption of smoothing techniques for areal interpolation by demographers are, first, that aggregation zones such as census tracts are not drawn randomly on an uninterrupted social landscape with smooth population density gradients, but rather are designed to use real and abrupt barriers in the landscape, such as elevated highways, as their boundaries. The tendency of such abrupt changes in urban geography and corresponding socioeconomic landscapes to coincide with aggregation zone boundaries introduces error in models that assume continuous data gradients. More of a drawback for many demographers, smoothing requires a high level of spatial statistics and geoprocessing skills.

Approaches to Spatially Mismatched Aggregate Data in Neighborhood Research

We might choose Duncan and Duncan (1957) and Taeuber and Taeuber (1965) as a reasonable starting point for contemporary methodological approaches to neighborhood change research. While neither book explicitly discusses their approach to mismatched tract geographies, Duncan and Duncan (p. 320) indicate that the restrictions on their analysis allowed them to avoid including any tracts that changed boundaries during the relevant intervals. Taeuber and Taeuber generally followed the Duncans' methodology in such matters. Among later studies that could not avoid mismatched tracts entirely, most studies to date attempt to deal with the problem by reaggregation, that is, by combining split and/or merged zones into larger units that are compatible with both zone systems. Compatibility in this context means that the larger units, where necessary, consist of territories made up of one or more whole zones in both zone systems, so that counts pertaining to both sets of original zones can be accurately computed for the custom geography by summation.

Examples of sociodemographic and economic geography studies using the reaggregation approach include Coulson and Leichenko (2004), Lobo et al. (2002), Reibel (2000), Clark (1996), Alba et al. (1995), Denton and Massey (1991), Lee and Wood (1991), Massey and Mullen (1984), and Massey (1983). While most investigators using reaggregation go to considerable lengths to reconcile tract boundary discrepancies, some simply omit from their analyses local areas in which census tract boundaries differ significantly between the two zone systems (Freeman and Rohe 2000; Lee 1985). This practice is ill-advised, particularly for the computation of local demographic trends. The parts of a region of interest (such as a metropolitan area) that experience the most sociodemographic and economic change are typically those that experience the most boundary change. Omitting such local areas from analysis will understate the social changes behind the geographic zone changes, leading to likely bias.

Another group of studies uses U.S. Census public use microdata (PUMS) to investigate metropolitan level changes over time (Odland and Ellis 2001; Myers 1999, Wright et al. 1997). Because PUMS data are geocoded to custom Public Use Microdata Areas (PUMAs) that sometimes span county boundaries, it is relatively painstaking to develop a protocol to geographically recode PUMS data from successive enumerations to a consistent metropolitan geography (see Ellis et al. 1999 for a discussion of the methodology).

Many studies of neighborhood change are not explicit about their methodology for dealing with spatially incompatible data (Fong and Shibuya 2003; Fong and Gulia 2000; Lauria and Baxter 1999; Galster 1990, 1998; Temkin and Rohe 1998; Galster et al. 1997; Smith 1991; Gober 1986). The lack of information on zone matching methodology in these papers is unfortunate, but it does not necessarily imply that adequate or even sophisticated approaches were not used. Rather, the omission may well reflect the low priority given to discussion of data geoprocessing by some in population studies.

Without GIS, reaggregation approaches to spatially incompatible data problems are labor-intensive. Investigators must execute each aggregation operation separately for the data file pertaining to each original zone system, and individually for each set of population and subpopulation counts in every local area requiring aggregation. In practice, this means hours of painstaking spreadsheet manipulations and/or hundreds of lines of computer code. Moreover, there is no reason to expect that data aggregated to two (or more) zone systems will always conform to the rule that for each local area, each zone in every system can be expressed as a simple split or merge with respect to zones in the other system(s). Indeed, in the typical case of U.S. Census tract boundary changes between decennial enumerations, individual tracts are often both split and merged.

