Environment and Planning A 2005, volume 37, pages 127 – 139

DOI:10.1068/a36202

Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems

Michael Reibel
Department of Geography and Anthropology, California State Polytechnic University, Pomona, CA 91768, USA; e-mail: mreibel@csupomona.edu

Michael E Bufalino
Center for Geographic Information Science Research, California State Polytechnic University, Pomona, CA 91768, USA; e-mail: mebufalino@csupomona.edu
Received 21 August 2003; in revised form 5 April 2004

Abstract. Data processing for the spatial analysis of small-area social, demographic, and economic data often requires the combination of data spatially aggregated to two or more incompatible zone systems in a region, such as a set of enumeration districts that changes over time. Such situations can be addressed by areal interpolation—the transfer of data between zonal systems according to spatial algorithms. The authors test a technique of areal interpolation using geographic information systems (GIS) that employs a digital map layer representing streets and roads to derive varying density weights for small areas within aggregation zones. The technique reduces errors in estimation compared with estimates derived using the commonly applied area-weighting technique, with its assumption of uniform density. The street-weighting technique is much easier to use than other interpolation techniques that have also been shown to reduce error compared with area-based weighting.

Introduction

Frequent changes in the geography of enumeration districts at the local scale are a constant frustration to analysts wishing to measure dynamic sociodemographic and economic trends. Our goal in this paper is to evaluate and test the errors associated with several techniques of interpolation which use geographic information systems (GIS). The street-weighted technique can be used to derive relatively accurate estimates of social, demographic, and economic trends for an exhaustive and complete set of local areas across a region, despite changes in the boundaries of the local areas during the trend interval. The technique can also be applied to cross-sectional problems in which data from two incompatible superimposed sets of areal units must be combined—a common problem for many spatial analysts, particularly in market research.

In many nations, local enumeration districts are designed for a target population size, and thus vary in area inversely with population density. For example, US census tracts are typically drawn to include approximately 4000 individuals, and can range in size from less than 1 km^2 in very dense urban zones to areas the size of Belgium in rural districts. The development of previously rural areas, or population decline in older urban areas necessitates the redrawing of such enumeration district boundaries to maintain a reasonable range of populations across the set of areal units. Splits, mergers, and complex territorial recombinations are the result.

The typical solution to the problem in which spatial data aggregated to incompatible, superimposed geographies must be combined is *areal interpolation*. This is the process by which data associated with one set of zones (the source layer) are assigned to the other set of zones (the target layer) according to defined algorithms. In cases where trend estimates must be computed during an interval in which enumeration district zones change, known counts at one boundary of the time interval are assigned to the areal units in effect at the other boundary of the time interval. To standardize to

the more recent set of zones, areal unit counts at time t_0 must be allocated, with minimum error, to the (different) zones in use for the same region at time t_1.

Areal interpolation and dasymetric mapping

The simplest approach to areal interpolation is to begin by joining (intersecting) the two zone-boundary layers, thus generating a set of zone fragments each of which has a unique pair of source and target zones [Flowerdew and Green (1992), call these fragments 'intersection zones']. Fragment-count estimates from the source time period can be generated by area weighting, that is, by multiplying the source-zone count by the ratio of the area of the fragment to the area of the source zone. The population estimate for a given target zone is the sum of these weighted counts across the set of source-zone fragments that exhaust the territory of the target zone[1] (Fisher and Langford, 1995):

$$\hat{P}_t = \sum \frac{A_{ts} P_s}{A_s}, \tag{1}$$

where $\hat{P}_t$ is the estimate target-zone count; A_{ts} is the area of the fragment belonging to a given pair of target and source zones, t and s; P_s is the source-zone count; and A_s is the source-zone area. The estimated fragment counts for the source-layer variable are then simply summed across all fragments belonging to each target zone to derive a set of source-layer count estimates for each target zone. In the case of enumeration districts changing over time, once corresponding sets of counts (or estimates) originating both in the source (time t_0) and in the target (time t_1) layers are available for a single set of areal units (the target layer), the trend can be computed by subtraction.

This form of area weighting meets the minimum criterion for areal interpolation—the pycnophylactic property, according to which source population is neither reduced nor increased by the process of weighting and assignment to a new set of zones (Tobler, 1979). In other words, for an areal-interpolation technique to be pycnophylactic, the (observed) size of the source population summed across all original source units in the region must be the same as the sum of the source-population estimates for the set of all target zones to which the source-unit populations are assigned. The problem with the basic area-weighting approach is that it assumes uniform count densities *within* the source-zone areas. Needless to say, this assumption is almost never accurate, and can be wildly inaccurate in areas (such as most cities in California) where steep hills and mountains create barriers to urban expansion. The result is inaccuracy and likely systematic bias in count estimates.

