Pergamon

Comput., Environ. and Urban Systems, Vol. 19, No. 4, pp. 287-306, 1995
Copyright © 1996 Elsevier Science Ltd.
Printed in the USA. All rights reserved
0198-9715/95 $9.50 + 0.00

0198-9715(95)00028-3

THE OVERLAID NETWORK ALGORITHMS FOR AREAL INTERPOLATION PROBLEM

Yichun Xie[1]
Department of Geography and Geology, Eastern Michigan University, Ypsilanti, Michigan 48197, U.S.A.

ABSTRACT. *In economic, social, and urban studies, areal units under analysis frequently differ from areal units over which data are compiled. Most area-based analyses, hence, face an unavoidable problem of transferring data across different zonal systems. This paper introduces a novel approach, the overlaid network algorithm, to develop a series of improved methods for tackling the population interpolation problem based on current GIS techniques and available digital information. The term network means that the partitioning of population for source zones is carried out over street segments. People are sheltered by houses which are located along the sides of streets or connected by roads. Thus, the street network provides an important information about the spatial distribution of population. The network length method discerns an even population distribution along one-dimensional line. The network hierarchical weighting method observes variations of residential density among different classes of streets. The network house bearing method breaks the assumptions of even population distribution over one and two dimensions and provides an automatic way of enumerating population over space. The application of these techniques to the interpolation of population in Erie County, New York improves the performance of areal interpolation significantly when compared with traditional methods.*

INTRODUCTION

Data of many kinds in the social, economic, and environmental sciences are collected and analyzed for areal units such as census tracts, block groups, blocks, service areas, school districts, election wards, watersheds, and soil regions. However, the analysis of these data is often made more difficult by the fact the areal units used differ among various data sets. The incompatibility of areal units arises from the fact that: (1) data come from different sources, for instance, boundaries of census tracts enumerated by U.S. Bureau of Census are usually different from those of administrative districts designated by local governments; (2) areas being studied change over time, for example, a city's boundary may expand because of annexation of suburban townships; and (3) specialized areas defined for particular applications, such as zoning by planning boards and redistricting by election commissions, are not consistent with existing district boundaries. As a result, many analyses of areal (geographical) data face unavoidable problems in comparing or transforming data collected for different zonal systems. In other words, the handling of geographical data often involves the transformation of data from one system of areal units (source zones) to another (target reporting zones), an activity referred to as *areal interpolation* in GIS terminology (Goodchild & Lam, 1980).

[1] Tel: 313-487-0218

This paper introduces a new approach that uses current GIS techniques (overlaying) and widely available digital databases (the U.S. Census Bureau's Topologically Integrated Geographical Encoding and Referencing—TIGER^{TM}/Line Files) to develop improved methods for dealing with the areal interpolation problem.

The paper is divided into five sections: (1) a brief introduction of some concepts pertaining to thematic data handling in map overlay; (2) a review of major existing algorithms for areal interpolation; (3) a step-by-step discussion of the new overlaid network algorithms, the network length (NL) algorithm, the network hierarchical weighting (NHW) algorithm, and the network housing-bearing (NHB) algorithm; (4) an analysis of errors derived from the overlaid network algorithms and those from commonly used methods for a real-world example, Erie County, New York; and (5) conclusions.

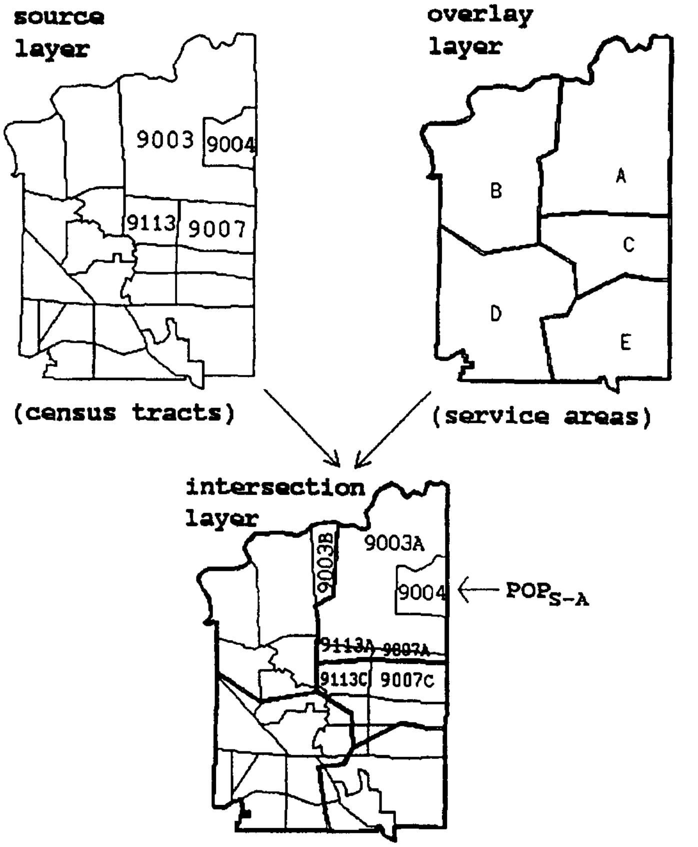
AREAL INTERPOLATION—AN IMPERATIVE PROCESSING OF THEMATIC DATA IN GIS OVERLAY OPERATION

What distinguishes a GIS from other types of information systems are its spatial analysis functions (Aronoff, 1989). These functions include the analysis of spatial data, the analysis of attribute (thematic) data, and, more importantly, the integrated analysis of spatial and attribute data. It is the analysis of complex, multiple types of spatial and non-spatial data in an integrated manner that cannot be done effectively with manual methods or with compute-aided design and drafting systems. The key operation in the integrated analysis of spatial and attribute data is *map overlay* (Goodchild & Gopal, 1989). Map overlay is one of the two GIS capacities (the other is changing map scales) that has excited enthusiasm among potential users (Abler, 1987).

The map overlay process involves the superimposition of two or more input maps, or data layers, with the aim of producing a composite map showing the intersection of the mapping units on the individual data layers (Veregin, 1989). An overlay operation usually requires two input data layers (a *source layer* and an *overlay layer*) and generates a composite layer (or *intersection layer*) (Fig. 1).

Applying the map overlay operation to input data layers creates both cartographic and thematic changes which are recorded on the composite layer. The cartographic reformation in the composite layer is created by intersecting polygons in the input layers (Fig. 1). For example, the created polygons often break the boundary unity of both source zones and overlay zones. Hence, a newly created intersection polygon usually does not coincide with the boundary of either a source zone or a overlay zone (Openshaw et al., 1986). Furthermore, this topological divergence makes it difficult for analysts to attach appropriate values of attributes to the newly created polygons. The estimation of attribute values for the intersection layer is frequently based on the values of the source layer. Particularly in the context of social, economic, and environmental studies, with resort to the overlay operation, analysts obtain attribute information for one system of areal units (target zones) from the known information of another system of areal units (source zones). This kind of data processing is a typical case of areal interpolation.

From the perspective of areal interpolation, the source zone and the target zone are equal to the source layer and the overlay layer in the terminology of GIS. However, the intersect layer, the final product of overlaying, is treated as an intermediate or transitional bridge for the purpose of transferring information from the source layer to the target (overlay) layer. For example, how many residents live in Service Area A can be calculated based on the population information in the census tracts which fall within the service region (Fig.



$$POP_{S-A} = 90\%(POP_{9003}) + 100\%(POP_{9004}) + 35\%(POP_{9113}) + 20\%(POP_{9007})$$

* All illustrations are based on Amherst, New York.