The designation of census tracts is complicated by the participation of local planning authorities, which sometimes tend to act conservatively to maintain existing boundaries or to maintain (or avoid) correspondences between some local areas as tracts change over time. But simultaneous tract splits-and-merges are, to some degree, a rational way of dealing with common patterns of growth and decline within metropolitan systems. For example, expansion at the urban fringe is often noncontiguous due to gaps in road access, topography, etc. A small part of a

previously lightly inhabited tract becomes developed during the interval, and that dense and compact newly developed area is split off as a separate tract. The balance of the parent tract, meanwhile, does not have enough population, community cohesion, or other criteria to stand alone as a new tract. It is distinct from the new development, however, and has more in common ecologically with an adjacent remaining undeveloped area outside the parent tract. This adjacent undeveloped area, moreover, may well have lost part of its area and population to a different development, also spun off as a new tract. The logical solution is to combine the undeveloped parts of the two rump tracts into a new third tract (see Fig. 3). Similar patterns of complex boundary changes can be observed in both declining and revitalizing older urban areas.

In sum, both the theoretically possible and the actual zone geographies confronting investigators attempting to cope with data aggregated to incompatible zone systems are complex in ways that can frustrate and defeat even the most conscientious efforts to enforce compatibility through reaggregation. Simultaneous split-and-merge situations as described above or equally complex superimpositions in the case of other zone system comparisons are quite common when spatially incompatible data must be combined. Such boundary changes require far more radical reaggregation to resolve than do simple merges and splits. The degree of reaggregation necessary to enforce compatibility is often such that a significant number of the resulting custom-made zones are at a completely different ecological scale than was intended. Reaggregating mismatched zones while leaving matched zones in their original state thus inflates and distorts the range of zone sizes in the desired study region. This range can quickly become so large that a single scale of analysis can no longer meaningfully be asserted.

One solution to the problem of complex tract boundary restructuring in the U.S. Census is to construct source period counts for target tracts by aggregating the counts of source period blocks to the target period tract geography. This approach

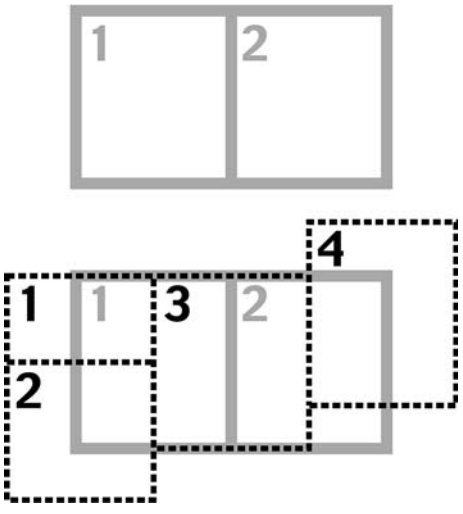

Fig. 3 Example of complex tract geography changes between decennial enumerations: Top, 1990 tracts; bottom 2000 tracts superimposed on 1990 tracts

was notably used by Allen and Turner (2002). It has the distinct advantage of a more fine-grained surface of source zones from which to assemble the target territories, a property that tends to minimize error (Sadahiro 2000). Inasmuch as block aggregation uses observed counts it is very accurate. Moreover, the reaggregation of source block counts to target tract zones can be automated with GIS. To assemble benchmark counts to error-test their street-weighted interpolated tract estimates, Reibel and Bufalino (2005) applied point-in-polygon aggregation to simultaneously aggregate whole blocks and assign split blocks resulting from the map layering to their respective target tracts. Their approach involved creating a new layer of point data (dimensionless locations each coded by a single pair of latitude/longitude coordinates) corresponding to the centroids (geometric center points) of the block areas. By attaching the block counts to their respective centroids in the new layer, Reibel and Bufalino were able, via geoprocessing in GIS, to assign block counts automatically to the target tracts in which the blocks' centroids are located.

Areal Interpolation of Aggregate Population Data

As an approach to reconciling and combining spatially mismatched data, areal interpolation is fundamentally different from the techniques described above because it does not rely on reaggregation, with its consequent loss of detail, to remove uncertainty. Areal interpolation therefore does not confront the dilemma described earlier, in which analysts attempting reaggregation in areas with complex intersections between zones (e.g., census tracts simultaneously splitting and merging) must either create enormous zones to enforce compatibility or drop local areas from the analysis. Rather, areal interpolation techniques seek to minimize estimation error for data interpolated to defined zones of any size. Areal interpolation of spatially mismatched data also has the advantage that it normally uses as a set of target zones a real zone system, such as the census tract geography pertaining to a given decade. This obviates the need for an ambiguously synthetic zone system with great variation in zone size.