Over the years, geographers have developed more sophisticated techniques to reduce error and systematic bias in estimates derived by areal interpolation. The techniques fall into two main categories: smoothing techniques; and dasymetric, or ancillary weighting, techniques. Tobler (1979) theorized the smoothing technique for areal interpolation. In the smoothing approach, the source areas are broken into a lattice. The lattice cells are then assigned a portion of their source-zone counts, weighted along a smooth density gradient computed by interpolating between the density of the source zone of the pixels and those of its nearest neighbor zones. The weights are constrained so as to preserve the pycnophylactic property. To the extent that aggregate density differences between adjacent areal units accurately reflect smooth underlying

[1] Note that, in order to preserve the pycnophylactic property, the set of source zones must completely overlie the entire territory of the target region. Such is the case in our example of Los Angeles County.

surface counts, the smoothing technique gives reliable estimates of within-zone density variations. Error is introduced when unit boundaries reflect relatively sharp changes in the true density surface or, more generally, when true density gradients do not follow smooth paths from the centroid of a given unit to the centroids of its adjacent neighbors.

Dasymetric mapping appears to be the general approach of choice in areal interpolation. Mennis (2003) defines dasymetric mapping as areal interpolation that uses ancillary spatial data to aid in the interpolation process. Typically, in dasymetric mapping source-layer zones in the region are first transformed into a surface lattice—in much the same way as in the smoothing techniques. But, rather than using purely mathematical algorithms to interpolate lattice-cell values from known data associated with the original source zones, an ancillary data layer is added to the lattice and a weighting scheme is applied to cell counts according to known or derived density levels associated with values in the ancillary data. It is a relatively simple matter to constrain the weighted cell-count estimates in such a way as to preserve the pycnophylactic property.

To date, dasymetric techniques have generally used remote sensing data to derive density weights from inferred urban land cover (Cockings et al, 1997; Eicher and Brewer, 2001; Langford et al, 1991). In this approach, the lattice resolution and rectification are selected to coincide with those of the satellite imagery used to generate the land-cover weights. In a direct comparison, Cockings et al reported that their land-cover weighted estimates were more accurate than area-weighted estimates for 84% of the target zones in their study area.

The obstacle to more widespread use of remote sensing data for areal interpolation is the need to process data in a raster geographic information system (GIS) environment, when the data-source zones (enumeration districts) are polygons—the digital map layers of which are made available in vector GIS format by government statistical agencies. Indeed, this is an obstacle to the widespread use of any technique, such as smoothing, that requires a lattice surface for weighting and computation. Vector GIS skills and installed software are more common than raster GIS, particularly among demographers, planners, local government technicians, and market analysts who use spatially referenced social and economic data. In applied settings, area weighting is ubiquitous when areal interpolation is necessary; dasymetric weighting using remote sensing has been almost completely restricted to computational experiments by geographers and allied spatial scientists. To promote the use of improved interpolated estimates in applied settings, an areal-interpolation technique must be developed that reduces count-estimate error and uses only ancillary data readily available in vector format.

The street-weighting technique

We statistically tested just such a dasymetric areal-interpolation technique that relies on readily available vector GIS data for ancillary local weighting, and that therefore does not require the raster GIS skills and capabilities needed to overlay remotely sensed data on enumeration geographies. The ancillary data we used are the digitally coded streets and roads, which in the USA are distributed by the Census Bureau and other sources as a vector GIS layer—the Topologically Integrated Geographic Encoding and Referencing, or TIGER, files (US Census Bureau, 1993). Although others have used this street-weighting technique, we believe there has been no previous statistical test of its accuracy.

The street-weighting technique has been used in at least one applied demographic series (Ong, 1996; Ong and Houston, 2003) and very likely others. In the methods-research literature, Xie (1996) describes the street-weighting technique, calling it the 'overlaid network algorithm'. Xie elaborated three approaches, including a network-length method that is identical to the technique used in this study. Xie did not, however, compute the errors in estimation for his techniques.

Methods and data

To test the street-weighting method statistically, we used Los Angeles County as a case-study area; interpolating the 1990 Census tract counts of persons and housing units to the year 2000 Census tract geography. The resulting count estimates were then analyzed for errors, using as a benchmark the 1990 counts associated with the 2000 Census tract areas computed by aggregating the city-block level 1990 counts to their corresponding 2000 Census tracts (see appendix A for details of our block-aggregation methodology). Our ability to aggregate benchmark counts in this way permits us to compute errors in estimation for each target-zone area (2000 Census tract) by subtracting the 1990 benchmark counts from our 1990 estimates for each target zone in the study area. We then analyzed distributions and statistics for the estimation errors directly. This permits us to establish the magnitude of estimation errors associated with each technique (area weighting versus street weighting), and to test formally whether the error reduction over area weighting which we achieve by means of the street-weighting technique is statistically significant.

