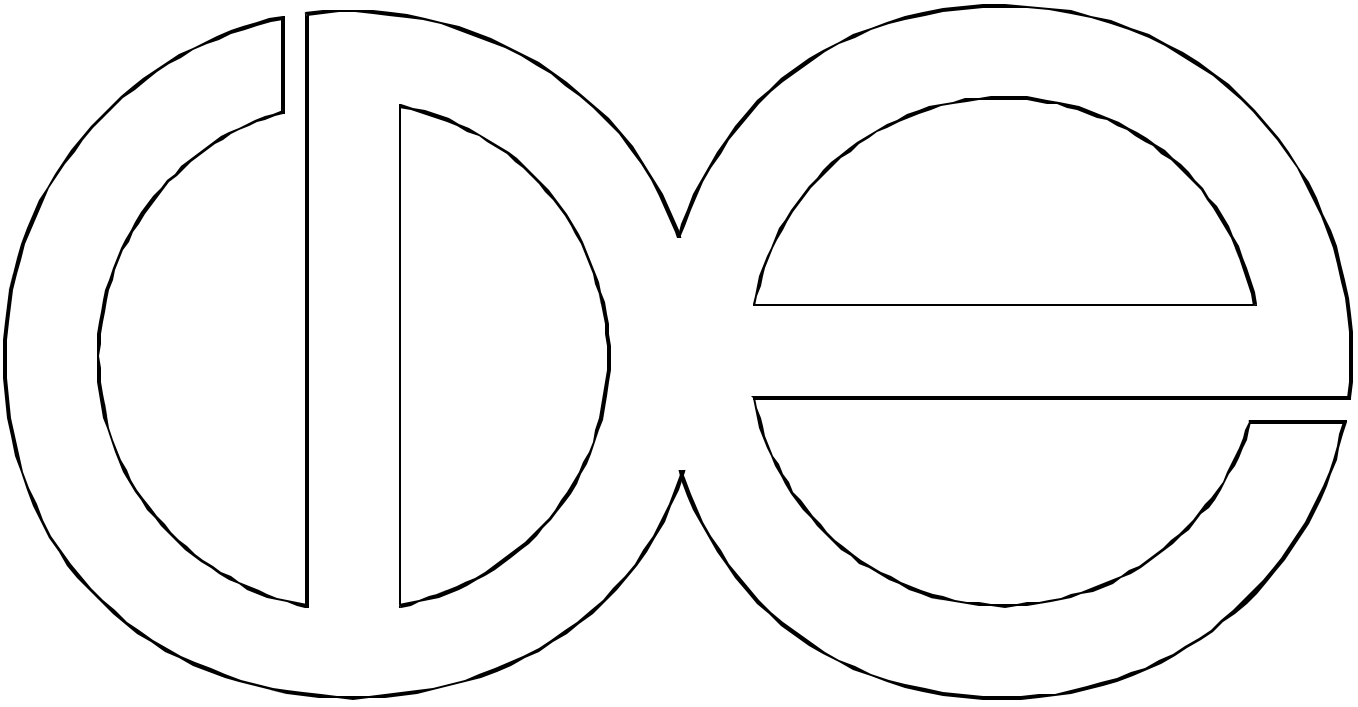
Center for Demography and Ecology

University of Wisconsin-Madison

When Census Geography Doesn't Work: Using Ancillary Information to Improve the Spatial Interpolation of Demographic Data

Paul R. Voss
David D. Long
Roger B. Hammer

CDE Working Paper No. 99-26



Environment & Planning A

Manuscript submitted 9/10/99

When Census Geography Doesn't Work: Using Ancillary Information to Improve the Spatial Interpolation of Demographic Data

Paul R. Voss

David D. Long

Roger B. Hammer

Department of Rural Sociology

University of Wisconsin-Madison

When Census Geography Doesn't Work:
Using Ancillary Information to Improve the Spatial Interpolation of Demographic Data