FIGURE 1. Map overlay operation.*

1). The computation involves two technical steps: (1) partitioning attribute values of source zones onto intersection zones; and (2) moving data from intersect zones to overlay zones. An illustration of this approach based on Fig. 1 will be given in the next section.

Most current methods of areal interpolation adopt the concept of overlay for moving data between source and target zones since tracking and indexing of overlaid features have been effectively solved by GIS techniques. However, how to partition values of attributes of the source zones remains controvertible and prevails the areal interpolation problem.

MAJOR EXISTING ALGORITHMS DEALING WITH AREAL INTERPOLATION

Many algorithms have been designed for areal interpolation. The *NCGIA CORE CURRICULUM* classifies these methods by six dichotomies into six pairs (Goodchild & Kemp, 1990): point versus areal; global versus local; exact versus approximate; stochastic versus deterministic; volume preserving versus non-volume preserving; and gradual versus abrupt.

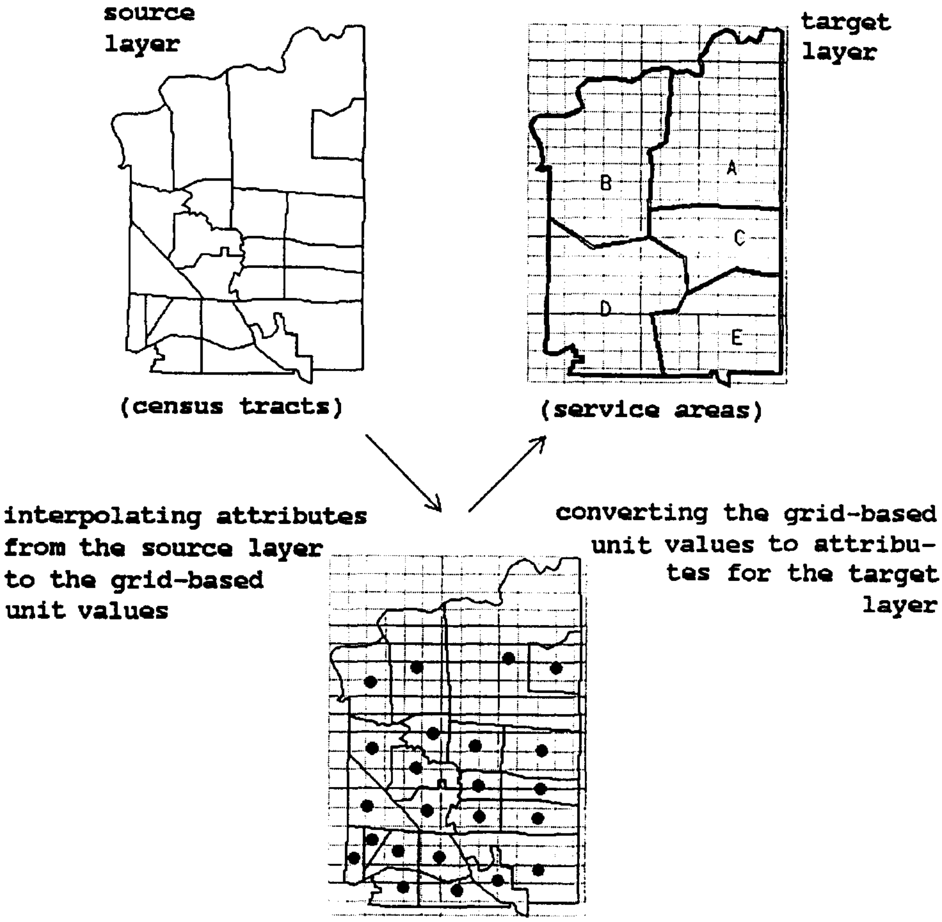

source layer
target layer
B
A
C
D
E
(census tracts)
(service areas)
interpolating attributes from the source layer to the grid-based unit values
converting the grid-based unit values to attributes for the target layer

FIGURE 2. The Global Smoothing Algorithm

However, these algorithms differ fundamentally in their methods of partitioning the values for the source layers. From the perspective of source value partitioning, three general types of spatial interpolation techniques, spatial-smoothing (point/raster-based), areal weighting (area/vector-based), and modeling (statistics-based) can be identified. These approaches are explained below.

Raster-based (spatial-smoothing) areal interpolators

These interpolators are based on regular grids of fine resolutions and divided into two categories from the view point of interpolation characteristics:

(1) *The global smoothing algorithm*: This group of methods usually produce **approximate** interpolation results. The main procedures include (Fig. 2),

- transfer the count variable into a ratio variable (divided by the area of each zone, such as zonal population density),
- identify a centroid for each zone and assign the zonal ratio value to this centroid,
- interpolate a grid surface by using the ratio values of this set of points and by one of the techniques, **Kriging, Polynomial Trend Surface Analysis, Fourier Series Analysis**, and **Moving Average Analysis** (Sampson, 1978; Lam, 1983; Burrough, 1986; Davis, 1986; McBratney & Webster, 1986; Dutton-Marion, 1988; Goodchild & Kemp, 1990; Oliver,

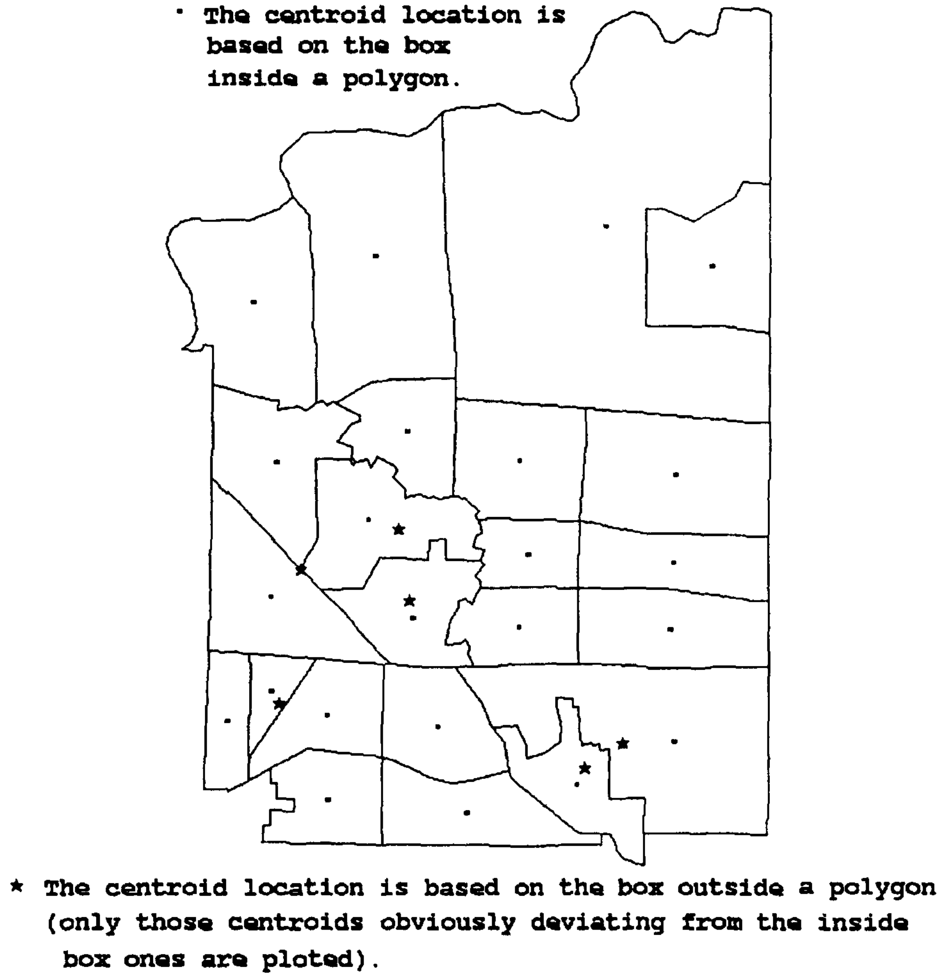

• The centroid location is based on the box inside a polygon.
★ The centroid location is based on the box outside a polygon (only those centroids obviously deviating from the inside box ones are ploted).