Computation of the street-weighted count estimates

The first stage of street-weighted areal interpolation is to superimpose in a vector GIS environment three digital maps: the source-zone boundaries, the target-zone boundaries, and the street layer (US Census Bureau, 1993). The layers must be in a common coordinate system and projection. The next task is to compute the street weight for each intersection-zone fragment. In order to preserve the pycnophylactic property, the street weights for each intersection zone are computed as the ratio of the aggregate length of the street vectors in the intersection zone to the aggregate length of the street vectors in the source zone:

$$W_{st} = \frac{\sum L_{st}}{\sum L_s}, \tag{2}$$

where W_{st} is the weight for a given intersection-zone fragment defined by its unique pair of source and target zones s and t; L_{st} is the length of each street vector in that intersection zone; and L_s is the length of each street vector in the source zone pertaining to that intersection zone.[(2)]

Once the weights for each intersection zone have been calculated, they are applied to the original source-zone counts attached to each one. The ratio formula of the weighting scheme preserves the pycnophylactic property when the weights are applied to the original counts. The weighted source-count estimates for each intersection zone are then summed across each target zone. The result is street-weighted estimates of the source-zone counts reassigned to the target geographic areas.

Error analysis

For this study, we computed both the area-weighted estimates and the street-weighted estimates for the Los Angeles County study area, as well as the 1990 total population and housing-unit counts for the 2000 Census tract areas (aggregated from 1990 block data, as explained in the appendix). Using the latter data as a benchmark, we computed errors in estimation both for the area-weighted and for the street-weighted estimates. Table 1 summarizes the distributions of errors for both of the two variables, each estimated according to the two weighting algorithms.

[(2)] The intersection zone defined by a unique pair of source and target zones need not be contiguous. In noncontiguous cases, the weighting process accurately assigns count estimates to the intersection zone as a set of zones. The set of noncontiguous intersection zones vanishes when it is reaggregated, without error, into the target zone in a later step.

Table 1. Error distributions—street-weighted and area-weighted estimates.

Percentile	Housing units		Total population	
	street weighted	area weighted	street weighted	area weighted
Minimum	−2 634	−4 606	−8 088	−10 027
1st	−923	−2 535	−2 669	−3406
5th	−471	−1 593	−1 271	−1 444
10th	−288	−1 231	−844	−851
20th	−152	−801	−443	−346
30th	−91	−496	−242	−184
40th	−39	−305	−109	−91
Median	0	−103	−13.5	−16
60th	41	106	105	52
70th	87	409	247	126
80th	156	759	428	302
90th	285	1 299	811	748
95th	445	1 823	1 320	1 397
99th	898	2 850	2 721	4 315
Maximum	2 160	6 225	7 668	10 022
Standard deviation	305.2	1 063.2	893.0	1 117.6

Table 1 clearly shows a contrast between the errors generated using area weighting and those from street weighting: the error distributions for both sets of street-weighted estimates are remarkably symmetrical and approximately normal; their medians are small in magnitude, and the category breaks are very similar in magnitude to their corresponding reflections across the median. For example, in the person-count distribution, the 30th and 70th percentiles are −242 and 247, respectively; the 10th and 90th percentiles for the housing unit counts are −288 and 285, respectively, and so forth.

The distributions of errors for the area-weighted estimates lack these desirable properties: their medians are of greater (negative) magnitude, particularly the housing-count estimates, indicating that the distributions are right skewed. Most importantly, numerous large outliers plague the area-weighted distributions. Indeed, the area-weighted errors for the count of persons have a larger standard deviation despite consistently *lower* errors over the middle sixty percentiles of their range.

Figures 1 and 2 (see over) show histograms of the errors in street-weighted estimation for the two test variables, total-population and housing-unit counts (respectively), with the frequencies of errors within the given ranges indicated on the vertical scale. The approximation of normality in the error distributions is again apparent from an inspection of figures 1 and 2: the distributions cluster narrowly and symmetrically around their modes, which are, in the respective categories, centered on zero. Outliers, although of considerable magnitude, are few in number. The change in the scale of the *x*-axis shows the considerably greater range of errors for population counts than for housing-unit counts derived from the street-weighting technique.