Abstract. This paper introduces two new spatial interpolation techniques that utilize the network of road segments and the resulting nodes to allocate aggregated demographic characteristics from one type of geographic boundaries (i.e., the geographic hierarchy of the U.S. Census) to another (e.g. watersheds) under conditions of "spatial incongruity." Spatial incongruity arises when spatially aggregated data are available for one set of geographic areal units but not the areal units of primary interest. Spatial incongruity presents a major obstacle to the integration of social and natural science data and consequently places limitations on interdisciplinary research efforts. In the natural sciences the geographic units of analysis frequently are areas defined by land use, land cover, soil type, watershed boundaries, and a variety of other biophysical and geophysical features. Given that census geography and its concomitant demographic data seldom correspond exactly to these areas, combining the data from different disciplines and disparate units of analysis becomes a crucial function. The road segment length interpolation method presented in this paper improves upon areal weighting, the most common method used to allocate characteristics from one geographic system to another, in limited circumstances while the nodal count method represents a substantial improvement.

Introduction

"Spatial incongruity" arises when spatially aggregated data are available for one set of geographic areal units but not the areal units of primary interest. The problem of spatial incongruity commonly arises in the context of interdisciplinary research, and is an impediment to such research despite the promise of geographic information systems (GIS) to provide integrated data structures. Although GIS has facilitated the utilization of spatial databases with incongruous boundaries through the basic

overlay process, the lack of correspondence often necessitates the use of spatial interpolation techniques in order to examine relationships between variables drawn from disparate units of analysis. The spatial incongruity problem is familiar to applied demographers addressing a research question that requires the tabulation of data from the decennial census, available for blocks, block groups, and census tracts, by customized geographic areas such as service territories, trade areas, or utility districts. This problem continues to present a major obstacle to the integration of social and natural science data and consequently places limitations on interdisciplinary research efforts. Antle and Just (1992; p. 314) maintain that a "major obstacle to integration of knowledge from various disciplines for informed policy analysis is an integrated data base." In the natural sciences the geographic units of analysis frequently are areas defined by land use, land cover, soil type, watershed boundaries, and a variety of other biophysical and geophysical features. Given that census geography and its concomitant demographic data seldom correspond exactly to these areas, combining the data from different disciplines and disparate units of analysis becomes a crucial function.

Various inferential techniques that attempt to reconcile the spatial incongruity between different spatial units of analysis have emerged. A common method of referencing the geography in this situation terms the geographic units in which data are available "source" geography, while "target" geography refers to the spatially incongruous units in which the data are needed (Goodchild and Lam 1980). Among the private data firms that tabulate demographic information for custom target areas, the most common approach involves rules of inclusion or exclusion based on the boundaries of a target geographic area and the centroids, or approximate centers, of the census source geography transected by those boundaries (Tordella 1987). This simple but crude technique can be characterized as centroid assignment. The term areal interpolation (Goodchild and Lam 1980) describes a variety of methods which generally apply a weight based on the area of intersection between source and target geographies in order to allocate characteristics to the target geography.

In this paper we introduce two closely related alternatives to centroid allocation and areal interpolation. These new interpolation techniques utilize the network of road segments and the resulting nodes located within the source and target geographies. The road segment and nodal interpolation methods have been developed and tested in a geographic information system environment. In order to implement these methods three conditions must be met: (1) use of data from the U.S. 1990 Census or another demographic data source using census defined geography, (2) use of the Census Bureau's Topologically Integrated Geographic Encoding and Referencing system (TIGER) as the geographic base file, and (3) allocation to target areas of interest that are not part of the census geographic hierarchy. In our test of these methods, we find that the road segment length interpolation method improves upon the areal weighting method in limited circumstances while the nodal count method represents a substantial improvement.