Areal interpolation may be defined as the transfer of local zone counts or other attribute magnitudes to other zones under conditions of uncertainty, using defined spatial algorithms. We have seen that when known counts from an original set of zones (the source zones) are interpolated to much smaller zones, or to a grid surface, the result is a population surface map. One additional step is needed to transform an interpolated population surface to a set of target zone counts. By reaggregating the small zones or grid cells on a population surface map and their interpolated count estimates back to a higher scale zone system, one can estimate populations of target zones under conditions of uncertainty. The typical example in demography is interpolating census tract counts from decade zero (e.g., 1990) to the corresponding tract geography for the same region in decade one (e.g., 2000), or vice versa, in order to compute trends by subtraction.

An example of an areal interpolation approach to combining spatially mismatched aggregate data that does not require a population surface map is area weighting (Goodchild and Lam 1980). In area weighting, source and target zone

systems are overlain and intersected. Source zone counts are fractionally assigned to their corresponding intersection zones based on the proportion of the source zone's area contained within the intersection zone. The intersection zones are then reaggregated to the target zone geography, and each target zone's fractional source counts are summed to yield the estimated source variable counts of the target zone units. Sociodemographic studies using data subjected to area weighted interpolation include Reibel (2003) and Vandell (1981).

The implicit assumption in area weighting is that population always varies directly with area; in other words, that there are no internal population density variations within any of the source tracts. This is clearly a dubious assumption. Large variations in population density with tracts are common, and they lead to relatively high rates of estimation error when area weighting is applied. On the other hand, area weighting, like all types of areal interpolation, can be automated in a GIS, and unlike reaggregation approaches to spatially mismatched data, all types of areal interpolation also preserve the scale and geographic precision of the intended study area.

To avoid the errors associated with the assumption of uniform densities within source zones, progress in the areal interpolation of spatially mismatched aggregate data has focused on dasymetric techniques incorporating proxy data layers to more accurately incorporate information about underlying population density when interpolating counts. One approach to combining spatially mismatched aggregate data that requires only the intersection of source and target zones, rather than the generation of a continuous population surface, is street weighting (Reibel and Bufalino 2005). In street weighted areal interpolation, the street and road grid is superimposed on the intersection zones of the source and target tracts, and source zone populations are fractionally assigned to intersection zones based on intersection zones' proportion of the aggregate length of their respective source zones' street and road grid. Finally, as in area weighting, the intersection zone counts are summed across their respective target zones to derive population estimates.

Other dasymetric approaches to combining spatially mismatched aggregate data involve the use of a third set of zones, called control zones, which are relatively homogeneous with respect to population density and which serve as an intermediate set of target zones to increase accuracy in estimation (Reibel and Agrawal 2005; Goodchild et al. 1993; Monmonier and Schnell 1984). In this approach areal interpolation to target zones is a two-step process: first from source zones to homogeneous control zones, then from control zones to target zones. A much more common approach for interpolating data between spatially mismatched zone systems is dasymetric mapping in a raster environment, using remotely sensed urban land cover data as a proxy measure of population. This is typically the type of data and processing approach used when, as described above, a raster (grid) population surface map created by dasymetric areal interpolation of source counts is reaggregated to a target zone geography to derive estimated populations of those target zones, e.g., incompatible census tracts. Land cover-weighted areal interpolation of data between spatially mismatched zone systems has proven to be relatively accurate (Reibel and Agrawal 2007; Cockings et al. 1997; Fisher and Langford 1995, 1996; Langford and Unwin 1994).

Special Geocoded and Specially Pre-matched Datasets

A final issue relevant to our discussion of spatial data processing in demography is the existence of special data sets in which geocoding or spatially mismatched data issues have been resolved. The tract coded Panel Study of Income Dynamics (PSID) microdata is a good example. Public release versions of the PSID include geocodes for region, state of residence, size of largest city in the county of residence, and the Beale rural-urban code, but no geocodes at the census tract or similar scale that would situate the microdata observations in neighborhood context. Special tract-coded extracts are available for the PSID, but the conditions are quite restrictive and the process takes months. Prospective users must submit a research plan, document a human subjects review clearance/waiver, negotiate a contract with a confidentiality agreement, and pay a fee. Investigators who have successfully navigated these requirements have used the data to produce excellent work on neighborhood effects in social demography and public health (Crowder 2000; Foster and McLanahan 1996; Harris 1999; Massey et al. 1994; Quillian 1999; South and Crowder 1997a, b, 1998a, b, 1999; Sucoff and Upchurch 1998).

