Spatial Errors in Small-Area Demographic Analysis: Estimating Population and Housing Characteristics for an Oregon School District

Irina V. Sharkova, Kenneth Radin

Population Research Center
Portland State University

Paper presented at the Population Association of America Annual Meeting, Los Angeles, California, March 30-April 1, 2006. Address correspondence to Irina V. Sharkova, Population Research Center, Portland State University, P.O. Box 751, Portland OR 97207-0751. E-mail: sharkovai@pdx.edu. Phone (503) 725-5160.

Spatial Errors in Small-Area Demographic Analysis: Estimating Population for an Oregon School District

Irina V. Sharkova, Kenneth Radin

Using a recent study to develop enrollment forecasts for the Medford School District (Oregon) as an example, this paper discusses common yet frequently overlooked sources of error in small area demographic analysis: positional errors. It identifies their sources and types and focuses on boundary mismatch errors arising from different spatial representations of the same study area. These errors are especially challenging when conflicting data are provided by trusted sources, such as GIS departments with the school district, city, or county. Using spatially referenced tax assessors' inventories, fine-resolution GIS imagery, and expert judgment, we corrected positional errors in the data and created "true" boundaries for the study area. Next, we developed population estimates for the school district and its attendance areas from "true" and conflicting boundary configurations. The paper compares gains in spatial accuracy with improvements in estimates' accuracy and discusses the results in light of efforts required to achieve them.

1. Introduction

Both academic and applied demographers have long been concerned with data quality. A solid body of research has produced standard, broadly accepted measures quantifying error in survey data such as US Census or Current Population Survey and their derivatives, as well as in estimates and forecasts of population, housing, and other characteristics. Yet most classical demographic studies on error have investigated large population groups representing the nation, states, or metropolitan areas, rarely examining areas smaller than counties (Smith 1987, Tayman et al. 1998, Smith and Cody 2004). Given well established locations of these big geographic entities, studying errors associated with their location would hardly seem important.

Meanwhile, on-going developments in geography and related disciplines have been generating new means to obtain data about smaller and smaller areas and their populations, and new methods of analyzing such data. This has been fueling a substantial interest in local (small-area) analysis of urban neighborhoods, school attendance areas, market areas, walking-distance accessibility zones, and so on. While attribute error measurements developed to-date continue to be useful and necessary, they are no longer sufficient, because they do not address a fundamental property of small-area data: location and errors associated with measuring the latter.

Geographic Information Science (GISc), the vehicle behind small-area data development and analysis of the last 15-20 years, has been studying effects of positional uncertainty and error on measurement and analysis of spatial phenomena. However, its findings have been slow to translate into demographic analysis, whether academic or applied. A welcome exception has been a recent increase in research to determine best ways of acquiring population and housing data for small-area demographic analysis when the study area's boundary does not match boundaries of census statistical areas for which the data was originally collected. Several areal interpolation methods have been proposed and evaluated. Yet these studies typically assume that boundaries of both the study area and census statistical areas are themselves correct, or positionally accurate, and spatial errors, if exist, are negligible. As experience shows, this assumption is often wrong.

In this paper we intend to partially fill this gap by focusing on spatial errors common in small-area demographic analysis. Using a project to develop population and enrollment forecasts for Medford (Oregon) School District as a case study, we quantify the amount of spatial error utilizing existing, but little-known measures. We do so at two geographic levels: regional (the entire district) and local (its attendance areas), and demonstrate that the amount of error depends on relative size of areas used in data interpolation. At the regional level, we investigate the relationship between positional errors resulting from boundary mismatch and respective demographic attribute errors. We explore this relationship for different data sources and methods of data interpolation. At the local level, we quantify spatial variations of error produced by boundary overlay of attendance areas and census statistical areas. In instances where exact amounts of attribute *error* cannot be established, we propose to use measures of *uncertainty* of demographic attributes. We also explore several correlates of spatial error including street network density and measures of boundary shape. The analysis demonstrates considerable variations in spatial error at the local level refuting a common misconception that spatial errors tend to balance themselves out. The study provides evidence that applied demographers should pay attention to positional accuracy when conducting small-area analyses.

The paper begins with a discussion of spatial error types, sources, and measurement issues, followed by the study area overview and study background. Next, District-level analysis of boundary mismatch and resulting errors is presented. An analysis of spatial errors at the local (attendance area) level follows. The paper concludes with a discussion section and recommendations for future studies.

2. Types, sources, and measures of spatial error

In GISc, "spatial error" is a broad term describing errors in observations, measurements, and analysis associated with location (Chrisman 1987, Longley et al 2005). It includes *positional* errors resulting from incorrect measurement of coordinates (as happens when student addresses get geocoded to a wrong school district), and *attribute* errors: misclassification or incorrect identification of characteristics of geographic (spatial) objects and phenomena. An example of the latter is an incorrect total of student residencies in a school attendance area due to errors in geocoding. Temporal accuracy, internal logical consistency, and completeness are other important characteristics of spatial data quality. Spatial error results from spatial *uncertainty*: lack of perfect knowledge about a geographic location and its characteristics.

Spatial errors can arise from many sources, including original measurements, data processing, and methods of analysis. In small-area demographic analysis, most common sources are utilization of poorly documented, "legacy" spatial data which may be inappropriate for the task at hand, uncorrected mismatch of coordinate systems and projections of spatial data sets, mismatch of scales at which data sets were first developed[1], and inattention to topological (internal logical) errors in vector spatial data, all of which can produce falsely non-matching boundaries. Contributing factors can be lack of coordination between agencies involved in data production and distribution, and lack of agreed-upon knowledge of the exact spatial extent of a given study area.

A particular strength of spatial data is its capacity for spatial overlay, or integration of geographic data sets based on their mutual location. Spatial overlay allows transfer of data from one zonal system (for example, Census statistical areas, municipal planning areas, point locations of new housing constriction) to another (for example, neighborhood boundaries, 10-minute walking accessibility zones of grocery stores). Several methods of spatial data transfer (areal interpolation) between *source* and *target* units (zones) have been developed and tested (Goodchild and Lam 1980, Goodchild et al 1993). They include simple areal interpolation, point-in-polygon aggregation, kernel method, and areal interpolation using supplementary data

[1] For example, U.S. Census TIGER data has been developed at a scale 1:100,000, while many common municipal applications rely on digital and paper maps with a scale 1:24,000.

(intelligent interpolation) (Goodchild et al 1993, Sadahiro 1999, Eicher and Brewer 2001).

The main difference between these methods is an assumption about the type of spatial distribution of the attribute (characteristic) of interest in the source zones. Thus, *simple areal interpolation* assumes that the attribute of interest, such as housing units, is evenly distributed in each source zone. The attribute of each source zone is then summed up to the target zone in proportion to the area of the source zone in a given target zone. An obvious shortcoming of this method is the fact that population and housing characteristics are almost never evenly distributed: group quarters, multi-family housing, and mobile homes, for instance, have a highly uneven spatial pattern, concentrating in some parts of a study area and absent in many others. In such cases, significant spatial error can result from interpolation (Sadahiro 1999).

Point-in-polygon method assumes that all spatial objects, such as people, are located exactly on the representative point of each source zone (called "centroid" in vector spatial data models); the interpolation then involves summing up all the counts allocated to representative points that are included in the target zone (Sadahiro 1999). For example, (Reibel and Bufalino 2005) calculated 1990 population of 2000 Census tracts by adding population of all 1990 Census blocks whose centroids were located inside a given 2000 Census tract and assigning it to that tract. While this is the fastest of all interpolation procedures, its assumptions regarding distribution of population and housing characteristics are even less realistic than in the simple areal interpolation. Additionally, a representative point of each source zone (its "centroid") can be located anywhere in the zone, and supplementary steps have to be taken to assure that it is assigned to the geographic center of the zone. Given a somewhat random location of a representative point in its zone, a source zone can be easily excluded from an overlay even though a good portion of it could be inside a target zone (Schlossberg 2003).