Spatial Interpolation

The problem of spatial incongruity has long been confronted by geographers, regional planners, and landscape ecologists. As noted, centroid assignment allocates the characteristics of a source polygon to a target polygon if the source polygon's centroid is located within the target polygon. Two general, but quite different, approaches to areal interpolation appear in the literature (for reviews of the literature see Lam 1983; and Flowerdew and Openshaw 1987). One approach, often referred to as "polygon overlay" (Markoff and Shapiro 1973) or somewhat more commonly as "areal weighting" (Flowerdew and Green 1994), combines source geography data weighted according to the area of the target geography, which they comprise. That is, the weights are determined by the extent of intersection between the source polygons and target polygons. This approach is greatly facilitated by basic GIS procedures that use functions for determining the area of intersection and assigning weights but has the disadvantage of assuming that the data of interest are distributed uniformly within the areas constituting the source

geography. A second approach, developed by Waldo Tobler (1979), fits a surface to the data in the source polygons and uses the surface to interpolate values for the target polygons. This latter approach has been used in several papers by British geographers Ian Bracken and David Martin (Bracken 1991; Bracken and Martin 1989, 1995; Martin 1996; Martin and Bracken 1991). Fitting a surface to the data for allocation to target geographies is itself a complex inferential process.

In recent years several papers by British geographers Robin Flowerdew and Mick Green have described an interesting new approach to the problem (Flowerdew and Green 1989, 1992, 1994; Flowerdew, Green, and Kehris 1991). Their method seeks to improve on simple areal interpolation by utilizing other relevant information regarding the target geography to improve the assignment of attributes to the target geography. The statistical concept behind their methods is based on an iterative expectation/maximization (EM) algorithm developed by Dempster, Laird, and Rubin (1977), a procedure originally designed to estimate missing data. Flowerdew and Green have adapted the EM method to address the spatial incongruity problem. Their method incorporates ancillary information for the target geography that is correlated with the characteristics of interest in the source geography. Flowerdew and Green do not formally compare the accuracy of their method to straight areal interpolation.

While our solution to the spatial incongruity problem incorporates ancillary information, it is more straightforward than the EM method and has certain features that make it superior to the statistical approach advocated by Flowerdew and Green. Their approach requires ancillary information for the target geography that is correlated with the characteristic of interest in the source geography. Our method uses either the length of road segments or the number of road nodes from the source geography. Second, in their method the relationship between the ancillary data and the characteristic of interest must be modeled correctly and in many cases must be

tested for linearity, possibly necessitating a more complex nonlinear specification.

Interpolation of Census Data

The U.S. decennial census is a massive undertaking that serves as the basis for political redistricting and as a basis for funding allocation and program administration. The census is designed to gather and report aggregated information for housing units, households, families, and individuals to support federal, state, county, city, and tribal government planning and policy making. For demographic data derived from the census, "small" geographic units of analysis commonly consist of statistical areas defined by the Census Bureau: blocks, block groups, tracts, and block numbering areas (U.S. Bureau of the Census 1991a). These units are geographically comprehensive and are linked to a prodigious amount of census data, principally in the series of Summary Tape Files 1 and 3 (U.S. Bureau of the Census 1991b and 1991c). Block groups are the smallest units of census geography for which the detailed "long-form" social and economic data from the census are tabulated while basic housing and population data are published for census blocks. In rural areas, these small-area census polygons are generally much larger geographically than their counterparts in urban areas. The large variation in the physical size and shape of rural blocks makes them an odd assortment of "building blocks" with which to make comparisons with non-census spatial units. That is, in rural areas the probability of census blocks nesting neatly within any non-census spatial units of interest is much lower than in urban areas. Moving up the hierarchy of census geography from blocks to block groups and block numbering areas compounds this problem. Spatial incongruity presents a greater problem in the study of rural areas and thus we have selected predominantly rural counties as our geography of interest.

In rural areas, the census block, the smallest unit in the census geographic

hierarchy for data tabulation purposes, does not correspond to the un-intersected city block found in urban areas but is geographically much more extensive. Census blocks, more than seven million of them in the 1990 Census, are simply polygons in the TIGER maps – polygons to which basic population and housing census information can be linked and mapped. However, there is additional information *within* census blocks, particularly in rural areas, that can be exploited to more accurately solve the problem of spatial incongruity. The TIGER line files include road segments (i.e., arcs). Some road segments penetrate census blocks but are not part of the line segments defining the census block boundaries. The road segments internal to census blocks include public and private roads, driveways, cul-de-sacs, and other access routes. Associated with these internal arcs are internal nodes, generally intersections and terminus points. A node is formally defined as a zero-dimensional object that is a topological junction of two or more links or chains, or an endpoint of a link or chain (U.S. Bureau of the Census, 1996, p. 1-7). Nodes are the markers in the TIGER Line Files that identify the intersection of lines (e.g., two roads) or the end of a line (or road). Figure 1 shows a sample block's road segment and node configurations. Given the association of housing units and their corresponding resident populations with road segments and nodes, we are able to use them in our alternatives to standard interpolation methods. Rather than applying an areal weight, our methods allocate demographic characteristics based on 1) the aggregate length of internal road segments or 2) the number of internal nodes. The road segment and node methods default to areal weighting in blocks with no internal road arcs or nodes.