FIGURE 3. Unstable Locations of Polygon Centroids

1990;),

- convert the ratio value of each grid cell to a count value by multiplying the cell's area,
- overlay the interpolated grid on the target map and calculate the value for each reporting zone by adding up grid values within its range.

This approach has two obvious limitations. First, the total value for each reporting zone may not be conserved because the spatial variation in the phenomenon represented by the ratio variable is assumed to be homogeneous statistically throughout the entire map area. Second, the centroid location has a substantial impact on the interpolated values but it depends heavily on the geometry of each zone, and may lay outside of the source zone (Fig. 3). That is, the interpolation results derived by this group of methods are not consistent or stable since they are influenced by many technical or computational details.

(2) *The local smoothing algorithm*: This is an improved version of the raster-based approach which attempts to address the limitations confronted by the previous global algorithm. This group includes the *Proximal Analysis*, *B-splines*, and *Pycnophylactic Analysis* approaches, which usually give exact interpolations (Goodchild & Kemp, 1990). The Pycnophylactic algorithm is a typical local smoothing method originally proposed by Tobler

(1979) and designed for count variables such as size, volume of research objects. This algorithm is comprised of the following procedures:

- overlay a dense raster (grid) on a source map,
- divide each zone's value equally among the overlaid raster cells,
- smooth the values by replacing each cell's value with the average of its neighbors,
- sum the values of the cells in each source zone and adjust the values of all cells within each zone respectively so that the zone's total is the same as the original value,
- iterate steps 3 and 4 until the difference between the sum and the original value falls within certain threshold,
- overlay the interpolated grid on the target map and calculate the value for each reporting zone by adding up grid values within its range.

Tobler's algorithm emphasizes creating a smooth surface but does not confine itself to homogeneity assumption, which compromises between two extremes, homogeneity and heterogeneity. Furthermore, this method tries hard to conserve the original value of each reporting zone, a property which is preferred by most users, though it commits greater demands on computation time.

The raster-based approaches are technically complicated in terms of operational design. They generally involve interactive or iterative investigations of the spatial behavior of the phenomenon being studied. The implementation of these approaches often requires that analysts have a precise knowledge about the data and a thorough understanding of spatial statistics. Moreover, few software packages include functions performing these spatial interpolations (such as ARC/INFO's GRID Module, and Idrisi???). Therefore, it is strongly recommended that analysts should be particularly careful when applying raster-based methods of spatial interpolation unless they have both conceptual awareness and technical skills (ESRI, 1991a,b).

Areal weighting interpolators

The areal weighting approach is most used in practice since the algorithm for this approach is defined clearly, included in most GIS software packages, and implemented easily (MacDougall, 1976; Goodchild & Lam, 1980). This method involves:

- overlaying the target zones on the source zones,
- determine the proportion of each source zone that falls into each target zone,
- apportion the attribute value for each source zone to target zones according to the areal proportions.

The methods used to apportion the attribute values into the new regions created in the intersection layer rely on the types of variables. Goodchild and Lam (1980) classify between intensive and extensive variables. When a zone is divided into a set of subzones, a variable is described as intensive if the value for the zone is a weighted average of the values for the subzones; it is extensive if its value for a zone is the sum of its values for the subzones. For example, the question of how many residents are served in Service Area I can be easily solved by the areal weighting method. The residents in Service Area I is simply equal to the sum of the population over those intersection zones which fall within Service Area A (Fig. 1),

$$\mathrm{POP}_{S-A} = 90\%(\mathrm{POP}_{9003}) + 100\%(\mathrm{POP}_{9004}) + 35\%(\mathrm{POP}_{9113}) + 20\%(\mathrm{POP}_{9007}).$$

The areal weighting algorithm is apparently based on the assumption that the values of the variable of interest are evenly distributed across each of the source zones. The areal weighting method is appropriate when there is no additional information available in the source zone although the assumption of even distribution is rarely hold in the real world.

Statistical interpolators

The statistical method is a new approach to the issue of areal interpolation. Some cartographers and GIS researchers attempt to apply statistical or mathematical models for areal interpolation (Goodchild & Hosage, 1983; Amrhein & Flowerdew, 1989). This method's proponents claim two important advantages over earlier methods of areal interpolation (Flowerdew, 1988; Flowerdew & Green, 1989; Green, 1989). First, the statistical methods take into account the values of other (predictor) variables to which the variable of interest may be related. Second, the new methods really are based on statistical assumptions, which provide maximum likelihood estimates of values for the target zones. Flowerdew and Green (1992) comment,

> "Essentially the method works through establishing a regression relationship between the variable of interest and one or more ancillary variables. Once this relationship has been established, it can be used, along with area, to estimate values for the variable of interest for the target zones."

A heavily used statistical interpolation method is the *Poisson Model* developed by Lovett and Flowerdew (1989), and Flowerdew and Green (1989). Flowerdew and Green (1990) made improvement to their Poisson Model through integrating with the Expectation and Maximum-likelihood (EM) algorithm developed by Dempster et al. (1977). More recently, the EM algorithm has been extended to deal with binomial distributions (Flowerdew et al., 1991) and continuous variables of normal distribution (Flowerdew & Green, 1992).

The statistical approach of areal interpolation, which incorporates additional information for interpolating values of source zones to target zones, suggests a fruitful direction for solving the areal interpolation problem. This approach identifies mathematical distribution patterns (such as poisson, binomial, and normal) for selected ancillary data and utilizes well acknowledged formulae to estimate values for the variables of interest. It employs stochastic concepts to tackle the uncertainty of the areal interpolation problem and attaches clearly defined statistical meanings to the solutions of the areal interpolation. It seems to be the best method when no information is available to derive deterministic interpolations.

There are some questions, however, about the statistical approach. First, the statistical method assumes that the variable of interest has a distribution which can be precisely described in mathematical language (or formula). This prerequisite increases the complexity of the areal interpolation task, and limits its application because some assumptions about data statistical properties can rarely be met in practice. Second, there is often an abundance of information available in practice. Especially in the United States, the US Bureau of Census and the United States Geological Survey (USGS) have released abundant digital and attribute (census) data sets, which can be used to develop uncomplicated and cogent methods for solving the areal interpolation problem. The statistical approach makes no attempt to use this important improvement of information accessibility. Third, the method adopts an approximate approach which defines a single function which is mapped across the whole research region. That is, it derives a general formula from individual (or sample) observations based on assumed statistical rules, and then applies that formula to the individual units being studied. In the process, the original information is usually lost to some extent, reducing the accuracy of the estimation for the target zones compared to the source zones. This defeats the purpose of areal interpolation in which researchers attempt to conserve the original information at the smallest area units (Goodchild & Kemp, 1990).