Kernel method is somewhat in-between the above two methods. It assumes that spatial objects are spread across the source zone, as in simple areal interpolation, but most of them are clustered around the representative point of each source zone, as in point-in-polygon method. The method requires a mathematical function to describe the distribution (Sadahiro 1999) and has been rarely, if ever, used in small-area demographic analysis. The method's reliability depends a lot on the type of the function chosen to represent the distribution of population and housing characteristics.

Finally, *intelligent* interpolation uses supplementary data such as satellite images or land use data to determine the distribution of a characteristic (spatial objects) in each source zone (Goodchild et al 1993, Sadahiro 1999). As a minimum, the goal is to exclude areas where no population and housing characteristics can be found (parks, industrial and most commercial land uses). Several other types of data can be used as supplementary (ancillary) data: a street network layer (Xie 1995, Reibel and Bufalino 2005), tax assessor's data, residential zoning data, and so on. Research has shown that, providing the ancillary data represents a distribution of population and housing characteristics fairly well, intelligent interpolation results in lowest spatial errors of all the methods discussed so far.

Overall, any areal interpolation method produces less attribute error when source zones are relatively small compared with the target zones (Sadahiro 2000, Gregory 2002). Additionally, the amount of error appears to vary depending on the shape of the source and target zones, type of the variable being interpolated, and characteristics of supplementary data (Flowerdew and Green 1994, Sadahiro 1999 & 2000, Gregory 2006).

The above studies have considered cases of "legitimate mismatch" between source and target zones: a true overlap of the boundaries. Yet in practice, a "legitimate mismatch" is often, if not always, accompanied by a misalignment of the boundaries (a true positional error). Misalignment of source and target zones' boundaries used in spatial overlay result in *slivers*, small, elongated polygons between misaligned boundaries. It's been often advised to merge such polygons smaller than a certain threshold with neighboring zones, although some researchers have found that such procedures may not be appropriate (Edwards and Lowell 1996, *cited in* Gaeuman et al. 2005). In a study area with large variations in shapes and sizes of source and target zones (e.g., a school district straddling both densely populated, small, regular-shaped urban census blocks and sparsely populated, large, irregular-shaped rural census blocks) it is difficult to establish a meaningful threshold to separate slivers from legitimate small source polygons. Additionally, when small spatial objects such as tax lots are used as supplementary data during intelligent interpolation, they may be incorrectly identified as slivers and eliminated.

A recent study by Gaeuman, Symanzik, and Schmidt (2005) develops a model that could be used to differentiate between true slivers and legitimate small polygons. While the authors' goal is to determine how to reliably identify areas of actual landscape change detected by a spatial overlay versus areas of false change produced by incorrect boundary positions, their model appears more broadly applicable. The authors examine boundary-scale errors: small errors resulting from overlay of slightly mismatching large polygons. They measure a length of a given boundary segment and an area of mismatch, and develop a formula to calculate probable error. They conclude that "estimated magnitude of the probable error caused by incorrect boundary positions must be less than 0.8 times the area of change measured on the overlay for the overlay to be useful". In other words, an area resulting from an overlay of a source zone and a target zone must be 0.8+ times larger than the estimated magnitude of the probable error for the area to be considered a legitimate small polygon rather than a sliver.

However, the above model is computationally intensive and may be impractical when several hundred of source zones such as Census blocks are involved. Several other measures developed in the GISc literature are used in this paper to evaluate positional error. At the *District level*, where we have a likely benchmark, or true, boundary, we calculate percent difference in area size of suspect boundary configurations vis-à-vis the benchmark, and Average Distance Error[2] between vertices of a given boundary and the benchmark. We also calculate attribute errors associated with positional errors: percent difference in population and housing characteristics in comparison with benchmark population and housing data. At the *local level*, where benchmark boundaries or socio-demographic data were not available, we calculate measures of spatial error that take into account both positional error and effects of legitimate overlap of boundaries of source and target zones. We also propose and calculate measures of attribute uncertainty resulting from positional error and legitimate overlap.

[2] In cases where only two boundaries are compared at a given time, only one distance measurement for each boundary vertex and its closest benchmark boundary location can exist. Therefore, Distance Root Means Square (dRMS), a measure commonly used to gauge positional errors (Siouris 1993, *cited in* Gaeuman et al. 2005) , is computationally identical to Average Distance Error.

3. Study area overview & background

Medford School District is located in Jackson County, Southern Oregon, and serves approximately 12,800 students in grades K-12. The area encompasses cities of Medford (population 70, 860) and Jacksonville (population 2,500) and a sizable, sparsely populated unincorporated area (approximately 3,500 persons). The District saw a strong growth in its housing, population, and school enrollment through the 1990s, however, in the recent years its enrollment growth has stagnated as the share of retirees and higher-income in-movers, - two populations with few or no school age children, has increased. The District asked the Population Research Center (PRC) to prepare an enrollment forecast for the 2005-2015 period to assist in long-term facilities planning. The forecasts were completed by grade for the District as a whole, as well as for its elementary, middle, and high schools.

The study utilized several methods; they included Cohort-Component model for District-wide forecasts, Housing Units method for elementary school attendance area forecasts, and a combination of Reside-Attend Ratios and Grade-Progression Ratios to establish a correspondence between forecasts for 14 elementary school attendance area (ESAAs) and individual schools. Thus, data for the models had to be developed at two geographic levels: the District level and the ESAA level. Among data sources we used were 1990 and 2000 Census, administrative records such as births by mother's place of residence, student addresses, and building permits, tax lot-level land use and zoning data, and household forecasts by transportation analysis zones (TAZs) produced by local planners. All of these data are location-specific and therefore, all are potentially influenced by positional errors arising in the process of data development (geoprocessing of the data).

In this paper we will discuss positional errors associated with a subset of geoprocessing tasks used in the study, namely, *areal interpolation* of population and housing characteristics between Census statistical areas (source zones) and the District as a whole and its elementary school attendance areas (target zones). PRC’s primary method for Census data interpolation was *intelligent dasymetric areal interpolation* (see Appendix for a detailed discussion of the method).

An obvious first step in any spatial analysis is establishing a geographic extent of the study area(s): its boundaries. Although it is still common for applied demographers to encounter lack of digital, GIS-compatible boundary data for areas of interest, dramatic progress has been achieved in recent years in spatial data development at the national, state, and local levels. In fact, the very first challenge we faced was availability of *too many sources* of boundary data for the District. Six sources of data were identified: the Medford School District GIS; the City of Medford planning department; Jackson County GIS department; Oregon Geospatial Enterprise Office (Oregon GEO); ESRI; and 2000 TIGER data. These data sources produced 4 different boundary configurations (see Fig. 1). The combined area of overlap between them takes only 3 percent of the combined study area (all areas in color on Fig. 1). While small percentage-wise, this nevertheless results in about 7,200 acres where uncertainty exists about the exact location of the District's boundary. Which one to use, and would it matter for the accuracy of the results?