[Figure 1 approx. here]

A Test of Alternative Interpolation Methods

Our interpolation approach assumes that the internal arcs representing access roads serve as proxies for the location of housing units and the resident population within a block or other polygon. That is, within a census block, population density is greater in areas with a high density of internal roads (and corresponding nodes) and lower in areas with a low density of internal roads. There are, of course, exceptions to this basic assumption but they are not particularly problematic, since our goal is to demonstrate in the aggregate that the use of internal roads and nodes provides a simple yet more reliable interpolation method than other existing approaches.

We do not create an estimate using Flowerdew and Green's EM method. Since the relationship between the ancillary target information and the source characteristic of interest must be carefully specified and modeled in the EM method, an objective test would be difficult. Although our method is probably easier to implement, there are certainly situations in which the EM method would yield more accurate results.

We chose Crawford County in southwestern Wisconsin to test these interpolation methods. Crawford is a primarily rural county containing six block numbering areas, 19 census block groups and 1,456 census blocks. Figure 2 shows portions of this geographic hierarchy and illustrates the significant variation in the shape and size of blocks, block groups and block numbering areas.

[Figure 2 approx. here]

From among the census blocks in Crawford County, we selected all "collection blocks" that were transected by a municipal (i.e., Minor Civil Division) boundary into two (or more) "tabulation blocks." Collection blocks are the small geographic polygons generally bounded by permanent, highly visible, physical features. They are used for census data collection by census enumerators. Frequently these

collection blocks are transected by an invisible political boundary. Before the data are tabulated, the Census Bureau inserts this boundary, splits the collection block into two or more tabulation blocks and correctly allocates the housing unit and population data from the collection block to the tabulation blocks. Using only collection blocks permitted us to ignore the municipal boundaries, treat these split collection blocks as single geographic entities and aggregate the number of persons and housing units. For our test we then transected these combined blocks with the municipal boundaries and estimated the number of persons for each of the constituent tabulation blocks using several interpolation methods (including road- and node-based methods). We then compared these estimates to the actual distribution of persons, as reported by the Census Bureau for each tabulation block, to evaluate the performance of each of the interpolation methods. In this test, the municipal boundaries dividing the collection blocks serve as a proxy for target geography boundaries that might conceivably split blocks. We illustrate this in Figure 3 by intersecting a watershed boundary with the census block groups. The watershed represents the "target" geographic unit for the census block group "source" data.

[Figure 3 approx. here]

A total of 116 collection blocks in Crawford County were transected by a municipal boundary and were thus suitable for the test. Because we selected collection blocks split by a minor civil division boundary, our sample is biased toward rural blocks. The county contains one small city with several dozen blocks, but only a small number of them are included in our sample. In addition, we elected to remove from our sample those collection blocks containing no housing units and/or no internal roads or nodes. This permitted us to perform each of the areal interpolation methods on the same sample of blocks. Since most of these blocks

were split into two (but often more than two) tabulation blocks, our final sample for the test consisted of 277 tabulation blocks. Ultimately we used the following four methods of interpolation:

1. Centroid Assignment. This method applied to block group polygons is commonly used in market research applications to define trade areas. When applied at the block level, the census block attributes (e.g., housing units or population) are assigned to whichever portion of the transected block contains the block's centroid or geographic center.
2. Areal Weighting. This is the traditional approach to areal interpolation that is built into the functionality of some GIS software. The block attributes are allocated to parts of the transected block based on the proportion of the block's total area contained within each part.
3. Road Segment Length. This method exploits the within-block road segment arc features in the TIGER line files. It allocates attributes to each part of a transected block based on the portion of the block's total internal road segment length located within each part.
4. Internal Node Counts. This method exploits the nodes of the within block road segment arc features in the TIGER line files. Census attributes are allocated to block parts according to the portion of the block's total internal nodes located in each part.