Another noteworthy special dataset is the Neighborhood Change Data Base (NCDB), a commercial data product released by GeoLytics in association with the Urban Institute. The NCDB is a nationwide (U.S.) dataset that contains census population and subpopulation counts for 1970, 1980, 1990, and 2000, all transferred to the 2000 census tract geography. The methodology used is highly complex, as it necessarily changed from one intercensal matching process to the next (see Tatian 2002, Appendix J). In general, however, we can assume that the older the data transferred to 2000 tract boundaries, the less accurate the estimates become. This is partly because the earlier decades must be transferred forward one decade at a time, and each successive transfer introduces error. It is also because small area census geography less exhaustively covered the nation's territory before the 1990 census. The transfer of the 1990 data to 2000 boundaries, at least, is extremely accurate: for this inter-censal transfer the NCDB aggregated 1990 blocks and used a version of street weighting to resolve 1990 blocks split by 2000 tracts—an excellent, if complicated, combination of aggregation and interpolation techniques that capitalizes on the advantages of both in terms of precision. For the 1980 to 1990 and 1970 to 1980 intercensal transfers, the creators had to rely at times on area weighted interpolation where aggregation from smaller units was not possible.

The NCDB evolved from the Urban Institute's Underclass Database (UDB). The UDB was not widely used by researchers; one example is Ellen (2000). The NCDB has been available only since 2002. This is not long enough to become widely known as a data source for research published in refereed articles, although the NCDB was used by Krol and Svarny (2005). The NCDB holds great promise for research on neighborhood change over multiple decades. Assuming the project will continue past 2010 and beyond, the series will grow not only in length but in quality relative to the early decades because presumably the future will continue to bring accurate and exhaustive digital geographic data layers for processing. Even if cautious investigators restrict themselves to the 1990 to 2000 trend data, the

resource is worth considering as it frees investigators examining sociodemographic trends from the task of processing spatially mismatched data.

Conclusion

The author has described GIS solutions to common problems in the processing of spatial population and socioeconomic data. Automated GIS techniques for microdata geoprocessing, and population surface mapping and block reaggregation for combining spatially incompatible tract data are highly accurate as well as far less labor-intensive than older approaches. GIS automated areal interpolation of spatially mismatched tract data across decades is much easier than reaggregation. Moreover, any properly executed areal interpolation, even the relatively crude area weighting technique, will better preserve the scale and exhaustiveness of the data being processed than will reaggregation.

Subsequently the review focused on available data sets offering special geocoded microdata and tract data from a series of decennial enumerations pre-matched to the 2000 tract geography. Unlike census data, neither data set is available free to institutionally affiliated researchers. The potential value of these data sets for investigators studying the social consequences of neighborhood change is great, however, and the cost (and difficulty, in the case of the PSID data) to such analysts of accessing these data sets seems eminently worthwhile.

It seems clear that, in particular, dynamic small area demography—that is, research on neighborhood change performed by population and related specialists—has been limited in the past by the sheer difficulty of processing data spatially aggregated to tracts (and other zone systems) that change over time. One consequence of this is that there are 10 or 20 articles on static neighborhood segregation patterns for every article on neighborhood transitions or succession. The reluctance to examine neighborhood transitions directly, doubtless due to data processing difficulties, has led to indirect forms of dynamic urban analysis in which increases or decreases over time in static segregation indices are substituted, and sometimes confused for, measures of neighborhood transition (see Alba et al. 1995 for a discussion of this type of error). Recent advances in spatial data processing using GIS hold out the hope that the direct analysis of neighborhood change and contextual effects will once again become common in demography and related sciences.

References

Alba, R. D., Denton, N. A., Leung, S. J., & Logan, J. R. (1995). Neighborhood change under conditions of mass immigration: The New York City Region 1970–1990. *International Migration Review, 29*, 625–656.

Allen, J. P., & Turner, E. (2002). *Changing faces, changing places: Mapping southern Californians*. Northridge, CA: The Center for Geographical Studies.

Bracken, I., & Martin, D. (1989). The generation of spatial population distributions from census centroid data. *Environment and Planning A, 21*, 537–543.

Chen, F. M., Brieman, R. F., Farley, M., Plikaytis, B., Deaver, K., & Cetron, M. S. (1998). Geocoding and linking data from population-based surveillance and the US census to evaluate the impact of median household income on the epidemiology of invasive Streptococcus pneumoniae infections. *American Journal of Epidemiology*, *148*, 1212–1218.