Each data set has certain strengths and limitation. District-wide and ESAA boundaries produced "in-house" by the *District GIS* imply the highest degree of accuracy since they are used on a daily basis for a variety of tasks including bussing of students, and therefore need to be correct and up-to-date. They also were produced at 1:24,000 scale, which preserves more locational details than a more common 1:100,000 scale used in creation of 2000 TIGER data. However, these boundaries were poorly aligned with 1990 and 2000 source zones (Census blocks and blockgroups derived from 2000 TIGER) and had internal topological errors; they were also misaligned with tax lot data available via Jackson County GIS.

District-wide and ESAA boundaries from *Jackson County GIS* and the *City of Medford*, produced at 1:24,000 scale, were well aligned with streets and tax lots, as well as 2000 Census statistical areas: Jackson County GIS has processed 2000 TIGER data to spatially match the rest of their GIS data. If we were to use these data, minimal discrepancies for 2000 Census data interpolation would be encountered. Yet the County had not updated these data layers since 2003, while the District changed attendance area boundary configurations in mid-2004. The County also did not develop a matching set of 1990 Census blocks and blockgroups.

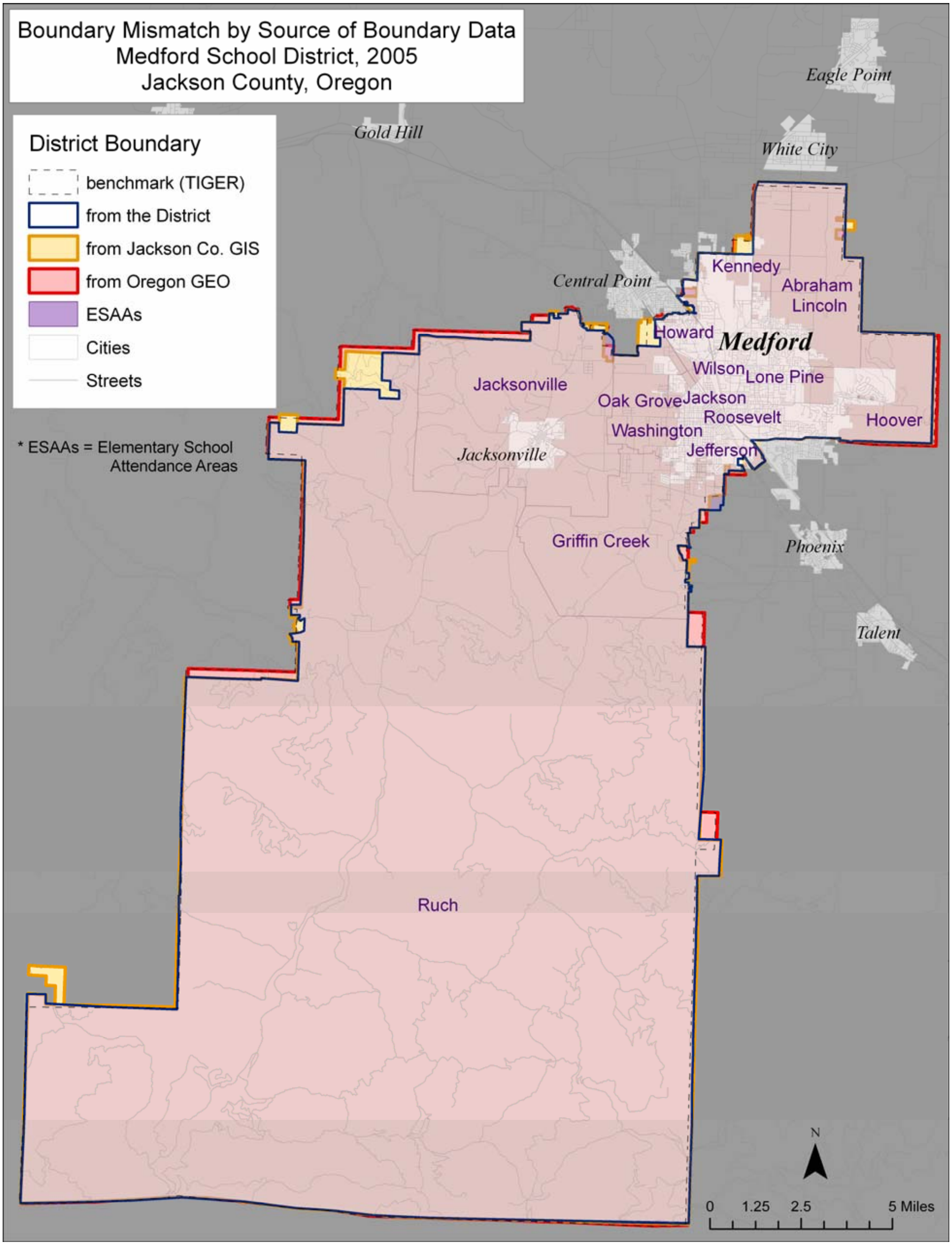
A boundary derived from *2000 TIGER* Census blocks dissolved by the District's unique ID (SDUNI = 08040) represented a territory for which District-wide data were readily available from the Census 2000 STP2[3], potentially saving data processing time, but attendance area attributes are not recorded in the Census data. Also, the 2000 TIGER boundary seemed misaligned with other features, which suggested a datum or another error of reprojecting the data to match the Oregon South State Plane coordinate system chosen as a common system for the analysis. Finally, *Oregon GEO* and *ESRI* data were aligned with both 1990 and 2000 Census statistical areas, but there again, no ESAA boundaries were available.

If the project had had a less-limited timeline and a more generous budget, it would have been worthwhile to update ESAA boundaries from Jackson County GIS thus minimizing known spatial mismatch. Under existing circumstances, a decision was made to use the boundaries supplied by Medford School District and correct topological errors and only most obvious positional errors in the data.

In the following two sections we attempt to quantify effects of positional errors (boundary mismatch) on estimates of 2000 Census population and housing characteristics first for the District as a whole, and then for ESAAs.

[3] Census 2000 School District Tabulation Data (http://nces.ed.gov/surveys/sdds/c2000.asp)

Figure 1. Study Area: Conflicting Boundary Configurations



4. District-level (regional) analysis

For the purposes of this analysis we assume that the boundary derived from dissolving 2000 TIGER blocks is the true, or benchmark, boundary against which other boundaries will be compared. The choice is somewhat arbitrary: we do not know what boundary is most accurate, or closest to reality. We do know, however, that the STP2 school district Census data are provided for aggregations of 2000 TIGER blocks with same school district IDs. This is important as it gives us benchmark population and housing data, or true counts of Census characteristics.

Two measures of spatial mismatch by data source are presented in Table 1: percent difference in area sizes of the District as defined by the benchmark boundary versus a comparator boundary, and Average Distance Error between vertices of a given boundary and the benchmark. While there are no identifiable trends and only 3 cases of data sources, it is worth noting that the boundary provided by Medford School District GIS produces a higher degree of overall spatial mismatch than alternatives.

Table 1. Spatial Error by Source of District's Boundary

Data source	Acres	% difference with TIGER	Average distance error, feet (**)
TIGER 2000 Blocks (*)	235,531	n/a	n/a
Medford SD GIS	233,580	-0.83	602.8
Jackson Co. GIS	235,438	-0.04	636.1
Oregon GEO	235,699	0.07	177.2

(*) Benchmark boundary derived from 2000 Census blocks with SDUNI = 08040.
(**) Average distance from vertices of a boundary to the benchmark boundary.