THE OVERLAID NETWORK ALGORITHMS

The areal interpolation problem still lacks a satisfactory solution even though the literature dealing with the issue is growing. The common limitation of the existing methods is that they are confined to the limited information on a single variable reported for source zones or on an approximate estimate based on assumed mathematical relationships between the variable of interest and ancillary variables. As Goodchild and Gopal (1989, p.219) point out,

> "It (areal interpolation) is also an information-limited problem—the more one knows about the area, the better the estimates."

This notion of incorporating more **relevant** and **directly related** information and utilizing available **GIS techniques** to the areal interpolation problem provides the rationale for the overlaid network algorithm, which has only become feasible recently.

Rapid and steady improvement of computer techniques and applications creates congenial conditions for developing improved approaches for solving the areal interpolation problem. First, the volumes of readily standardized spatial and attribute data in digital format are exploding. For instance, US Bureau of Census's TIGER/Line files provide abundant information with many potential uses for many applications in social and natural sciences (US Bureau of Census, 1991a; Klosterman & Xie, 1992). The Bureau of Census also provides several series of attribute data in computer readable form, such as Public Law 94-171 Counts File (redistricting data) and 1990 Census of Population and Housing Summary Tape Files (US Bureau of Census, 1990, 1991b). These data can easily be combined with the TIGER/LINE data.

Second, a variety of commercial GIS packages are capable of processing large volume of spatial and attribute data in an integrated manner, such as ARC/INFO, MapInfo, Atlas-GIS, ArcCad, TransCad, Erdas, and Idrisi. More profoundly, current software development enables flexible and versatile manipulations of spatial data. For instance, a technique for partition and aggregation of spatial data rather than simple overlay of maps has been developed in NCGIA at Buffalo (Batty & Xie, 1994a, 1994b). The availability of these data and the power of modern GIS software allow the adoption of new approaches for areal interpolation.

The TIGER/Line files contain all map features, including information on street segments which are particularly useful for areal population interpolation. People are sheltered by houses. Houses are usually located along the sides of streets or along roads. As a result, the distribution of population in an areal unit is closely related to the street network. This suggests that the street network provides an important supplementary information about a population's distribution over an area. This supplement of direct related information forms the basis for constructing the *overlaid network algorithms.*

The term *network* means that partitioning of the attribute values (population in this example) for the source layer is based on street segments. In other words, the population of a source zone is allocated (or distributed) to all street segments within this zone. Three methods of allocating the population to the street segments (by street segment length, by street segment classes, and by house bearings of street segments) will be discussed in this paper (Fig. 4).

In the terminology of GIS, the network is characterized by a set of one-dimensional objects, i.e. lines. The source or target zones are comprised of two-dimensional objects, i.e. polygons. The transferring of data between source and target zones is achieved with assistance of a series of GIS overlay operations: (1) overlaying source and target layers to get an intersection layer; and (2) overlaying the intersection layer and the network layer to realize

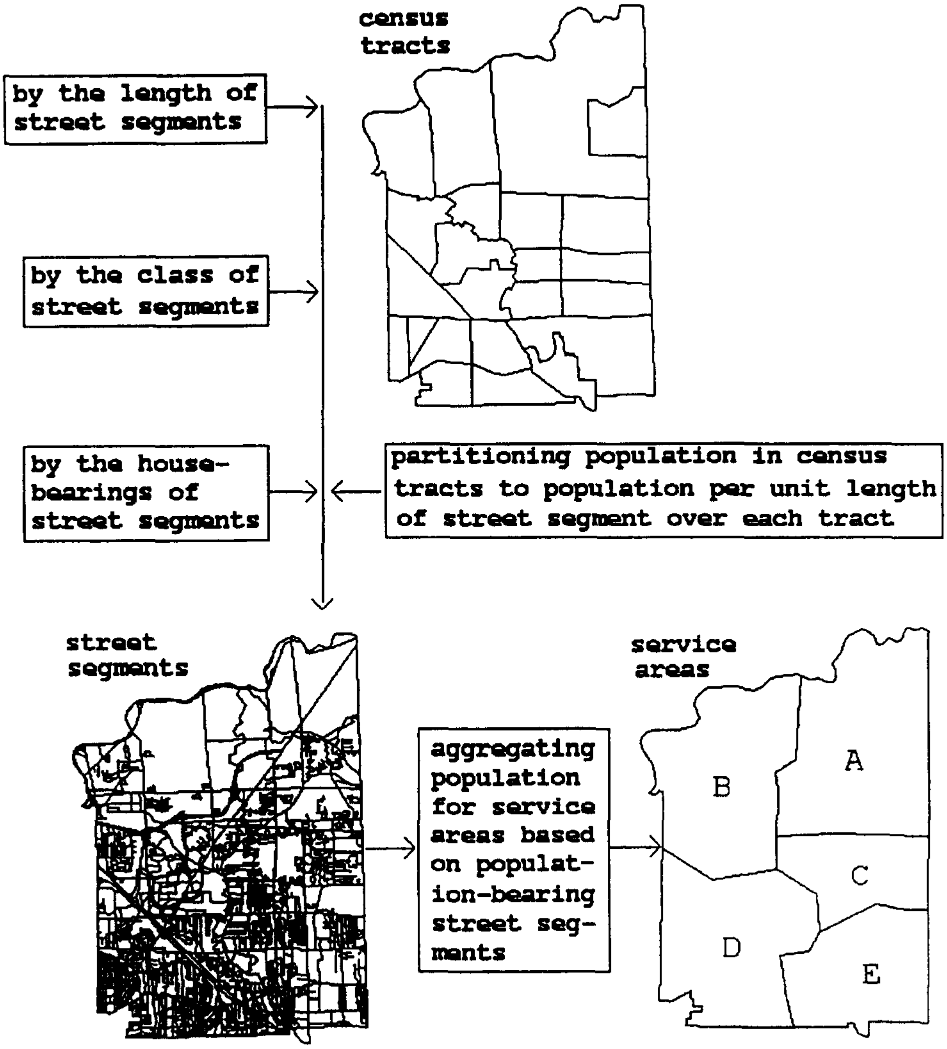
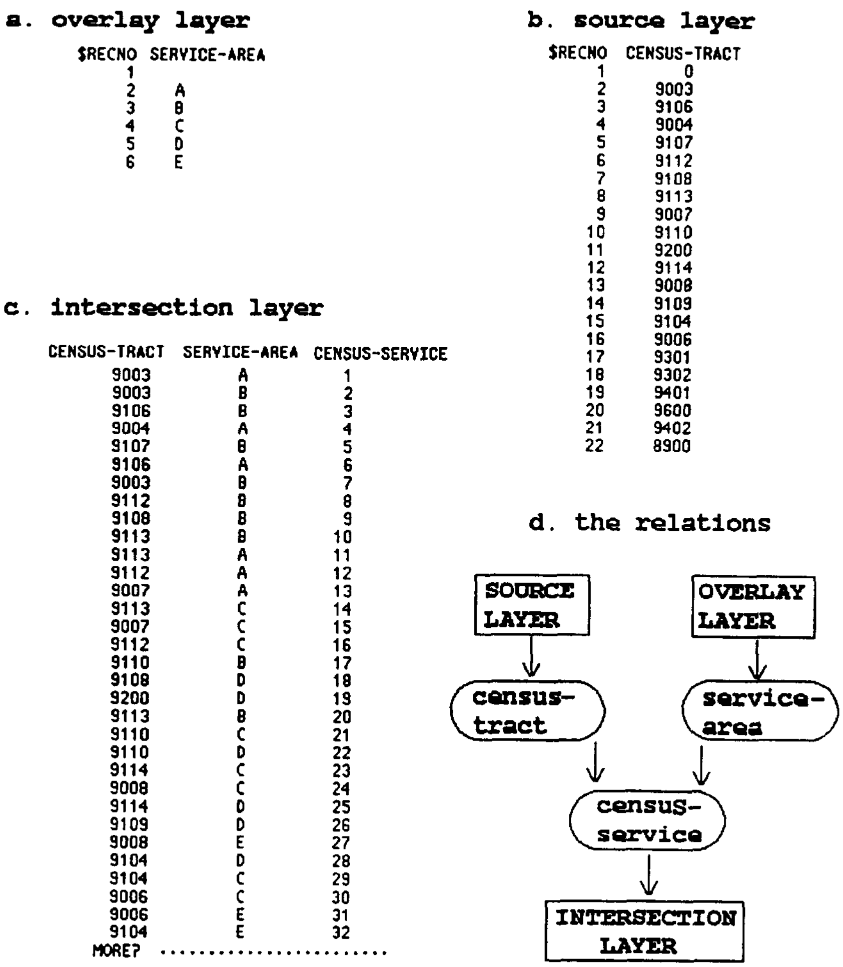