Results from the test of interpolation methods are summarized in Table 1. We use two measures of error to assess the accuracy of the interpolated estimates. The Mean Absolute Error (MAE) is the average number of persons that were incorrectly allocated to split blocks and the Mean Absolute Percent Error (MAPE) is the average proportion of persons that were incorrectly allocated to split block groups. To facilitate comparisons of error among the interpolation methods, we calculated a

ratio of the error for each method compared to the error of the areal weighting method. As expected, centroid assignment was the least accurate method, incorrectly allocating 14.5 persons on average and incorrectly allocating 25% of the population on average. The level of error for centroid assignment was 1.5 times greater than areal weighting in terms of the number of persons and 1.4 times greater in terms of the percentage of persons. Areal weighting, our comparison method, on average incorrectly allocated 9.5 persons or 17.9% of the block's population.

[Table 1 approx. here]

The performance of the road segment length method was slightly better than areal weighting in terms of the MAE, 8.6 with a ratio of 0.9, while it was comparable in terms of the MAPE, 18.3% with a ratio of 1.02. The test results indicate that the node count method for allocating population has the lowest error both in numeric (MAE) and proportional terms (MAPE). The error for this method was only 7.1 persons per block and 16.6% of the population per block. By exploiting the internal nodes located within census blocks in the TIGER file, this method afforded a 25% improvement over the conventional areal weighting method and a 51% improvement over centroid assignment in terms of the number of persons mis-allocated on average. In terms of the percentage of the population that was not correctly allocated, the error ratio of the internal node method to areal weighting was 0.93, representing a 7% improvement. The node count method also compared favorably to the road segment length method and to several combinations, taken as simple means, of the other interpolation methods (not shown).

The discrepancy in the areal weighting ratios (the numbers show in parentheses in Table 1) between the proportional level of error (MAPE) and the error in the number of persons (MAE) is a result of the heterogeneity in the population size of blocks and variation in the accuracy of the interpolation methods across this size

range. The road segment and node interpolation methods gained much of their predictive advantage in blocks containing larger populations. Thus, using ancillary information in the interpolation method has the greatest advantage among the more populous blocks, for which prediction accuracy may be more important.

Discussion

Although the test of our method demonstrates the efficacy of using road segment and node ancillary information to improve spatial interpolation, it has some obvious limitations. We assume that the county from which the blocks for the test were chosen is representative of rural counties elsewhere. The number of blocks selected for the test was relatively small, and they were predominantly, but not exclusively, rural. Regardless of these limitations, this method possesses an intuitive appeal: roads provide access to housing units, road segments and nodes indicate the location of housing units, and the vast majority of people live in housing units. Allocating housing and population attributes within a block using node counts improves upon allocation methods that assume housing and population are uniformly distributed within a block. The extent of improvement over the areal weighting method is substantial, suggesting that the method generally should work in rural areas.

The contribution of this method to interdisciplinary research is not only the more accurate block level interpolations it affords but also its ability to scale up the spatial interpolation from limited block level demographic attributes to more comprehensive block group level attributes. The test of our interpolation method only allocates the number of housing units and persons among parts of split census blocks. However, only rarely would it be important to study the distribution of people within a single block. Returning to our original application, that of allocating housing and population attributes from census source geography to non-census target geography, this method offers some refinement to that process. Moreover, such refinement has

implications beyond the allocation of the limited array of block level attributes. Rather than interpolating directly from the larger aggregated block groups, this method eliminates the need to interpolate blocks that fall fully within a target zone and those that fall fully outside the target zone. Then employing the node-based interpolation for split blocks allows us to determine with finer precision the extent of housing units and population located within the boundaries of a target zone. The imputation of population characteristics from the block group level requires the assumption that population *characteristics* are homogeneously distributed across blocks within the block group, but when our method brings to bear more accurate estimates of the distribution of the population associated with these characteristics.