To test what difference this mismatch could make, we allocated commonly used SF1 block-level data and SF3 blockgroup-level data to three District boundary configurations using simple areal interpolation and point-in-polygon aggregation. While these methods are simplistic: the former assumes an even spatial distribution of source zones' characteristics, the latter assigns data to a target zone if a source zone's centroid falls inside it, they are quick to implement. For one boundary configuration (from the District GIS), we also used "intelligent interpolation": a combination of dasymetric areal interpolation and expert judgment (these data were used for enrollment forecasts). We then compared the resulting estimates of population and housing characteristics with 2000 Census STP2 tabulations for Medford SD ("Census benchmarks"). The results are presented in tables 2 and 3.

Table 2. Attribute Error by Source of District's Boundary and Interpolation Method

	-----------------SF1 data-------------------			----------------SF3 data----------------				
Data source	**Population**	**HU**	**Households**	**SFHU**	**MFHU**	**MHO**	***Total HU***	*Comments*
NCES Census 2000	76,667	31,615	29,928	22,530	6,710	2,375	31,615	Benchmarks from STP2
Medford SD (PRC)	76,777	31,426	29,955	22,430	6,706	2,290	31,426	"Intelligent interpolation"
Medford SD GIS	76,612	31,364	29,897	22,184	6,858	2,635	31,676	Areal interpolation
Jackson Co. GIS	76,632	31,370	29,903	22,316	6,880	2,660	31,856	Areal interpolation
Oregon GEO	76,657	31,394	29,923	22,344	6,877	2,674	31,896	Areal interpolation
	----------------*Percent difference with Census benchmarks*--------------							
Medford SD (PRC)	0.1	-0.6	0.1	-0.4	-0.1	-3.6	-0.6	
Medford SD GIS	-0.1	-0.8	-0.1	-1.5	2.2	10.9	0.2	
Jackson Co. GIS	0.0	-0.8	-0.1	-0.9	2.5	12.0	0.8	
Oregon GEO	0.0	-0.7	0.0	-0.8	2.5	12.6	0.9	

Key: HU = Housing units; SFHU = Single-family units; MFHU = Multi-family units; MHO = Manufactured homes.

The most obvious observation following from the table is that even simple areal interpolation of SF1 data from Census blocks produces District totals practically identical to the benchmarks (differences of less than 1 percent). PRC's "intelligent interpolation", while methodologically more sound, makes no noticeable difference for block data interpolation at the District level. This result is important in practical terms: it may take days and weeks to produce data using the "intelligent interpolation" method, while simple areal interpolation would take just a few hours. However, in our study the data had to be produced for sub-areas (ESAAs) as well. In such cases, as we will argue in the following section of the paper, the "intelligent interpolation" method is worth the effort.

As expected, SF3 blockgroup data processing has produced larger errors than interpolation from SF1 block data: an assumption of even spatial distribution of demographic characteristics inside source units becomes less reliable as their size and internal heterogeneity increase. Here, "intelligent interpolation" delivers far better results relatively to simple areal interpolation.

Although, in theory, variations in the amount of error between characteristics should be expected, it is unclear why population and household counts came a lot closer to the benchmarks than either SF1 or SF3 housing characteristics. In fact, some SF3 housing data differ by as much as 12.6 percent from the benchmarks, a high percentage by any measure. Counts of Multi-family units and Manufactured homes display particularly high discrepancies with benchmarks, perhaps due to their highly uneven patterns of spatial distribution in most areas.

Data interpolation using aggregation of source zones' centroids (point-in-polygon method) is the least time-consuming of the methods discussed here. As evident in Table 3, block centroid aggregation of SF1 data to the District boundaries works almost as well as simple areal interpolation from blocks, although slightly worse than "intelligent interpolation" presented in Table 2: it has produced an average error of only 0.3 percent. Not surprisingly, blockgroup centroid aggregation produces larger errors: an average of 2.5 percent, up to 6.7 percent for counts of single-family housing units.

Table 3. Attribute Error of Point-in-Polygon Aggregation by Source of District's Boundary

	------------------SF1 data--------------------			----------------SF3 data-------------------			
Data source	**Population**	**HU**	**Households**	**SFHU**	**MFHU**	**MHO**	***Total HU***
NCES Census 2000	76,667	31,615	29,928	22,530	6,710	2,375	31,615
Medford SD GIS	76,503	31,321	29,857	21,700	6,692	2,324	30,716
Jackson Co. GIS	76,627	31,374	29,906	21,021	6,655	2,278	29,954
Oregon GEO	76,667	31,399	29,928	21,700	6,692	2,324	30,716
	----------------*Percent difference with Census benchmarks*---------------						
Medford SD GIS	-0.2	-0.9	-0.2	-3.7	-0.3	-2.1	-2.8
Jackson Co. GIS	-0.1	-0.8	-0.1	-6.7	-0.8	-4.1	-5.3
Oregon GEO	0.0	-0.7	0.0	-3.7	-0.3	-2.1	-2.8

Key: HU = Housing units; SFHU = Single-family units; MFHU = Multi-family units; MHO = Manufactured homes.

Overall, it appears that at the District level of analysis errors in the boundary configuration have had no practical effect on accuracy of aggregated SF1 attribute data. Errors of SF3 data interpolation (aggregation) are higher and seem to justify a more sophisticated approach to data processing, such as "intelligent interpolation". We propose to use a ratio of average population of source zones and target zones as a simple and practical way to choose data interpolation method. In this study, the ratio of average population of blocks with centroids inside the District, and the District's population is 0.0005. For block groups, the ratio is 50 times higher: 0.025[4]. It seems reasonable to assume that the higher the ratio, the higher the likelihood of misallocation of the data when simple methods of data interpolation are used. However, additional analysis will be necessary to fully establish parameters of the relationship.

[4] The focus here is not on the relative population size of blocks and block groups, but on the relationship between their sizes and that of the target unit(s).

5. ESAA-level (local) analysis

No true (benchmark) boundaries or benchmark demographic data were available at the ESAA level of analysis. While this has prevented us from establishing an exact amount of attribute errors in interpolated population and housing characteristics, it still allows to evaluate the magnitude of spatial errors and how much they vary from one school attendance area to another.

As Table 4 demonstrates, ESAAs differ in their population, area, and other characteristics. The ESAA population varies from approximately 3,000 people in Ruch ESAA to 8,700 people in Wilson ESAA. The smallest ESAA, Washington, occupies only 348 acres, while the largest, Ruch, encompasses 179,522 acres. Not surprisingly, the smallest ESAA is also the most urban, with 15.5 persons per net residential acre and 193 feet of streets per gross acre, while the largest ESAA is the most rural, with only 0.2 persons per net residential acre and 14 feet of streets per gross acre.

Table 4. Population and Spatial Characteristics of Elementary School Attendance Areas

Target Units	Acres	Total population	Net population density, per acre	Street density, feet/acre	Number of vertices
ESAAs:					
Abraham Lincoln	6563.1	4,900	2.4	26.5	126
Griffin Creek	10722.8	5,958	1.0	26.1	170
Hoover	6204.8	6,463	2.2	42.9	138
Howard	707.6	5,148	8.5	124.4	89
Jackson	765.8	5,175	11.3	166.3	53
Jacksonville	15794.9	4,710	0.9	25.3	129
Jefferson	1357.6	5,135	6.5	136.7	114
Kennedy	2629.7	5,978	4.6	57.7	86
Lone Pine	1114.2	6,470	5.9	131.8	101
Oak Grove	3316.1	4,538	4.0	46.4	128
Roosevelt	749.8	5,819	9.6	152.8	92
Ruch	179521.5	3,078	0.2	14.1	151
Washington	347.8	4,701	15.5	193.1	68
Wilson	3784.5	8,704	8.2	79.8	166
District:	233580.1	76,777	1.9	21.3	408

The number of vertices (points determining location and shape of a boundary line) can be considered a measure of complexity of the boundary: when a boundary segment is a straight line, only two vertices are necessary to define its location, shape, and length. However, when a segment is bent or irregular, several vertices are needed to approximate its shape. Boundary complexity tend to increase with area's size, although there are variations in this relationship: the largest ESAA (Ruch) is only 3rd as far as the number of vertices is concerned, and the smallest ESAA (Washington) has the 2nd lowest number of vertices.