Clark, W. A. V. (1996). Residential patterns: Avoidance, assimilation and succession. In R. Waldinger, & M. Bozorgmehr (Eds.), *Ethnic Los Angeles* (pp. 109–138). New York: Russell Sage Foundation.

Cockings, S., Fisher P., & Langford, M. (1997). Parameterization and visualization of the errors in areal interpolation. *Geographical Analysis*, *29*, 314–328.

Coulson, N. E., & Leichenko, R. M. (2004). Historic preservation and neighborhood change. *Urban Studies, 41*, 1587–1600.

Crowder, K. (2000). The racial context of white mobility: An individual-level assessment of the white flight hypothesis. *Social Science Research, 29*, 223–257.

Denton, N. A., & Massey, D. S. (1991). Patterns of neighborhood transition in a multi-ethnic world: U.S. metropolitan areas, 1970–1980. *Demography, 28*, 41–63.

Dragicevic, S., & Marceau, D. J. (1999). Spatio-temporal interpolation and fuzzy logic for GIS simulation of rural-to-urban transition. *Cartography and Geographic Information Science*, *26*, 125–137.

Duncan, O. D., & Duncan, B. (1957). *The Negro population of Chicago*. Chicago, IL: University of Chicago Press.

Eicher, C., & Brewer, C. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, *28*, 125–138.

Ellen, I. G. (2000). Race-based neighborhood projection: A proposed framework for understanding new data on racial integration. *Urban Studies, 37*, 1513–1533.

Elliott, D. S., Quinn, M. A., & Mendelson, R. E. (1985). Maintenance behavior of large-scale landlords and theories of neighborhood succession. *AREUEA Journal, 13*, 424–445.

Ellis, M., Reibel, M., & Wright, R. (1999). Comparative metropolitan area analysis: Matching the 1980 and 1990 Census Public Use Microdata Samples for metropolitan areas. *Urban Geography, 20*, 75–92.

Fisher, P., & Langford, M. (1995). Modeling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A, 27*, 211–224.

Fisher, P., & Langford, M. (1996). Modeling sensitivity to accuracy in classified imagery. *Professional Geographer*, *48*, 299–309.

Flowerdew, R., & Green, M. (1989). Statistical methods for inference between incompatible zone systems. In M. Goodchild, & S. Gopal (Eds.), *The accuracy of spatial databases* (pp. 239–247). London, England: Taylor and Francis.

Flowerdew, R., & Green, M. (1992). Developments in areal interpolation methods and GIS. *Annals of Regional Science, 26*, 67–78.

Fong, E., & Gulia, M. (2000). Neighborhood changed within the Canadian ethnic mosaic, 1986–1991. *Population Research and Policy Review, 19*, 155–177.

Fong, E., & Shibuya, K. (2003). Economic changes in Canadian neighborhoods. *Population Research and Policy Review, 22*, 147–170.

Foster, E. M., & McLanahan, S. (1996). An illustration of the use of instrumental variables: Do neighborhood conditions affect a young person's chance of finishing high school? *Psychological Methods*, *1*, 249–260.

Freeman, L., & Rohe, W. (2000). Subsidized housing and neighborhood racial transition: An empirical investigation. *Housing Policy Debate, 11*, 67–89.

Galster, G. (1990). White flight from racially integrated neighborhoods in the 1970s: The Cleveland experience. *Urban Studies, 27*, 385–399.

Galster G. (1998). A stock/flow model of defining racially integrated neighborhoods. *Journal of Urban Affairs, 20*, 43–51.

Galster, G., Mincy, R., & Tobin, M. (1997). The disparate racial neighborhood impacts of metropolitan economic restructuring. *Urban Affairs Review, 32*, 797–824.

Gober, P. (1986). How and why Phoenix households changed: 1970–1980. *Annals of the Association of American Geographers, 76*, 536–549.

Goodchild, M., Anselin, L., & Deichmann, U. (1993). A framework for the aerial interpolation of socioeconomic data. *Environment and Planning A, 25*, 383–397.

Goodchild, M., & Lam, N. (1980). Aerial interpolation: A variant of the traditional spatial problem. *Geo-Processing, 1*, 297–312.

Harris, D. R. (1999). "Property values drop when blacks move in, because...": Racial and socioeconomic determinants of neighborhood desirability. *American Sociological Review, 64*, 461–479.