We have automated the node count interpolation method discussed in this paper in a robust "Extension" for ArcView® 3.1 GIS software distributed by Environmental Systems Research Institute, Inc.. We are currently beta testing the complete and well documented extension prior to its release. The application produces a table of the proportion of housing units or population, or the proportion of another user-specified block-level population attribute for each block group located completely or partially within the boundaries of some target area (see Figure 4). These proportions can be used to weight census block group attributes (e.g., those in STF-3A) to generate detailed demographic profiles for non-census target geographies. We have most frequently applied the method to watersheds and sub-watersheds adding detailed population data available at the block group level to natural science data, facilitating interdisciplinary research. The extension would not necessarily be limited to census source geography but could calculate the aggregate length of road segments and the number of nodes within source polygons not defined by features available in TIGER given that an ArcView® compatible digital representation of the polygon boundaries was available. However, this non-census source geography interpolation would only be practicable for demographic characteristics associated with housing units and population.

[Figure 4 approx. here]

Acknowledgements. An earlier version of this paper was presented at the Annual Meeting of the Population Association of America, New York City, March 27, 1999. The research was supported in part by the following agencies: The Wisconsin Agricultural Experiment Station, through USDA Hatch Project No. 3865; the USDA through NRI Research Grant No. 98-35401-6158; the NSF through the North Temperate Lakes Long-Term Ecological Research (LTER/NTL) grant to the Center for Limnology, University of Wisconsin-Madison, Grant No. DEB-9632853; and by the National Institute of Child Health and Human Development through the Center for Demography and Ecology, University of Wisconsin-Madison, Grant No. P30-HD05876. Comments or questions may be addressed to David Long, Department of Rural Sociology, University of Wisconsin-Madison, 1450 Linden Drive, Madison, WI 53706. E-mail: dlong@ssc.wisc.edu.

References

Antle J M, Just R E, 1992, "Conceptual and empirical foundations for agricultural-environmental policy analysis" *Journal of Environmental Quality* **21** 307-316

Bracken I, 1991, "A Surface Model Approach to Small Area Population Estimation" *The Town Planning Review* **62** 225-237

Bracken I, Martin D, 1989, "The Generation of Spatial Population Distributions from Census Centroid Data" *Environment and Planning A* **21** 537-543

Bracken I, Martin D, 1995, "Linkage of the 1981 and 1991 UK Censuses Using Surface Modelling Concepts" *Environment and Planning A* **27** 379-390

Dempster A P, Laird N M, Rubin D, 1977, "Maximum Likelihood from Incomplete Data via the EM Algorithm" *Journal of the Royal Statistical Society B* **39** 1-38

Flowerdew R, and Green M, 1989, "Statistical Methods for Inference between Incompatible Zonal Systems," in *Accuracy of Spatial Databases* Ed. M Goodchild, S Gopal (Taylor & Francis, London) pp 239-247

Flowerdew R, Green M, 1992, "Developments in Areal Interpolation Methods and

GIS" *The Annals of Regional Science* **26** 67-78

Flowerdew R, Green M, 1994, "Areal Interpolation and Types of Data" in *Spatial Analysis and GIS* Ed. S Fotheringham, P Rogers (Taylor & Francis, London) pp 121-145

Flowerdew R, Green M, Kehris E, 1991, "Using Areal Interpolation Methods in Geographic Information Systems" *Journal of the Regional Science Association International* **70(3)** 303-315

Flowerdew R, Openshaw S, 1987, "A Review of the Problems of Transferring Data from One Set of Areal Units to Another Incompatible Set" *Research Report 4* (Northern Regional Research Laboratory, Lancaster and Newcastle)

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

Lam N S, 1983, "Spatial Interpolation Methods: A Review" *The American Cartographer* **10(2)** 129-149

Markoff, J, Shapiro G, 1973, "The linkage of data describing overlapping geographical units" *Historical Methods Newsletter* **7** 34-46

Martin D, 1996, "An Assessment of Surface and Zonal Models of Population" *International Journal for Geographical Information Systems* **10(8)** 973-989