census tracts
by the length of street segments
by the class of street segments
by the house-bearings of street segments
partitioning population in census tracts to population per unit length of street segment over each tract
street segments
aggregating population for service areas based on populat-ion-bearing street seg-ments
service areas
A
B
C
D
E

FIGURE 4. The Overlaid Network Algorithm

the data transferring.

Applying overlay of two incompatible zonal layers will create a layer of intersected polygons (Fig. 1). The relationships between the new polygons and the source and target polygons can be traced by listing the New-Polygon-ID, the Source-Polygon-ID, and the Target-Polygon-ID (Fig. 5). Then, overlaying the intersection layer to the network (line) layer creates intersected lines for each intersection polygon (Fig. 6). The newly created line segments bear relationships with the intersecting polygons through two data items, *Left-Polygon*, and *Right-Polygon* (Fig. 7). Hence, the topological relations between polygon and line layers, and between source and target layers can be easily established. In turn, the attribute exchanges between these layers can be identified through the topological relations. For instance, the population information in a source zone can be ***partitioned*** to the street segments within the source zone by one of the ***standardized*** methods: *street-length, street-hirarchy-weighting, street-house-bearing*. Then the standardized segment-based quantity is treated as a **universal constant** and can thus be used easily to transfer population information from the source



a. overlay layer

```
$RECNO SERVICE-AREA
   1
   2     A
   3     B
   4     C
   5     D
   6     E
```

b. source layer

```
$RECNO  CENSUS-TRACT
   1        0
   2      9003
   3      9106
   4      9004
   5      9107
   6      9112
   7      9108
   8      9113
   9      9007
  10      9110
  11      9200
  12      9114
  13      9008
  14      9109
  15      9104
  16      9006
  17      9301
  18      9302
  19      9401
  20      9600
  21      9402
  22      8900
```

c. intersection layer

```
CENSUS-TRACT  SERVICE-AREA  CENSUS-SERVICE
    9003           A              1
    9003           B              2
    9106           B              3
    9004           A              4
    9107           B              5
    9106           A              6
    9003           B              7
    9112           B              8
    9108           B              9
    9113           B             10
    9113           A             11
    9112           A             12
    9007           A             13
    9113           C             14
    9007           C             15
    9112           C             16
    9110           B             17
    9108           D             18
    9200           D             19
    9113           B             20
    9110           C             21
    9110           D             22
    9114           C             23
    9008           C             24
    9114           D             25
    9109           D             26
    9008           E             27
    9104           D             28
    9104           C             29
    9006           C             30
    9006           E             31
    9104           E             32
  MORE? ........................
```

d. the relations

FIGURE 5. Topological Relations of Polygon Overlay

zones to the target zones through the intersected street segments associated with the intersection zones. Therefore, the transferring of information between source and target zones can be accomplished through overlay operations, that is, intersected zones and corresponding intersected line segments. How to partition, standardize, and transfer data by the three *overlaid network algorithms* is explained below.

Network length method (NL)

The length of a street segment is a rudimentary quantity of street network. Most GIS packages automatically record the lengths of line features. Thus, it is easy to obtain the total length of all of street segments in a GIS processing.

The NL is the simplest of the overlaid network algorithms and is easy to be implemented. The procedures include:

(1) Prepare three data layers or coverages: a source layer consisting of *s* zones (or polygons), a target layer made up of *t* zones, and a street network made up of *n* lines
(2) Use standard GIS function to overlay the source-zone layer with the target-zone layer to create a new polygon layer, the intersection-zone layer(*i*).
(3) Overlay the intersection-zone layer with the network layer to create a new line layer,

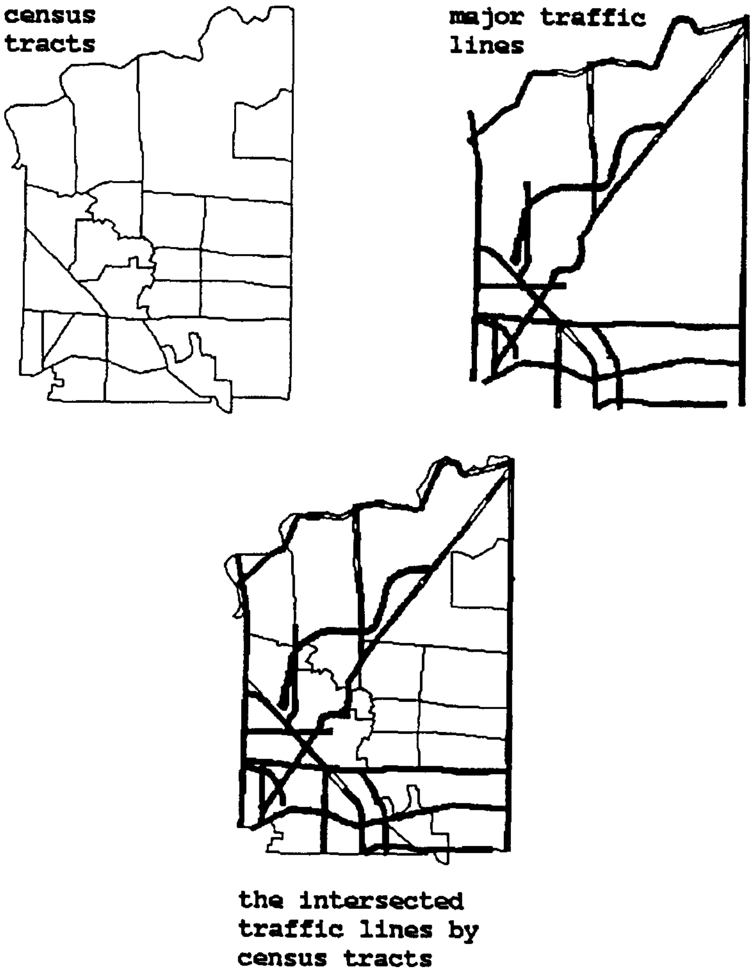


FIGURE 6. Polygon and Line Overlay Operation

the control-net (*c*).