Standardized error measures

To establish measures of estimate error that can be used to compare errors from several variables, and errors in estimates for a given variable derived from multiple interpolation methods, we follow Eicher and Brewer (2001) and Fisher and Langford (1995) in using root mean squared (RMS) error. RMS error is a somewhat more rigorous measure than the mean absolute error used by Goodchild et al (1993), because it is more sensitive to outliers in the error distributions. Table 2 (see over) shows the

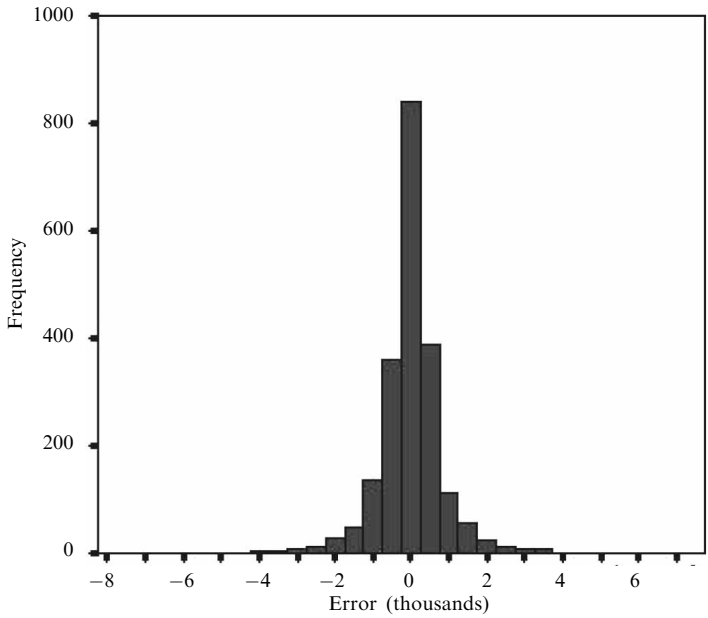

Figure 1. Estimation errors—street-weighted person counts for 1990 Los Angeles County Census tracts interpolated to 2000 Census tract geography.

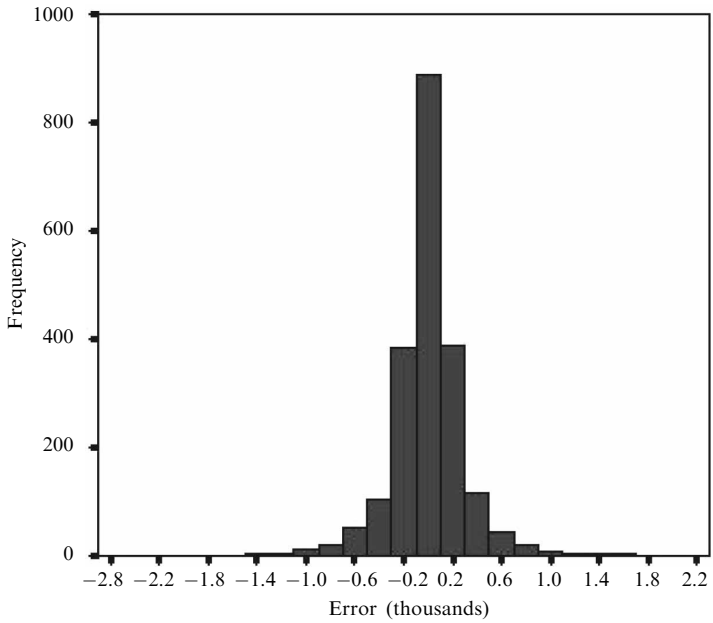

Figure 2. Estimation errors—street-weighted housing-unit counts for 1990 Los Angeles County Census tracts interpolated to 2000 Census tract geography.

statistics for population and housing-unit counts and estimates for 1990 Los Angeles County census tracts. Table 2 shows large differences in estimation error both for population and for housing-unit counts when the two different weighted areal interpolation techniques (area weighting versus street weighting) are used. Examining the RMS error levels for the two variables separately, we see that the area-weighted errors for housing-unit counts are much larger than the corresponding street-weighted

Table 2. Root mean square (RMS) error analysis, 1990 Los Angeles County Census tract count estimates interpolated to 2000 Census tract geography.

	Housing units		Total population	
	street weighted	area weighted	street weighted	area weighted
RMS errors	305.15	1062.99	892.76	1117.42
T-test, difference of means for RMS	31.02 (<0.001*)		7.11 (<0.001*)	
Mean, benchmark from block counts	1 541.88		4 321.47	
Coefficient of variation (RMS)	0.198	0.689	0.207	0.259
Error reduction (improvement over area weighting) (%)	71.26		20.08	

*Probability that observed difference is due to chance.

estimates for this variable, but the RMS errors for the population counts derived from the two techniques, although significantly quite different, are less divergent.