Martin D, Braken I, 1991, "Techniques for Modelling Population-Related Raster Databases" *Environment and Planning A* **23** 1069-1075

Tobler W R, 1979, "Smooth Pycnophylactic Interpolation for Geographical Regions" *Journal of the American Statistical Association* **74** 519-530

Tordella S J, 1987, "How to Relate to Centroids" *American Demographics* **9(May)** 46-50

U.S. Bureau of the Census, 1991a, *Census Geography – Concepts and Products, Census Factfinder No. 8 (Revised)* (U.S. Bureau of the Census, Washington, DC)

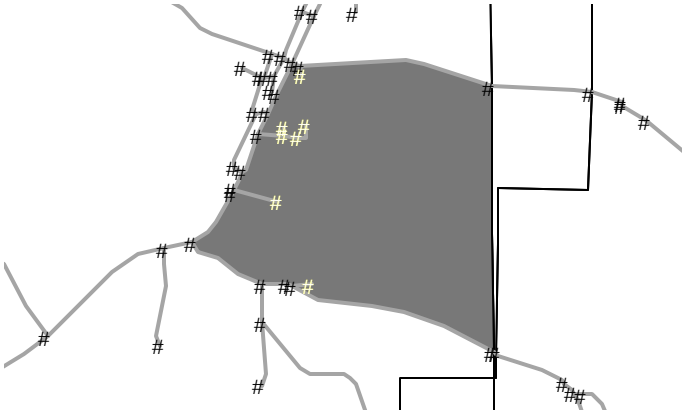
U.S. Bureau of the Census, 1991b, *Census of Population and Housing, 1990: Summary Tape File 1 [machine-readable data files]* (U.S. Bureau of the Census, Washington, DC)

U.S. Bureau of the Census, 1991c, *Census of Population and Housing, 1990: Summary Tape File 3 [machine-readable data files]* (U.S. Bureau of the Census, Washington, DC)

U.S. Bureau of the Census, 1996, *TIGER/Line Files, 1995: Technical Documentation* (U.S. Bureau of the Census, Washington, DC)