This study uses several measures of spatial error. Two of them were first proposed by Simpson (2002): *Degree of Hierarchy* and *Degree of Fit*. These measures quantify the amount of error associated with interpolation of source units (in our study, 2000 Census blocks) to target units (the District as a whole; individual ESAA). Degree of Hierarchy (DH) is a ratio of the number of source units (zones) that completely fit into a target zone, and the total number of source units (zones). The higher the ratio, the lower the mismatch between source and target zones, and the higher the accuracy of resulting estimates. We chose to express it as percent, to be comparable with the second measure, Degree of Fit (DF). The latter evaluates similarity between source and target zones[5]. The higher the value, the lower the error. The first two columns of Table 5 display these measures by ESAA.

Table 5. Measures of Spatial Error by Elementary School Attendance Area

Target Units	Degree of hierarchy (C1)	Degree of fit (C2)	Average distance error (C3)	Percent in split source units				Percent in split source units with Interpolation coefficient < 0.5			
				Area (C5)	Persons (C6)	Households (C7)	Housing Units (C8)	Area (C9)	Persons (C10)	Households (C11)	Housing Units (C12)
ESAAs:											
Abraham Lincoln	55.4	75.1	132.69	46.3	44.8	44.8	44.5	7.8	5.6	5.5	5.5
Griffin Creek	68.1	83.0	428.76	76.8	32.2	32.5	32.9	2.6	3.9	4.1	4.2
Hoover	63.4	79.2	167.13	60.1	48.7	49.4	49.8	2.1	2.5	2.5	2.5
Howard	76.4	85.9	105.55	27.6	32.1	36.9	37.2	1.2	1.0	1.0	1.0
Jackson	71.1	89.6	38.18	35.1	26.1	25.3	25.5	1.8	1.3	1.2	1.2
Jacksonville	83.7	91.2	530.85	50.6	26.5	24.3	23.9	16.0	6.7	6.1	6.1
Jefferson	72.7	88.3	84.42	28.4	20.1	18.9	18.5	3.2	2.9	3.0	3.0
Kennedy	58.8	79.1	120.37	41.3	41.8	39.9	39.8	5.3	3.0	2.8	2.9
Lone Pine	47.9	80.5	76.17	48.5	42.1	42.0	42.1	2.4	2.6	2.7	2.8
Oak Grove	52.7	69.3	157.71	48.6	33.3	34.7	34.4	7.5	8.6	8.5	8.6
Roosevelt	75.5	85.9	106.02	32.5	20.9	22.6	22.4	0.6	0.7	0.5	0.6
Ruch	85.0	90.3	491.34	15.8	15.3	14.3	14.0	5.9	5.4	5.2	5.1
Washington	68.9	87.8	76.86	24.5	22.4	23.0	23.1	1.2	1.3	1.1	1.1
Wilson	61.9	76.6	98.63	34.7	28.6	29.4	29.0	8.1	9.2	9.8	9.5
District:	88.4	94.2	263.9	38.6	31.9	31.9	31.8	6.2	4.0	4.0	4.0

As evident from the table, almost 9 out of 10 Census blocks interpolated to the District completely nest into it (DH 88.4), and there is a high degree of similarity between the source

[5] It calculates the ratio between the sum of each source zone's highest (maximum) interpolation weight (coefficient), and the total number of source zones. In simple areal interpolation, the weight is the proportion of the source zone's area in the total area of the target zone. If there is no split , the weight equals 1 and the measure equals 100.

zones (blocks) and the District (DF 94.5). This support the conclusion of the earlier analysis: a fairly low spatial error is associated with SF1 block data interpolation at the District level.

However, once the analysis moves a step down geographical hierarchy, to ESAAs, and source units become more likely to be split by boundaries of target units, higher spatial errors are observed. Degree of Hierarchy ranges from 47.9 in Lone Pine ESAA (less than half of interpolated blocks fit completely (nest) into the ESAAs) to 85.0 in Ruch ESAA (8.5 out of 10 blocks nest into it). In other words, spatial error resulting from both positional errors and legitimate mismatch of the boundaries affects more than a half of blocks and related variables interpolated into Lone Pine ESAA. At least two more ESAAs show similarly high levels of error. Degree of Fit follows a similar, although not identical pattern: overall, it increases with an increase of DH. The measure varies from 69.3 in Oak Grove ESAA to 91.2 in Jacksonville ESAA.

An interesting pattern emerges once spatial distribution of the two error measures are displayed on a map (Figures 2 & 3). On both maps, there is an obvious cluster of low values in the north-eastern corner of the map, in the City of Medford and its immediate vicinity. To the contrary, two ESAAs in the rural and mountainous southwestern corner of the study area show high(est) values. It appears that block data interpolation in more densely populated urban and suburban areas produces higher levels of spatial error than interpolation from rural, sparsely-populated blocks. This finding seems counterintuitive at first: in cities, street locations and other TIGER features are well-known and have been mapped for decades, which should result in more accurate representation of block boundaries based on these features. However, if we take into consideration the reality of boundary mismatch, the results start making sense. With most administrative boundaries drawn along street centerlines or tax lot (property) boundaries, even a small misalignment of Census blocks and, in our case, ESAAs, would result in slivers. The number and relative area of slivers are taken into account by both measures of spatial error discussed here. When boundaries of source and target zones are misaligned, more slivers would result from a polygon overlay of source and target zones in areas with relatively denser street pattern.