To compare directly the relative error reduction when the different techniques are applied to two (or more) count variables, we refer to the respective coefficients of variation for the set of error-estimate distributions. Comparison of these statistics shows the area-weighted errors for housing-unit counts are much larger than the area-weighted population-estimate errors. This volatility in area-weighted estimates indicates extreme sensitivity to the ubiquitous internal variations in density which violate the equal-density assumption upon which area weighting relies—a volatility which, depending upon the underlying surface geography of the construct being interpolated, can produce such large variations in area-weighted error distributions as between the person counts and housing-unit counts in our example.

Errors in the street-weighted estimates, by contrast, are far more consistent for the two variables in our example—their respective coefficients of variation differing by only 0.009. For both variables, the use of the street-weighting interpolation algorithm reduces error significantly over the area-weighting technique. Because the area-weighted errors for the housing-unit counts are so large, the improvement in estimation for housing-unit counts achieved by street weighting is much greater, by a factor of three, than the corresponding improvement for person counts.

Interpretation of error maps

Figures 3 and 4 (see over) map the population estimation error rates for the 1990 population counts interpolated to 2000 Census tracts for Los Angeles County derived from the area-weighting and street-weighting techniques, respectively. The four data intervals on each map are determined by two standard deviations in the error distribution of estimates generated from the street-weighting method: negative errors greater than two standard deviations; negative errors zero to two standard deviations; and similarly for positive errors. The use of the same absolute intervals permits direct comparison between the two maps.

Figure 3 shows the spatial pattern of errors obtained using area-based weighting. The most dramatic feature is the diagonal swath of high positive errors running from the northwest corner of Los Angeles County through to the eastern border. Readers familiar with southern California will recognize this region as corresponding closely to the San Gabriel Mountains and the associated highlands of the transverse ranges.

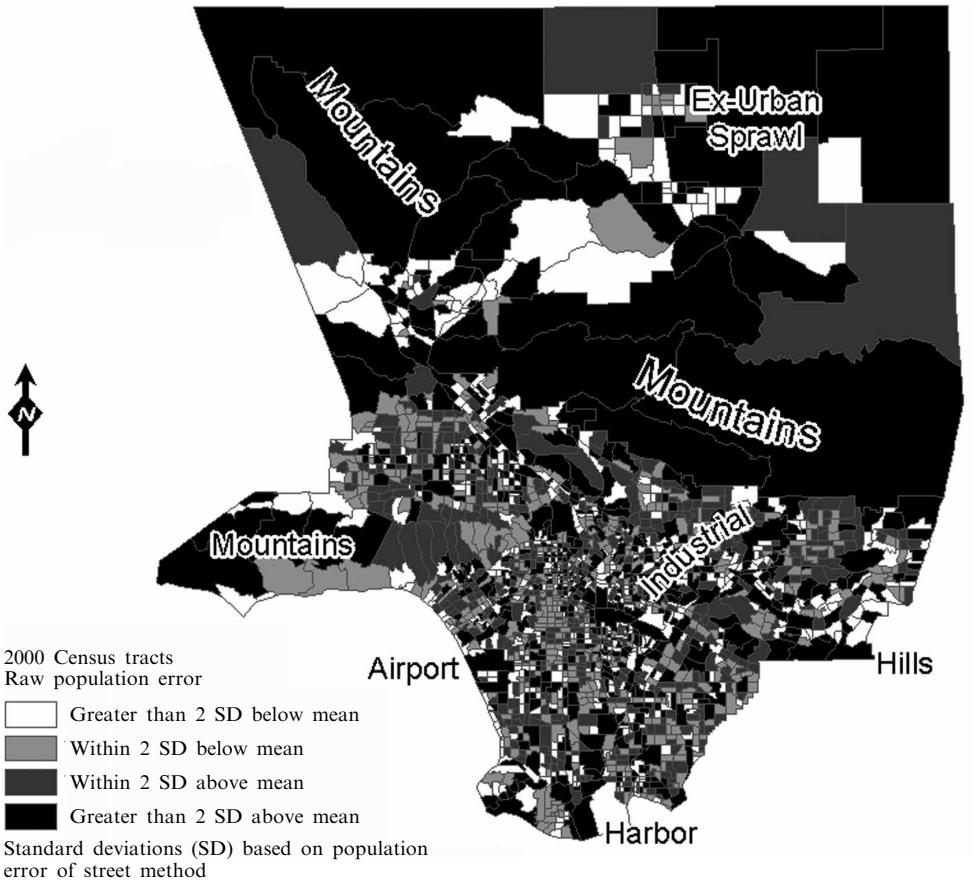

Figure 3. Map of area-weighted estimation errors.

Because both the area-weighting and street-weighting techniques preserve the pycnophylactic property, errors in a given 2000 Census tract are mirrored by corresponding errors of the opposite sign in some sets of nearby (typically adjacent) tracts that share one or more 1990 source tracts with the 2000 tract in question. Such corresponding 2000 tracts showing high negative errors can be seen in some fast-growing foothill communities immediately south of the large mountain region. Similar patterns of high error in which mountainous areas are erroneously assigned population at the expense of rapidly growing, adjacent, foothill areas can be seen in the Santa Monica Mountains in the extreme western part of the county, and in the Puente Hills (labeled 'Hills') at the east-by-southeast edge of the county.