(4) Overlay the source-zone layer and the network layer to determine the length of street segments within each source zone; then allocate the population to the street segments in each source zone by length to get *the population per unit length* (based on the polygon-id information from the source-zone layer and the right-poly and left-poly information from the network layer in the terminology of ARC/INFO),

$$\mathrm{LEN1}_s = \sum_s \sum_n \mathrm{LENGTH1},$$

$$\mathrm{POPLEN}_s = \mathrm{POP1}_s / \mathrm{LEN1}_s$$

where $\mathrm{LEN1}_S$ is the total length of street segments in a source zone, *s*; LENGTH1 is the length of a street segment in the network layer. (If a street segment lies in the common border between two adjacent zones, one half of the segment length is counted for each zone. This solution is also used in the network hierarchical weighting method.) POPLEN_S is the population per unit length (or the linear density) for a source zone, *s*; and POP1 is the population of the source zone.

a. polygon layer

TRACT	POLY#
0	1
9003	2
9106	3
9004	4
9107	5
9112	6
9108	7
9113	8
9007	9
9110	10
9200	11
9114	12
9008	13
9109	14
9104	15
9006	16
9301	17
9302	18
9401	19
9600	20
9402	21
8900	22
9501	23
9502	24

b. the intersected line layer

LINE#	L-POLY#	R-POLY#
1	2	1
2	2	1
3	2	2
4	2	2
5	2	2
6	3	2
7	2	2
8	2	2
9	2	2
10	3	2
11	2	2
12	3	3
13	3	3
14	2	1
15	2	1
16	2	1
17	2	2
18	3	3
19	2	2
20	5	5
21	2	2
22	5	5
23	3	2
24	5	5
25	5	5
26	2	2
27	2	2
28	2	2
29	2	2
30	5	5
....		

c. the relations

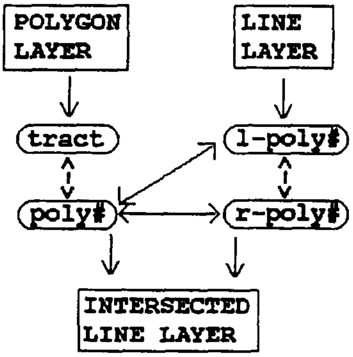


FIGURE 7. Topological Relations of Polygon and Line Overlay

(5) Within each intersection zone, determine the length of subsets of street segments based on the information from the layers of intersect-zone and control-net,

$$\mathrm{LEN2}_i = \sum_i \sum_c \mathrm{LENGTH2},$$

where $\mathrm{LEN2}_i$ is the total length of street segments in an intersection zone, *i*, and LENGTH2 is the length of a street segment in the control network layer.

(6) Multiply the network length in the intersection zones by the population per unit length in the corresponding source zones, and sum these values to get the interpolated population counts for the target zones,

$$\mathrm{POP2}_i = \mathrm{LEN2}_i * \mathrm{POPLEN}_s,$$

$$\mathrm{POP}_t = \sum_t \sum_i \mathrm{POP2}_i,$$

where $\mathrm{POP2}_i$ is the population for an intersection zone, $i \in s$; POP_t is the population in a target zone, $i \in s$.

Table 1. Road features (CFCC Classification A)

CFCC	ROAD FEATURES
A0X*	road, classification unknown or not elsewhere classified
A1X	primary road (interstate highway) category
A2X	secondary road (state road) category
A3X	connecting road category
A40	neighborhood roads, city streets and unimproved roads
A41	neighborhood roads, city streets and unimproved roads, undivided
A42	neighborhood roads, city streets and unimproved roads, undivided, in tunnel
A43	neighborhood roads, city streets and unimproved roads, undivided, underpassing
A44	neighborhood roads, city streets and unimproved roads, undivided, rail line in center
A45	neighborhood roads, city streets and unimproved roads, divided
A46	neighborhood roads, city streets and unimproved roads, divided, in tunnel
A47	neighborhood roads, city streets and unimproved roads, divided, underpassing
A48	neighborhood roads, city streets and unimproved roads, divided, rail line in center
A5X	jeep trail
A6X	special road category
A7X	other thoroughfare category

[a] * X is used here to indicate a group of classes.

Source: US Bureau of the Census, 1991, *TIGER/LINE^{TM} CENSUS FILES, 1990: TECHNICAL DOCUMENTATION.*

Network hierarchial weighting method

The network length method assumes an equal distribution of population along the streets in each zone. This is questionable although an assumed equal distribution along lines is preferable to that over areas. However, we know streets are of various classes. The census feature class code (CFCC) defines eight categories of road features, each consisting of several classes (Table 1). Common sense suggests that residential densities or population concentrations differ along different categories of roads. That is, very few people live along interstate highways. Few people dwell along state roads or connecting roads. The majority of people reside along neighborhood roads, even though there exist deviations among different classes of neighborhood roads. This suggests that the classifications of roads should be taken into account to yield more accurate population estimates.

A weight matrix W_{sc} can thus be introduced into the network methods of areal interpolation. The first three procedures of the network hierarchical weighting method are the same as those of the network length method. The remaining steps are:

(1) Construct a weight matrix, W_{sc}, for the source layer according to the following rules: (i) the row sum, $\sum_s W_{sc} = 1$; (ii) $W_{sc} = 0$ when the road class c is missing in the source zone s; and (iii) assign a higher positive value to W_{sc} if the road class c is assumed to be densely resided, which depends on the field survey in a specific area or on an averaged value over a large reference region.

(2) Overlay the source-zone layer with the network layer, then determine the lengths of streets of various classes $LEN1_{sc}$ within each source zone, calculate the weighted length $WEILEN1_s$ by multiplying $LEN1_{sc}$ by W_{sc} element by element and then summing them up, and allocate the population counts to the network by the weighted length based on the polygon-id information from the source-zone layer and the right-poly and left-poly information from the network layer,

$$LEN1_{sc} = \sum_{nc} LENGTH1,$$

$$\mathrm{WEILEN1_s} = \sum_{sc}(\mathrm{LEN1_{sc}}\,.*\,\mathrm{W_{sc}}),$$

$$\mathrm{POPLEN_s} = \mathrm{POP1_s/WEILEN1_s};$$

(3) Within each intersection zone, determine the lengths of subsets of street network of various classes based on the information from the layers of intersect-zone and control-net, and then calculate the weighted length by multiplying $\mathrm{LEN2_{ic}}$ with $\mathrm{W_{sc}}$ element by element and then summing them up,

$$\mathrm{LEN2_{ic}} = \sum_{cc}\mathrm{LENGTH2},$$

$$\mathrm{WEILEN2_i} = \sum_{ic}(\mathrm{LEN2_{ic}}\,.*\,\mathrm{W_{sc}})$$

(4) Multiply the weighted network length in the intersection zones with the population counts per unit weighted length in corresponding source zones, and add up to get interpolated population counts for target zones,

$$\mathrm{POP_i} = \mathrm{WEILEN2_i} * \mathrm{POPLEN_s}$$

$$\mathrm{POP_t} = \sum_{i}\mathrm{POP_i}.$$

Network housing-bearing method

The network hierarchical weighting method considers the differences in residence densities along various classes of roads and streets. The approach seems more reasonable in theory and seems to approximate the population distribution along streets more accurately than the NL approach. However, the hierarchical weighting method is based only on a reasonable interpolation of the real world, not on the observation of the real world. With the flood of available digital information and the assistance of modern GIS techniques, a method analogous to real inspection is now available.