Figure 2. Degree of Hierarchy by ESAA

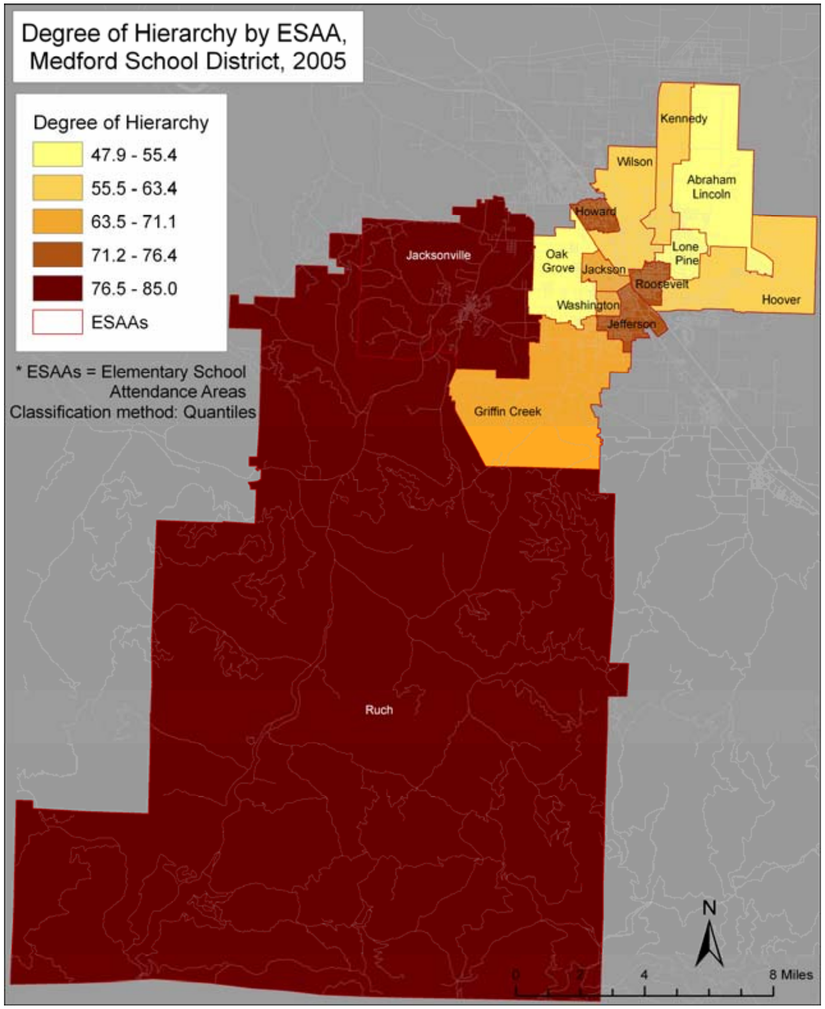
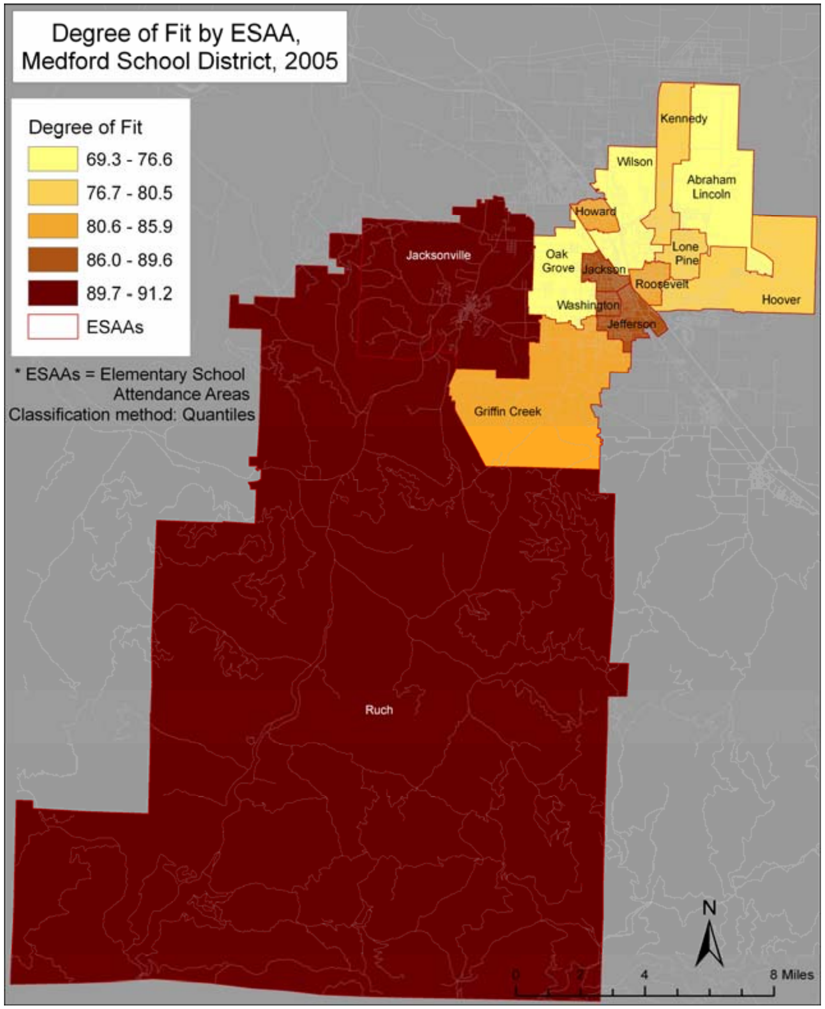


Figure 3. Degree of Fit by ESAA



To investigate the amount of spatial error more fully, several additional measures were calculated (Table 5). *Average Distance Error* (C3) evaluates the linear extent of spatial mismatch by computing an average of shortest distances between each vertex of a target zone boundary and its closest source zone boundary[6]; the higher the distance, the larger the error.

The next 8 measures quantify effects of boundary mismatch on several key ESAA characteristics (area, population, households, and housing units). They show *percentage of a target zone's characteristic derived from split source units*: Census blocks split by ESAA boundaries. For example, 60 percent of Hoover ESAA's total area (variable C5) is aggregated from split Census blocks, while only 15.8 percent of Ruch ESAA's area comes from such blocks. In Hoover ESAA, 49 percent of persons and households and 50 percent of housing units are found in split blocks, while only 14-15 percent of these characteristics come from split blocks in Ruch ESAA[7]. While we cannot claim that ESAA population and housing characteristics resulting from interpolation are 14, 49, or 60 percent wrong, we propose to use these numbers as *measures of uncertainty* associated with the variables. It seems reasonable that the higher the share of a variable that is calculated from split blocks, the larger associated potential interpolation error, and the less reliable (more uncertain) resulting attributes become.

Variables C9 through C12 are similar to C5 through C8 in that they also identify percentage of a characteristic in split source areas. However, the criteria are narrower here: only split blocks that are less than 50 percent of their original area are used for calculation. The limit is somewhat arbitrary and chosen under an assumption that an interpolation coefficient 50 percent or higher would result in more reliable interpolated characteristics.

It is obvious from a quick analysis of Table 5 that spatial errors and spatial uncertainty vary considerably by ESAA, and some values are quite large. Overall, one third of all persons, households, and housing units were derived from split source zones at the ESAA level of analysis. For individual ESAAs, between 14 and 50 percent of these characteristics came from

[6] This is different from the District-wide analysis where the distance was measured between vertices of a given District boundary and the benchmark boundary.

[7] "Intelligent interpolation" method was used to derive characteristics of target units. The Interpolation Coefficient is a proportion of a source zone's area in the target zone's area where area calculations are limited to residential land only (see Appendix).

split blocks. Since any interpolation method requires assumptions about distribution of Census characteristics within source zones, even "intelligent interpolation" weakens reliability of the resulting data. In our study, for 10 out of 14 ESAAs, more than a quarter of their population and housing characteristics were affected by interpolation error; for one half of ESSAs, more than 30 percent were affected.

Predictably, variables C9 through C12 are not as high as variables C5 through C8: only a subset of split blocks and their characteristics is considered. Still, in some cases share of ESAA characteristics derived from block pieces less than a half of their original size raises up to 10-16 percent.

To evaluate if there are any commonalities between the measures of spatial error, correlation coefficients (Spearman's rho) were calculated for pairs of error measures as well as other ESAA characteristics (see Table 6).