The explanation for high errors in mountainous regions generated using the area-weighting method is straightforward: these are regions in which many zones split over the given time interval, reflecting rapid development in the foothill areas. Given the assumption of the area-weighting algorithm of uniform population density, the algorithm wrongly assigns most of the presplit population to the large, high-elevation fragments, which are sparsely populated, rather than to the smaller foothill fragments that were more densely populated even before the zone split.

A similar pattern is found in northeast Los Angeles County, where the cities of Lancaster and Palmdate are sites of rapid exurban development. Although topography is not an issue in this relatively flat, high-desert region, very large source zones are

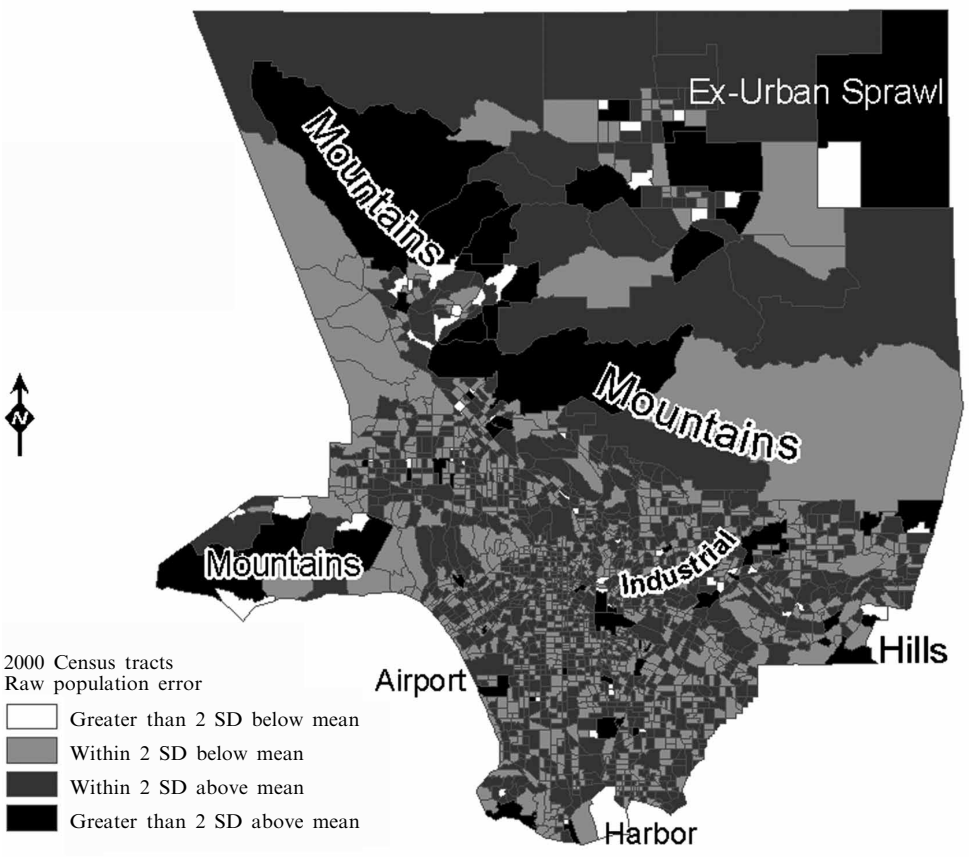

Figure 4. Map of street-weighted estimation errors.

once again split between small target zones representing areas of rapid development and larger target zones that correspond to the remaining sparsely populated areas. As in the mountainous regions, the developing parts of the desert region were already more densely populated before the split, leading to error when the area-weighting algorithm was applied.

The other types of regions showing clusters of high errors are those dominated by industry and transportation. Development near Los Angeles International Airport in the southwest led to a tract split in this area: because the airport itself has virtually no permanent residents, it is wrongly assigned a large population. Conversely, the tract (labeled 'Harbor') containing the Port of Long Beach in the far south is overbounded. Development in residential parts of the tract bordering the harbor is quite dense. Meanwhile, the land area of the target tract is much smaller than the boundary drawn on the map, which extends far into the harbor. The result is an underestimation of tract population from the use of area weighting, and erroneous assignment of harborside population to the adjacent landward tracts.