Krieger, N. (1992). Overcoming the absence of socioeconomic data in medical records: Validation and application of a census-based methodology. *American Journal of Public Health*, *82*, 703–710.

Krieger, N., Chen, J. T., Waterman, P. D., Soobader, M. J., Subramanian, S. V., & Carson, R. (2002). Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: Does the choice of area-based measure and geographic level matter? The Public Health Disparities Geocoding Project. *American Journal of Epidemiology, 156*, 471–482.

Krieger, N., Chen, J. T., Waterman, P. D., Soobader, M. J., Subramanian, S. V., & Carson, R. (2003a). Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). *Journal of Epidemiology and Community Health, 57*, 186–199.

Krieger, N., Waterman, P. D., Chen, J. T., Soobader, M. J., & Subramanian, S. (2003b). Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violence: Geocoding and choice of area-based socioeconomic measures—The Public Health Disparities Geocoding Project (US). *Public Health Reports, 118*, 240–260.

Krol, R., & Svarny, S. (2005). The effect of rent control on commute times. *Journal of Urban Economics, 58*, 421–436.

Langford, M., Maguire, D., & Unwin, D. (1991). The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In I. Masser, & M. Blakemore (Eds.), *Handling geographic information: Methodology and potential applications* (pp. 55–77). Harlow, Essex: Longman.

Langford, M., & Unwin, D. (1994). Generating and mapping population density surfaces within a geographical information system. *Cartographic Journal, 31*, 21–26.

Lauria, M., & Baxter, V. (1999). Residential mortgage foreclosure and racial transition in New Orleans. *Urban Affairs Review, 34*, 757–786.

Lee, B. A. (1985). Racially mixed neighborhoods during the 1970s. *Social Science Quarterly, 66*, 346–364.

Lee, B. A., & Wood, P. B. (1991). Is neighborhood racial succession place-specific? *Demography, 28*, 21–40.

Lobo, A. P., Flores, R. J. O., & Salvo, J. J. (2002). The impact of Hispanic growth on the racial/ethnic composition of New York City neighborhoods. *Urban Affairs Review, 37*, 703–727.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2005). *Geographic information systems and science* (2nd ed.). New York: John Wiley & Sons.

Martin, D. (1989). Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers, 14*, 90–97.

Massey, D. S. (1983). A research note on residential succession: The Hispanic case. *Social Forces*, *61*, 825–833.

Massey, D. S., Gross, A. B., & Shibuya, K. (1994). Migration, segregation, and the geographic concentration of poverty. *American Sociological Review*, *59*, 425–444.

Massey, D. S., & Mullen, B. P. (1984). Processes of Hispanic and black spatial assimilation. *American Journal of Sociology, 89*, 836–873.

McHarg, I. L. (1969). *Design with nature*. New York: Doubleday/Natural History Press.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *Professional Geographer, 55*, 31–42.

Monmonier, M., & Schnell, G. (1984). Land use and land cover data and the mapping of population density. *International Yearbook of Cartography, 24*, 115–121.

Myers, D. (1999). Demographic dynamism and metropolitan change: Comparing Los Angeles, New York, Chicago, and Washington, DC. *Housing Policy Debate, 10*, 919–954.

O'Campo, P., Xue, X., Wang, M. C., & Caughy, M. (1997). Neighborhood risk factors for low birth weight in Baltimore: A multilevel analysis. *American Journal of Public Health, 87*, 1113–1118.

Odland, J., & Ellis, M. (2001). Changes in the inequality of earnings for young men in metropolitan labor markets, 1979–1989: The effects of declining wages and sectoral shifts within an efficiency wage framework. *Economic Geography, 77*, 148–179.

Openshaw, S. (1983). *The modifiable areal unit problem. Concepts and techniques in modern geography*, Vol. 38. Norwich, England: Geobooks.

Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning A, 16*, 17–31.

Pozzi, F., Small, C., & Yetman, G. (2003). Modeling the distribution of human population with night-time satellite imagery and gridded population of the world. *Earth Observation Magazine*, *12*(4), 24–30.

Qiu, F., Woller, K., & Briggs, R. (2003). Modeling urban population growth from remotely sensed imagery and TIGER GIS road data. *Photogrammetric Engineering and Remote Sensing*, *69*, 1031–1042.