The TIGER/Line files contain four items that can be used for address matching: the "from" address left (FRADDL), the "to" address left (TOADDL), the "from" address right (FRADDR), and the "to" address right (TOADDR). That is, these address ranges provide a better measure of how many houses locate along a street segment. Moreover, the STF-1A file includes detailed information about housing. The information from the two sources can be checked with each other so that a clear picture of the association between housing and population can be postulated. In other words, population can be allocated to houses and houses can be attached to street segments according to the address ranges in each side of the streets in each zone. Thus, this algorithm is named, *network housing-bearing method.* The implementation of the housing-bearing method is quite straightforward. The first three steps are the same as the other two algorithms. Additional steps involve:

(1) Determine the average number of persons per housing in each source zone by dividing the population counts by the number of house units,

$$\mathrm{PERHOU_s} = \mathrm{POP1_s/HOU1_s}$$

(2) Calculate the number of housing units per unit length of street segment (the house bearing) according to the right-address-range and left-address-range information from the network layer,

$$\mathrm{NETHBEAR} - \mathrm{L1}_n = \mathrm{abs}(\mathrm{FRADDL}_n - \mathrm{TOADDL}_n)/(2 * \mathrm{LENGTH1}_n),$$

$$\mathrm{NETHBEAR} - \mathrm{R1}_n = \mathrm{abs}(\mathrm{FRADDR}_n - \mathrm{TOADDR}_n)/(2 * \mathrm{LENGTH1}_n),$$

where NETHBEAR-L1$_n$ is the house bearing of the left side of a street segment in the network layer; NETHBEAR-R1$_n$ is the house bearing of the right side of a street segment in the network layer; LENGTH1$_n$ is the length of a street segment in the network layer.

(3) Calculate the number of housing units for street segments in the layer of control-net by multiplying the segment length with the corresponding housing bearing calculated from Step (2), and by linking the layers of network and control-net through the items, the original left or right polygons and the newly-overlaid left or right polygons,

$$\mathrm{HOUSENUM} - \mathrm{L2}_c = \mathrm{LENGTH2}_c * \mathrm{NETHBEAR} - \mathrm{L1}_c,$$

$$\mathrm{HOUSENUM} - \mathrm{R2}_c = \mathrm{LENGTH2}_c * \mathrm{NETHBEAR} - \mathrm{R1}_c$$

(4) Multiply the number of housing units in the control network by the number of persons per house in the corresponding source zones, and sum up to get the interpolated population counts for the target zones,

$$\mathrm{POP} - \mathrm{L}_c = \mathrm{NETHBEAR} - \mathrm{L2}_c * \mathrm{PERHOU}_s,$$

$$\mathrm{POP} - \mathrm{R}_c = \mathrm{NETHBEAR} - \mathrm{R2}_c * \mathrm{PERHOU}_s,$$

$$\mathrm{POP}_t = \sum_c (\mathrm{POP} - \mathrm{L}_c + \mathrm{POP} - \mathrm{R}_c).$$

EVALUATION OF ALTERNATE METHODS

Conceptually the assumptions of the network-based techniques are more reasonable. The traditional areal-weighting method assumes an even population distribution over two-dimensional space. The network length method assumes an even distribution of population along one-dimensional line. This is much more appealing in distinguishing between areas of the same size that may have few or many traversing roads and streets. The network hierarchial weighting method distinguishes between streets of various classes and assigns population to network by different weights. The network housing-bearing method breaks the assumptions of an even distribution over one and two dimensions, and comes closer to real observations of population. These considerations suggest that the applications of these new techniques should improve the performance of areal interpolation when compared with the traditional ones. A test of this assumption is provided in this section.

The test area includes Erie County, New York, which includes the City of Buffalo and its nearby suburbs. The county consists of 236 census tracts, 973 block groups, and 40,574 line segments (Fig. 8). For comparison purposes, the census tracts are used as the source zones, and the block groups are used as the target zones. Incompatible zonal systems are intentionally avoided in the test because the observed population values are available for the block

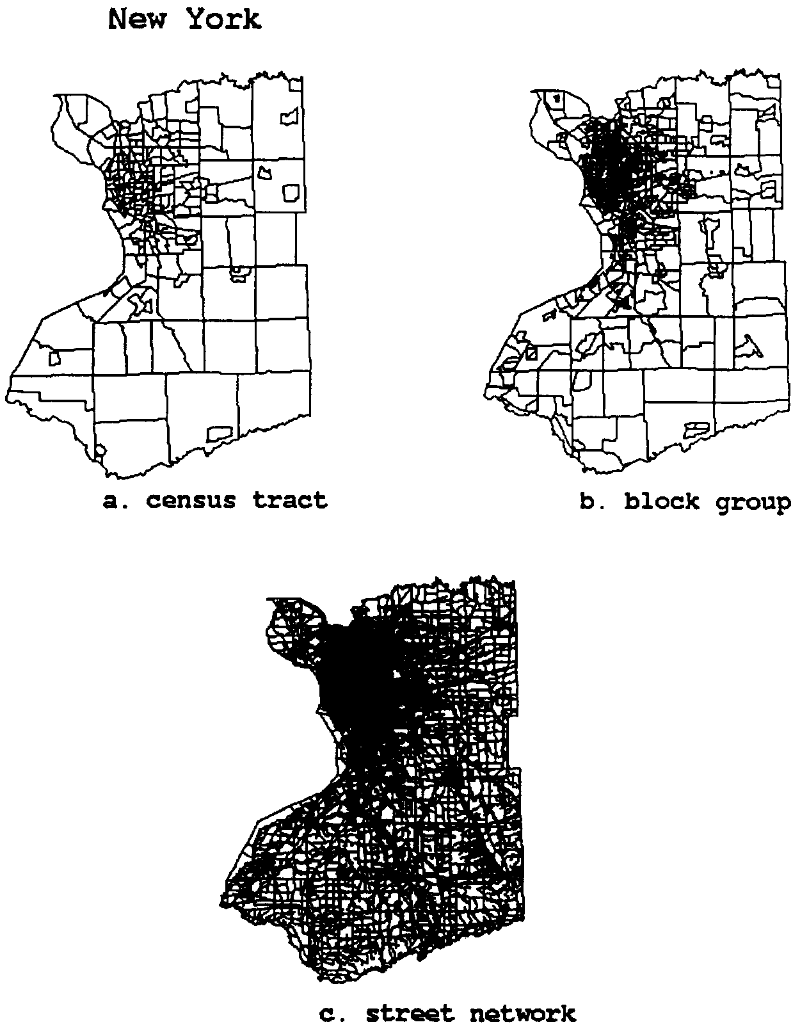


FIGURE 8. The Experimental Area—Erie County, New York

groups and can be compared to interpolated results so that the performance of different methods can be easily judged. Considering the popularity of the vector-based structure, the area-weighting technique is used to represent the traditional methods. The test is conducted by using ARC/INFO along with user-written C codes. The results are given in Table 2.

The first test compared the mean values of the original dataset for block groups with the interpolated outputs from the area-weighting method and the overlaid network methods. The mean values are exactly the same for all five sets of values, indicating that the total population of the study region is preserved and the procedure's computation accuracy is reliable.

The second test examined the *variance, standard deviation*, and *standard error of mean* statistics. The values of these statistics for the area-weighting method is much larger than for the network methods. The improvement for the network length method is particularly significant. The network hierarchical weighting method has the lowest values of the variance-related statistics. The statistics derived from the NHW method are even better than those from the original census data (Table 2).

Table 2. Comparison of the interpolation results by area-weighting and overlaid network methods

Statistics	Observed	Area-Ratio	Net-Length	Net-Weig.	Net-House
	1	2	3	4	5
Mean	994.39	994.39	994.39	994.39	994.39
Variance	446933.26	722695.38	519696.71	433868.81	453022.90
Std Dev	668.53	850.12	720.90	658.69	673.07
S.E. Mean	21.42	27.24	23.11	21.11	21.57
Maximum	5354.00	7591.00	6606.00	5922.00	5722.00
Minimum	0.00	0.00	0.00	0.00	0.00
Range	5354.00	7591.00	6606.00	5922.00	5722.00
Kurtosis	8.27	9.27	8.87	8.86	8.98
Skewness	2.21	2.48	2.36	2.34	2.36
S.E. Kurt	0.16	0.16	0.16	0.16	0.16
S.E. Skew	0.08	0.08	0.08	0.08	0.08

[a] Source: calculated by the authors.