The industrial tracts with large positive errors are strung out in corridors along river basins. The first runs along the Los Angeles River from just east of downtown Los Angeles (due west of the 'I' in the label 'Industrial') down river to the south and east, through the industrial suburbs of Vernon and Commerce, and then turns south-southwest through Southgate, Paramount, and northwest Long Beach on its way to the harbor. The other corridor, as indicated by the label 'Industrial', is along the upper reaches of San Gabriel River to the northeast, beginning in Irwindale and flowing

southwest through the Whittier Narrows. All these areas are sparsely populated and heavily industrial, leading the area-weighting method to overestimate their populations.

Figure 4 maps population-estimate errors generated using the street-weighting technique. Comparing the spatial patterns of errors in figure 4 with those in figure 3, we see a similar pattern in which errors are concentrated in desert and upland regions experiencing rapid growth, as well as in regions that are devoted to transportation or industrial uses. Upon closer inspection, however, certain differences emerge: overall, there are fewer areas of error exceeding plus or minus two standard deviations.

The most consistent improvements in estimation are in the upland regions of mountains, hills, and foothills. This was expected, as the highest and most rugged (hence, most sparsely populated) parts of these regions are usually aggregated into large zones with few roads. The large land areas in such zones generate high positive errors when the area-weighting algorithm is applied, whereas the relative absence of roads causes the street-weighted estimates to be smaller and thus more accurate. Less dramatic improvement is also visible along the industrial corridors and in transportation-related areas showing high positive errors; presumably because the density of roads in such areas is lower than in residential areas, leading to correspondingly lower population estimates.

On the other hand, the street-weighting method brings no apparent improvement to estimation in the exurban desert sprawl region in northeastern Los Angeles County. Indeed, figures 3 and 4 show that the pattern of errors for this region, as indicated by the intervals chosen, is identical when the two estimation techniques are used. The explanation for this pattern appears to be the high density of roads in this region, the great majority of them unpaved, that are included in the Census street-layer data.

Discussion and conclusions

This study provides the first error tests of street weighting, a promising technique for dasymetric areal interpolation using a vector data layer for ancillary weighting. We have provided evidence that the street-weighting technique produces significantly lower errors in estimation than the area-weighting technique overall, and also produces more consistent errors when applied to different variable counts in a given study area. A visual comparison of the spatial distribution of population-estimate errors generated from the area-weighting and street-weighting techniques confirms that comparative error reduction is inversely proportional to the density of nonresidential roads. In other words, the street-weighting method appears to reduce errors most compared with the area-weighting method in those areas where the lack of population is reflected in the lack of roads, and least in those areas (such as industrial areas) with a more developed, but nonresidential, transportation infrastructure.

The overall reduction in error afforded by the street-weighting technique was expected, given that it (like all dasymetric techniques) incorporates some information on the internal density variations within source-area zones—information that is completely lacking in estimates computed using the area-weighting technique. That said, the street-weighting technique was superior to the area-weighting technique *across all intervals* of their respective error distributions *only* for the housing-unit counts; the statistically significant overall error reduction observed for the population counts is achieved entirely at the extremes of the error distributions. Indeed, in this initial case study using a relatively attribute-poor street layer, the street-weighted population-count errors were larger than the corresponding area-weighted errors over approximately the middle 70% of their respective distributions (see table 1).

The use of the street and road grid (aggregate segment length) as a proxy for approximate population density surfaces, as used in this study, is far from perfect: it requires the assumption that the residential population density gradient at a given

distance from the nearest street or road is constant. Moreover, within the constraints of this study, this assumption of constant density gradients is made without bringing to bear any attribute information about the streets and roads, such as traffic capacity or access, nor any information about the densities of structures within a given distance from streets, such as the proportion of those structures that are residences, or the population density per residence. Finally, this and any other case study of interpolation errors in a single study area can only provide evidence for that area which, as Fisher and Langford (1995) point out, may be anomalous.

Based on this initial, limited test, we suggest that the street-weighting technique provides better *overall* population estimates than does area weighting, and therefore is preferable if the choice is either/or. But the better performance of the area-weighting method in zones less prone to large errors implies that analysts can improve their results further by only using street weighting in those types of regions associated with large errors in this study. Such regions include fast-growing, previously undeveloped, areas bordering on areas that remain very sparsely populated, particularly in upland areas, as well as transportation and industrial complexes. It should be noted, however, that this selective approach to street weighting would require additional data processing based on decision rules which, in the absence of research on the matter, must be imposed a priori. Without a theoretically robust threshold rule for the use of one versus the other technique, such a step could well introduce more error than it removes.