Figures and Tables

Figure 1
Block w/ Road Segments and Nodes Shown



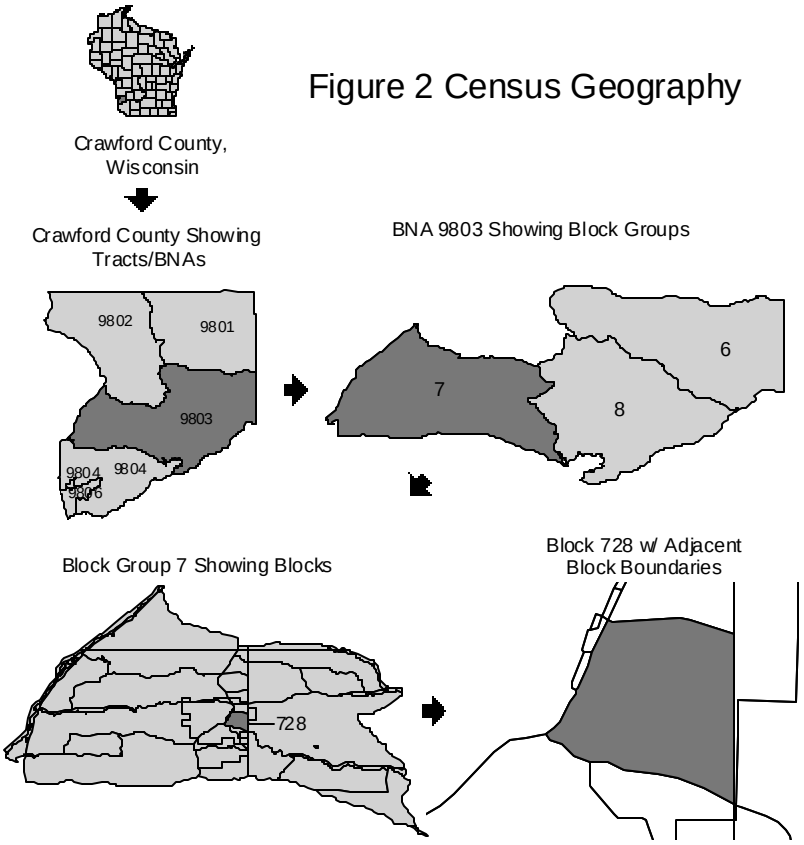


Figure 2 Census Geography

Figure 3
Crawford County Block Groups and
the Lower Kickapoo Subwatershed

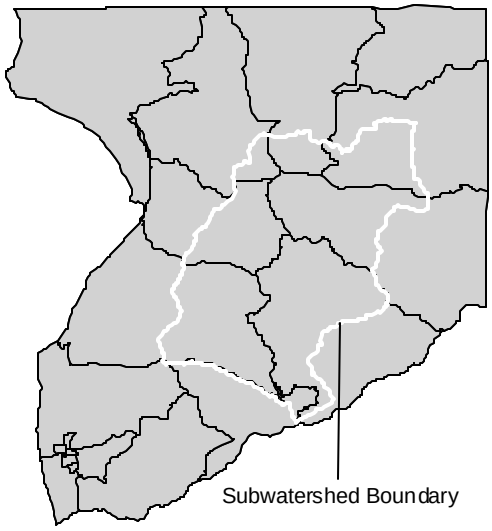


Table 1. Test Results Comparing Four Methods of Areal Interpolation

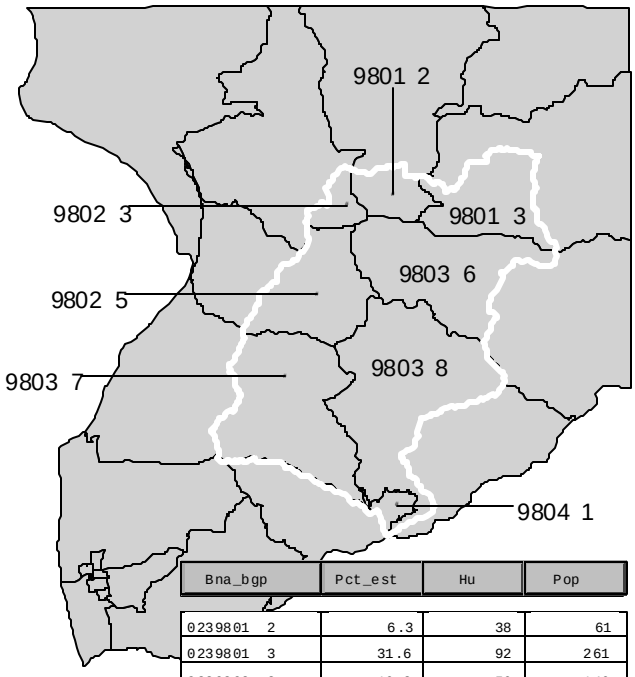
Method	Mean Absolute Error[1] (Error Ratio of Method to Areal Weighting)	Mean Absolute Percent Error[2] (Error Ratio of Method to Areal Weighting)[3]
Centroid Assignment	14.5 (1.53)	25.2% (1.41)
Areal Weighting	9.5 (1.00)	17.9% (1.00)
Road Segment Length	8.6 (0.90)	18.3% (1.02)
Internal Node Counts	7.1 (0.75)	16.6% (0.93)

[1]The equation is expressed as: MAE = (3| estimated Population minus actual Population |)/number of blocks.

[2] The equation is expressed as: MAPE = [(3| estimated proportion of the population minus actual proportion of population |)/number of blocks] x 100.

[3]The equation is expressed as: MAE = [(3| estimated Population minus actual Population |) / (3| estimated Population using areal weighting method minus actual Population |)] x 100.

Figure 4
Block Group Proportions for
Lower Kickapoo Subwatershed



Bna_bgp	Pct_est	Hu	Pop
0239801 2	6.3	38	61
0239801 3	31.6	92	261
0239802 3	16.2	56	146
0239802 5	47.0	114	331
0239803 6	38.8	140	329
0239803 7	51.0	184	552
0239803 8	58.8	152	397
0239804 1	60.5	204	493

Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400
comments to: dlong@ssc.wisc.edu
requests to: cdepubs@ssc.wisc.edu