Quillian, L. (1999). Migration patterns and the growth of high-poverty neighborhoods, 1970–1990. *American Journal of Sociology*, *105*, 1–37.

Radeloff, V. C., Hammer, R. B., Voss, P. R., Hagen, A. E., Field, D. R., & Mladenhoff, D. J. (2001). Human demographic trends and landscape level forest management in the northwest Wisconsin pine barrens. *Forest Science*, *47*, 229–241.

Reibel, M. (2000). Geographic variation in mortgage lending: Evidence from Los Angeles. *Urban Geography, 21*, 45–60.

Reibel, M. (2003). Measures of geographically uneven subpopulation group change and local group transitions: Examples from Los Angeles. *Geographical Analysis*, *35*, 257–271.

Reibel, M., & Agrawal, A. (2005). *Land use weighted areal interpolation.* Paper presented at the GIS Planet 2005 International Conference, Estoril, Portugal, May.

Reibel, M., & Agrawal, A. (2007). Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review, 26*(5–6), doi:10.1007/s11113-007-9050-9.

Reibel, M., & Bufalino, M. E. (2005). A test of street weighted areal interpolation using geographic information systems. *Environment and Planning A*, *37*, 127–139.

Ryznar, R. M., & Wagner, T. W. (2001). Using remotely sensed imagery to detect urban change: Viewing Detroit from space. *Journal of the American Planning Association*, *67*, 327–336.

Sadahiro, Y. (2000). Accuracy of count data transferred through the areal weighting interpolation method. *International Journal of Geographical Information Science*, *14*, 25–50.

Smith, R. A. (1991). The measurement of segregation change through integration and deconcentration, 1970–1980. *Urban Affairs Quarterly*, *26*, 477–496.

South, S. J., & Crowder, K. D. (1997a). Residential mobility between cities and suburbs: Race, suburbanization, and back-to-the-city moves. *Demography, 34*, 525–538.

South, S. J., & Crowder, K. D. (1997b). Escaping distressed neighborhoods: Individual, community and metropolitan influences. *American Journal of Sociology, 103*, 1040–1084.

South, S. J., & Crowder, K. D. (1998a). Leaving the 'hood: Residential mobility between black, white, and integrated neighborhoods. *American Sociological Review*, *63*, 17–26.

South, S. J., & Crowder, K. D. (1998b). Housing discrimination and residential mobility: Impacts for blacks and whites. *Population Research and Policy Review*, *17*, 369–387.

South, S. J., & Crowder, K. D. (1999). Neighborhood effects on family formation: Concentrated poverty and beyond. *American Sociological Review, 64*, 113–132.

Subramanian, S. V., Chen, J. T., Rehkopf, D. H., Waterman, P. D., & Krieger, N. (2005). Racial disparities in context: A multilevel analysis of neighborhood variations in poverty and excess mortality among black populations in Massachusetts. *American Journal of Public Health*, *95*, 260–265.

Sucoff, C. A., & Upchurch, D. M. (1998). Neighborhood context and the risk of childbearing among metropolitan-area black adolescents. *American Sociological Review*, *63*, 571–585.

Sucoff, C., Upchurch, D., & Aneshensel, C. (1999). Neighborhood influences on parent-child relations: Implications for adolescent health. *Journal of Adolescent Health*, *24*, 113–113.

Taeuber, K. E., & Taeuber, A. F. (1965). *Negroes in cities.* Chicago, IL: Aldine.

Tatian, P. A. (2002). *Neighborhood change data base 1970-2000 tract data. Data users' guide, short form release*. Washington, DC: The Urban Institute.

Temkin, K., & Rohe, W. M. (1998). Social capital and neighborhood stability: An empirical investigation. *Housing Policy Debate*, *9*, 61–88.

Tobler, W. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association, 74*, 519–536.

Vandell, K. (1981). The effects of racial composition on neighborhood succession. *Urban Studies*, *18*, 315–333.

Ward, D., Phinn, S. R., & Murray, A. T. (2000). Monitoring growth in rapidly urbanizing areas using remotely sensed data. *Professional Geographer*, *52*, 371–386.

Wright, J. K. (1936). A method of mapping densities of population with Cape Cod as an example. *Geographical Review*, *26*, 103–110.

Wright, R., Ellis, M., & Reibel, M. (1997). The linkage between immigration and internal migration in large metropolitan areas in the United States. *Economic Geography*, *73*, 234–254.