Table 6. Correlation coefficients (Spearman's rho)

Variable	Name	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
Degree of hierarchy	C1	1.000														
Degree of fit	C2	0.854**	1.000													
Average distance error	C3	0.240	0.000	1.000												
Number of vertices	C4	-0.011	-0.180	0.679**	1.000											
% area in split blocks	C5	-0.486	-0.389	0.402	0.407	1.000										
% persons in split blocks	C6	-0.758**	-0.744**	0.130	0.125	0.705**	1.000									
% households in split blocks	C7	-0.710**	-0.713**	0.029	0.020	0.600*	0.982**	1.000								
% housing units in split blocks	C8	-0.710**	-0.713**	0.029	0.020	0.600*	0.982**	1.000**	1.000							
% area in split blocks, IC < 0.5	C9	-0.196	-0.224	0.484	0.583*	0.308	0.161	0.059	0.059	1.000						
% persons in split blocks, IC < 0.5	C10	-0.295	-0.339	0.482	0.673**	0.343	0.180	0.064	0.064	0.961**	1.000					
% households in split blocks, IC < 0.5	C11	-0.266	-0.304	0.459	0.688**	0.336	0.143	0.029	0.029	0.959**	0.995**	1.000				
% housing units in split blocks, IC < 0.5	C12	-0.266	-0.304	0.459	0.688**	0.336	0.143	0.029	0.029	0.959**	0.995**	1.000**	1.000			
Street density, feet per acre	C13	-0.068	0.092	-0.908**	-0.763**	-0.437	-0.270	-0.182	-0.182	-0.673**	-0.660*	-0.644*	-0.644*	1.000		
Area (in acres)	C14	0.042	-0.035	0.811**	0.824**	0.451	0.160	0.051	0.051	0.761**	0.748**	0.749**	0.749**	-0.934**	1.000	
Net population density per acre	C15	-0.029	0.031	-0.873**	-0.741**	-0.499	-0.266	-0.160	-0.160	-0.629*	-0.607*	-0.600*	-0.600*	0.960**	-0.930**	1.000

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Many measures of spatial error are significantly correlated. As expected, Degree of Hierarchy (C1) and Degree of Fit (C2) are positively and significantly correlated with each other, and are negatively correlated with nearly all other spatial measures: C1 and C2 get higher when boundaries are more simple (closer to straight lines), fewer blocks are split, and fewer and smaller slivers result. This negative correlation becomes highly significant with measures C6, C7, and C8: percentages of a target zone's persons, households, and housing units derived from split source units. However, C1 and C2 quantify somewhat different aspects of spatial mismatch: C1 tends to be higher where blocks are larger, less urban, and less densely populated while C2 appears to be lower in these areas (but the relationships are not significant).

Average Distance Error (C3) is higher where target zones' boundaries are more complex (C4), blocks are larger (C14), less urban (C13), and less densely populated (C15); these relationships are significant. This makes sense: with many block boundaries drawn along straight street centerlines, more irregular ESAA boundaries would produce a bigger mismatch. Blocks in rural and mountainous areas are often very large and irregular, increasing the likelihood that other boundaries would not match with them. While correlation between C3 and measures C5-C12 is positive, as expected, it is not significant.

Measure C4 (number of vertices) behaves in a similar manner, except that its correlation gets much stronger (and significant) with measures C9 through C12. More complex ESAA boundaries are less likely to follow streets and more likely to split blocks into slivers of larger size.

Not surprisingly, measures C5-C8 are positively and significantly correlated amongst themselves, and so are measures C9-C12: measures in each of these groups were calculated in the same manner. Additionally, the latter group is negatively and significantly correlated with street density and net population density: areal size of slivers as a proportion of target zones tend to be lower in more urban areas resulting in lower values for measures C9-C12.

6. Discussion and conclusions

This study has examined a real-life example of small area demographic analysis and some of its challenges commonly encountered by applied demographers. Choices with regard to spatial data sources and geoprocessing methods of data development have been discussed. The analyses of spatial errors resulting from boundary mismatch have been conducted at two geographical levels: regional, for the District as a whole, and local, for its elementary schools attendance areas (ESAAs). At the regional level, we have been able to quantify both the extent of positional errors and the effect they have on population and housing characteristics vis-à-vis benchmark data when interpolated from source to target zones using several common geoprocessing methods. At the local level, no benchmark data was available; however, we have been able to investigate variations by ESAA of different measures of spatial errors.

The analysis presented here supports the authors' claim that applied demographers should seriously consider issues of positional accuracy when conducting small-area analysis. The predominant expectation that errors of spatial interpolation just "wash out" when the data is aggregated is not supported by our results. Only block data interpolation, and only at the District level has produced negligible errors in population and housing characteristics. With blockgroup data, errors have increased to 10-13 percent depending on the interpolation method, data source, and the variable under consideration; this suggests that "intelligent interpolation" becomes a necessity for this scale of analysis.

These results, however, are likely to vary with the size of a study area and its population. Areas comparable with ESAAs (an urban neighborhood, a 10-minute walking zone around a store) may need to use "intelligent interpolation" even for block data, while much larger areas (a big school district, a part of a large county, an area within an urban growth boundary of a large city) might "get away" with simpler methods of spatial data aggregation. We have proposed to use a simple measure of relative population sizes of source and target zones (a ratio of average population of source zones and a target zone) to determine what method is more appropriate. In our study, the ratio for blocks vis-à-vis the District was 0.0005; for block groups it was 0.025. Incidentally, the ratios calculated at the ESAA level (average population of blocks inside an ESAA to ESAA's total population) were a lot closer to the latter number than the former: while varying from 0.020

to 0.009, an average of the ESAA ratios was 0.014. This lands additional support to our original choice of the "intelligent interpolation" method for data development at the ESAA level.

Several measures of spatial error were examined at the ESAA level. They evaluate how well source and target zones fit (nest) into each other, how similar their spatial configurations and boundary shape and locations are, and what share of a target zone's Census characteristics is affected by uncertainty associated with mismatch between source and target zones. While most measures are significantly correlated, they nevertheless describe somewhat different aspects of spatial error. The measures show substantial variation by ESAA providing additional evidence that errors do not tend to "wash out" at the local level.

The analysis shows that up to 50 percent of population and housing characteristics in a given ESAA can be derived from split source zones; on average, up to 32 percent of ESAAs' Census characteristics are affected by boundary mismatch. While we do not know how much error this introduces into interpolated counts without comparing the measures to Census benchmarks, the amount of *uncertainty* associated with them appears non-trivial and worth further investigation.

Lower levels of both Degree of Hierarchy and Degree of Fit in most of urban Medford ESAAs challenge a common assumption that spatial errors should be lower in more developed and denser populated areas due to more reliable data infrastructure. When boundaries of source and target zones (i.e., Census blocks and ESAAs) are misaligned, errors appear higher in more urban areas than in rural and less populated ones.

While this study has not directly measured the effect of varying geographic scales at which spatial data was originally developed on subsequent interpolation errors, its results suggest the need for Census TIGER data to be developed at a scale more common to municipal and other local government application (1:24,000 versus current 1: 100,000). This may be especially important once small-area American Community Survey data will become available. It would greatly help data users if they would have to deal with fewer than three types of error that are likely to be associated with ACS data: common survey errors (sample and non-sample), errors of averaging data over time (3 to 5-year time period), and spatial errors of boundary misalignment.

There are several *limitations* to the study and its conclusions. This is a case study, which raises the issue of applicability of its results to other situations. It would make it stronger if several school districts were analyzed, and the number of ESAAs increased from the current N of 14 to at least 50. As noted above, due to data limitations we could evaluate the degree of positional error associated with boundary mismatch at the ESAA level, but not the effect of positional error on estimated population and housing characteristics. We also could not separate the effects of a "legitimate mismatch" (a true overlap of the boundaries) at the ESAA level from a misalignment of the boundaries (a true positional error); to do so, "benchmark" ESAA boundaries are necessary. Also, we estimated errors at the polygon level (District, ESAA), but not at the boundary level. It has been shown by (Gaeuman et al. 2005) that significant local positional error can be associated with local boundary forms (sinuosity).