Further research on the street-weighting method ought thus to investigate two areas initially: developing threshold rules based on local characteristics for the selective application of street weighting; and testing further improvements in error reduction using richer street and road data which distinguish, at a minimum, paved from unpaved roads. Moreover, it is quite possible, and perhaps likely, that in a direct comparison dasymetric areal interpolation using remotely sensed urban land-cover data, or vector land-cover information, would produce more accurate estimates of sociodemographic counts than the street-weighting technique discussed in this paper. Further research will directly compare street-weighted to land-cover-weighted estimates of sociodemographic-count variables over a common study area to determine their respective error levels.

The authors maintain, however, that the street-weighted areal-interpolation technique used here is valuable even should future research establish that it yields moderately larger errors than dasymetric areal interpolation with remotely sensed urban land-cover data. The rationale for this assertion is the evident improvement of the technique over area-weighted estimates, combined with its considerably greater ease of use when compared with classified urban land-cover-weighting techniques, whether raster or vector. Unlike the latter, the street-weighting technique employs easily accessible data and is feasible for anyone equipped to compute area-weighted estimates—namely, that vast majority of applied sociodemographic analysts whose knowledge and access are restricted to vector GIS.

References

Cockings S, Fisher P, Langford M, 1997, "Parameterization and visualization of the errors in area interpolation" *Geographical Analysis* **29** 314 – 328

Eicher C, Brewer C, 2001, "Dasymetric mapping and areal interpolation: implementation and evaluation" *Cartography and Geographic Information Science* **28** 125 – 138

Fisher P F, Langford M, 1995, "Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation" *Environment and Planning A* **27** 211 – 224

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

Goodchild M F, Anselin L, Deichmann U, 1993, "A framework for the areal interpolation of socioeconomic data" *Environment and Planning A* **25** 383 – 397

Langford M, Maguire D, Unwin D, 1991, "The areal interpolation problem: estimating population using remote sensing in a GIS framework", in *Handling Geographic Information: Methodology and Potential Applications* Eds I Masser, M Blakemore (Longman, Harlow, Essex) pp 55 – 77

Mennis J, 2003, "Generating surface models of population using dasymetric mapping" *The Professional Geographer* **55** 31 – 42

Ong P, 1996 *Final Socioeconomic Report for 1997 Air Quality Management Plan* South Coast Air Quality Management District, Diamond Bar, CA

Ong P, Houston D, 2003 *Draft Socioeconomic Report for 2003 Air Quality Management Plan* South Coast Air Quality Management District, Diamond Bar, CA

Sadahiro Y, 2000, "Accuracy of count data estimated by the point-in-polygon method" *Geographical Analysis* **32** 64 – 89

Tobler W, 1979, "Smooth pycnophylactic interpolation for geographic regions" *Journal of the American Statistical Association* **74** 519 – 530

US Census Bureau, 1993 *TIGER/Line Files, 1992* Department of Commerce, US Census Bureau, Geography Division, Washington, DC

Xie Y, 1996, "The overlaid network algorithms for areal interpolation problem" *Computers, Environment and Urban Systems* **19** 287 – 306

Appendix

Computation of benchmark counts for 1990 population of 2000 Census tracts

An imperfect match in the superimposition of the two layers complicated the aggregation of 1990 blocks into 2000 tracts. There were many small discrepancies in the position of boundaries in the two layers, generally due to increased accuracy of the 2000 Census boundary layer data. In the course of the computation, this led to many apparent gaps between boundaries, creating slivers of territory, whereas in reality the true boundaries were identical. The result was 47 814 fragments created by the layering. Upon inspection, however, 49% of these fragments constituted less than 5% of their source 1990 block area, and 36% constituted greater than 95% of their source 1990 block area. The concentration of frequencies at the extremes of fragment area proportion values confirms that the great majority of apparent splits are in fact artifacts of boundary-line discrepancies in the layering and geoprocessing.

Our solution to this problem was point-in-polygon aggregation. The 1990 blocks were aggregated into 2000 tract territories as determined by the location of their tract centroids. The use of point-in-polygon processing in this case is different from typical point-in-polygon interpolation: because the great majority of divided-source polygons are artifacts of spatial data mismatch and, because this technique accurately removes such errors, the processing step is better understood as being one primarily of 'data cleaning', rather than of 'data estimation'. In the few cases where 1990 blocks were in fact split by 2000 tract boundaries, point-in-polygon interpolation is expected to be highly accurate because of the very large size of the target zones relative to the source zones; because of the typically regular shape of block areas; and because of the availability of spatial centroid reference points. For a discussion of the importance of these factors, see Sadahiro (2000).

Point-in-polygon aggregation in this context is therefore expected to yield benchmark values which, with few exceptions, are identical to the observed 1990 counts, inasmuch as the technique corrects mismatches resulting from false fragments that are artifacts of boundary layering. It is therefore the best solution to an intractable spatial data processing problem, particularly because all alternative techniques require some type of interpolation of block data which would propagate the errors inherent in the false block splits that result from layering.