Nevertheless, we hope that this study advances an understanding among applied demographers of spatial errors and their importance for accuracy of estimated population and housing characteristics. We believe that, similarly to reporting confidentiality intervals for sample attribute data, measures of spatial error and/or spatial uncertainty should accompany analyses involving spatial interpolation of the data, be they of applied or academic nature.

References

Chrisman, N. R., 1987, "The Accuracy of Map Overlays: A Reassessment." *Landscape and Urban Planning* 14, 427–39.

Eicher, C. and Brewer, C., 2001, "Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation." *Cartography and Geographic Information Science,* 28, pp. 125-138.

Flowerdew, R. and Green, M., 1994, "Areal Interpolation and Types of Data." In *Spatial Analysis and GIS,* A.S. Fotheringham and P.A. Rogerson (Eds), pp. 121–145 (London: Taylor & Francis).

Gaeuman, D., Symanzik, J., and Schmidt, J.C., 2005, "A Map Overlay Error Model Based on Boundary Geometry." *Geographical Analysis,* 37, pp. 350–369

Goodchild, M. F. and Lam, N. S., 1980, "Areal Interpolation: a Variant of the Traditional Spatial Problem." *Geo-Processing,* 1, pp. 297-312

Goodchild, M.F., Anselin, L. and Deichmann, U., 1993, "A Framework for the Areal Interpretation of Socioeconomic Data." *Environment and Planning A*, 25, pp. 383-397

Gregory, I.N., 2002, "The Accuracy of Areal Interpolation Techniques: Standardising 19th and 20th Century Census Data to Allow Long-Term Comparisons." *Computers, Environment and Urban Systems*, 26, pp. 293–314

Gregory, I.N., and Ell, P.S., 2006, "Error-sensitive Historical GIS: Identifying Areal Interpolation Errors in Time-series Data." *International Journal of Geographical Information Science*, 20 (2), pp. 135–152

Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W., 2005. *Geographic Information Systems and Science*. 2nd Ed. John Wiley & Sons and ESRI Press.

Reibel, M., and Bufalino, M.E., 2005, "Street-Weighted Interpolation Techniques for Demographic Count Estimation In Incompatible Zone Systems." *Environment and Planning A*, 37, pp. 127–39.

Sadahiro, Y., 1999, "Accuracy of Areal Interpolation: a Comparison of Alternative Methods." *Journal of Geographical Systems*, 1, pp. 323–346.

Sadahiro, Y., 2000, "Accuracy of Count Data Estimated by the Point-in-Polygon Method." *Geographical Analysis*, 32, pp. 64-89.

Schlossberg, M., 2003, "GIS, the US Census and Neighbourhood Scale Analysis." *Planning, Practice & Research*, 18 (2–3), pp. 213–217

Sharkova, I. V., Radin, K., and Lycan R., 2005, "Medford School District 549 School Enrollment Forecast, 2005 to 2015." Population Research Center, Portland State University.

Simpson, L., 2002, "Geography Conversion Tables: a Framework for Conversion of Data between Geographical Units." *International Journal of Population Geography*, 8, pp. 69–82.

Smith, S., 1987, "Tests of Accuracy and Bias for County Population Projections." *Journal of the American Statistical Association*, 82, pp. 991-1003.

Smith, S., and Cody, S., 2004, "An Evaluation of Population Estimates in Florida: April 1, 2000." *Population Research and Policy Review*, 23, pp. 1-24.

Tayman, J., Schafer, E., and Carter, L., 1998, "The Role of Population Size in the Determination and Prediction of Population Forecast Errors: An Evaluation Using Confidence Intervals For Subcounty Areas." *Population Research and Policy Review*, 17, pp. 1-20.

Xie,Y., 1995, "The Overlaid Network Algorithms for Areal Interpolation Problem." *Computers, Environment and Urban Systems*, 19, pp. 287- 306

Appendix. Interpolation of Census data[8]

Interpolating Census demographic and housing data to 'custom' boundaries, such as attendance areas, is the most complex of the GIS methods utilized for the forecast. In addition to standard GIS geoprocessing techniques, PRC used "dasymetric areal interpolation" for most of the Census data, though interpolation of housing units by structure type, for 2000 Census data, required ad-hoc uses of primarily tax lot attributes.

Dasymetric mapping relies on ancillary geospatial data to find the right locations for some other, geographically aggregated, data set. In the Medford case, Census blocks are the primary level of aggregation, and zoning, for example, is one ancillary, "helper" geospatial dataset that helped us find more realistic distributions of data corresponding to a given block. Blocks themselves are fairly small units and in most cases provide the necessary level of disaggregation as-is. However, sometimes, particularly with large blocks, it is possible and useful to identify portions that are not likely to be populated. It is not likely, for example, that people live on commercially zoned land. In such a case, the portion of a given Census block that intersects the commercial zone is 'erased' and it is assumed that people live in the non-commercial part. This is the "dasymetric" part of the method: one ancillary geospatial dataset, the zoning, is used to find more realistic locations for another geographically aggregated dataset, population by Census blocks.

A fairly exhaustive list of helper-data was used to develop a spatial theme distinguishing residential areas from non-residential areas. The theme was developed to be inclusive rather than precise; thus, for example, in most cases zoning was used rather than parcel-level attributes.

Once Census blocks are adjusted by the residential theme, attribute data can be interpolated to custom boundaries – the attendance areas. The attendance areas are used to select blocks that fall completely within a given attendance area boundary, or to cut blocks, as if using a cookie-cutter, when they do not fall completely within. This is where areal interpolation is used. Here it is assumed that block-level attribute data are evenly distributed throughout the residential portion of a given census block. For those blocks that are cut by attendance area boundaries, attribute values, such as population, are weighted by the proportion of the block that ends up in a given attendance area. For example, if an attendance area boundary cuts through the middle of a Census block, half the population is allocated to one side of the boundary while the other half is allocated to the other side of the boundary.

After adjusting for residential versus non-residential areas (dasymetric mapping) and areal interpolation, Census variables can then be summed by attendance areas and the District as whole.

The two primary sources of error for this method arise from spatial mismatches between the Census block boundaries and the District attendance area boundaries, and the accuracy of the identification of residential versus non-residential areas. For the former, the spatial representation of a given block boundary can typically be 50-100 feet off from more precise tax lot themes, for example, which might place some of the block's population in one attendance area when it might belong to another. Beyond using matched 1990 and 2000 Census blocks, and

[8] From (Sharkova, Radin, and Lycan 2005)

visual scans for locations that would obviously pose a problem, no special methods were developed to eliminate this type of error. For the latter, identification of residential areas erred on the side of caution such that no population could possibly go uncounted.

Finally, an exception to the above method was necessary to interpolate 2000 Census housing by structure type to attendance areas. This variable is available at the block-level for 1990, yet only at the block group level for 2000. Here the tax lot attribute "building code" was used to develop interpolation coefficients. Tax lots were selected by building codes that identify residential units and the locations were coded for number of units and unit-type (SFR, MFR or 'Other'). Interpolation coefficients, by which the block group-level variables (SFR, MFR, and Other) could be multiplied, were then derived by dividing the number of units (or "unit factor") by type by tax lot, by the sum of unit factor by type by block group. In other words, housing units by type by tax lot, divided by the sum of housing units by type by tax lot by block group geography, were used to allocate block group-level housing units by type to more realistic locations with a given block group. Once this is completed, the units can then be spatially joined to attendance areas. Ultimately, this interpolation was used only as an adjustment coefficient to reallocated block-level housing unit data at the attendance area level, not as the basis for the number of units.