These findings suggest that there is room for improvements in existing census population estimation and data processing techniques. The network hierarchical weighting algorithm has the potential of making an important contributions to census data processing. However, the network housing–bearing method does not exhibit the refinement that would be expected by its underlying rationale. The unexpected performance of the network housing–bearing method may be ascribed to the following causes: (1) some street segments in newly developed suburbs have not been assigned addresses; (2) there are no street addresses in rural areas (US Department of Commerce, 1988); (3) some street segments are numbered by post box numbers where there are no street addresses; and (4) street addresses do not always remain sequential along the street and some parcels have many street addresses (e.g. apartments and condominiums). But the overall performance of the network housing–bearing method is very satisfactory. The range of values for the housing–bearing method is the narrowest (Table 2). In short, the performance of the network methods are satisfactory since all statistics related to variance analysis are improved obviously when compared with the area–weighting method.

The third test surveyed the statistics measuring the distribution and shape of the data (the last four indices in Table 2), which have no significant variations. These findings suggest that the sample distributions for the observed values and the estimation methods are similar, which indicates that the trends and patterns of error disturbances from the two types of algorithms are very similar. The improvement of the network methods is reflected in the reduced dispersion of the values, compared with the areal weighting method. This indicates that the interpolated values from the network methods are more accurate.

CONCLUSIONS AND RESEARCH DIRECTIONS IN FUTURE

The areal interpolation problem occurs when data from two or more incompatible zonal (areal) systems are combined. This is an important issue for GIS data processing and for the spatial analysis of socio-economic data in a wide range of practical applications.

The generally available methods attempt to partition the values of an attribute on the basis of the information on the variable itself. This approach is constrained by the very limited information that is available and produces a very crude spatial interpolation. Methods which use grid or raster-based techniques are generally unavailable in practice. Statistical approaches incorporate ancillary information and attempt to construct a statistical model to obtain estimates for the target zones. These methods' nature depends heavily on the characteristics of ancillary variables and limits their application to certain cases.

Areal interpolation approaches based on *the overlaid network algorithms* try to incorporate the most relevant population related information—street networks—to develop new ways of partitioning population values. It also uses an important GIS technique—overlaying —to facilitate the partition of the attribute values and the transfer of data between the source and target zones. As a result, the overlaid network approach makes efficient use of modern analytic methods and the abundance of digital information in an integrated way. It encourages a thorough understanding of the research problem, the availability of data, and the feasibility of techniques.

The overlaid network algorithms produce much more accurate interpolations than the generally used areal weighting method. Though the network length algorithm is theoretically unsophisticated, its estimation accuracy is quite impressive in contrast to the area-weighting method. The assignment of hierarchical weights to various classes of streets and roads provides the best estimates. As a result, this method may become a very useful method for handling spatially-related data. Further experiment should be conducted on the hierarchical weighting method so that the suitability and stability of its estimation could be clarified. Moreover, the network housing–bearing method generates satisfactory results though it does not produce the best approximates to the areal population interpolations. Because of the complexities of the real world, its performance is a little worse than expected in theory.

Another potential application of the overlaid network algorithms is providing a new solution to the modifiable areal unit problem (MAUP) in spatial analysis and dynamic modeling. MAUP refers to the sensitivity of analytic results to the definition of areal units for which the data are collected (Batty & Sikdar, 1982a, 1982b, 1982c, 1982d; Openshaw, 1977, 1978, 1984; Fotheringham & Wong, 1991). However, the current literature that deals with MAUP subjects to the constraint of keeping unity of all levels of areal units when processing spatial aggregation. It will be preferable if the spatial aggregation is based on social and economic constraints and not constrained by the limitation of areal boundaries. However, few attempts have been made to assign or apportion data values from an areal unit to several zones because of uncertainty and controversy in solving the areal interpolation problem. This may be the annoying root of MAUP. The overlaid network algorithm disaggregates an areal value to a set of basic (very short line) segments. These numericalized line segments can be treated as basic units for inventive aggregations, which can form new areas according to user-defined purposes. The overlaid network approach thus enables the construction of *scientific areal cohorts* on the basis of socio-economic constraints and may overcome the problem of MAUP.

The overlaid network algorithm is a deterministic method. It is built on the accessibility of available data. At present it is only investigated on the data of population. Its applicability for other types of data needs to be clarified in future.

ACKNOWLEDGEMENTS–This research was supported by a grant from the National Science Foundation (SBR88-10917) to the National Center for Geographic Information and Analysis and the support of the NCGIA at Buffalo is gratefully acknowledged. The comments on the manuscript provided by Dr Richard E. Klosterman of the University of Akron is also appreciated.

REFERENCES

Abler, R. F. (1987). The National Science Foundation National Center for Geographic Information and Analysis.

International Journal of Geographical Information Systems, 1, 303–326.

Aronoff, S. (1989). *Geographic information systems: a management perspective.* Ottawa: WDL Publications.

Amrhein, C. G. & Flowerdew, R. (1989). The effect of data aggregation on a Poisson regression model of Canadian migration. In M. F. Goodchild and S. Gopal (Eds)*Accuracy of Spatial Databases* (pp. 21–34), London: Taylor and Francis Ltd.

Batty, M. & Xie, Y. (1994a). Modeling inside GIS: part I: model structures, exploratory spatial data analysis and aggregation. *International Journal of Geographical Information Systems, 8*, 291–307.

Batty, M. & Xie, Y. (1994b). Modeling inside GIS: part II: selecting and calibrating urban models using arc-info. *International Journal of Geographical Information Systems, 8*, 451–470.

Batty, M. & Sikdar, P. K. (1982a). Spatial aggregation in gravity model: 1. An information-theoretic framework. *Environment and Planning A, 14*, 525–553.

Batty, M. & Sikdar, P. K. (1982b). Spatial aggregation in gravity model: 2. One-dimensional population density models. *Environment and Planning A, 14*, 629–658.

Batty, M. & Sikdar, P. K. (1982c). Spatial aggregation in gravity model: 3. Two-dimensional trip distribution and location models. *Environment and Planning A, 14*, 795–822.

Batty, M. & Sikdar, P. K. (1982d). Spatial aggregation in gravity model: 4. Generalizations and large-scale applications. *Environment and Planning A, 14*, 795–822.

Bureau of the Census, US Department of Commerce (1988). Tiger/Line file: Boone County, Missouri. Technical document, Washington, D.C.

Bureau of the Census, US Department of Commerce (1990). Census'90 basics. Washington, D.C.

Bureau of the Census, US Department of Commerce (1991a). TIGER/LINE^{TM} census files, 1990: technical documentation. Washington, D.C.

Bureau of the Census, US Department of Commerce (1991b). 1990 census of population and housing: summary tape file 1: technical documentation. Washington, D.C.

Burrough, P. A. (1986). *Principles of Geographic Information Systems for Landuse Resources Assessment.* Oxford: Clarendon Press.

Davis, J. C. (1986). *Statistics and Data Analysis in Geology*, 2nd edition. New York: Wiley.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal, Royal Statistical Society B, 39*, 1–38.

Dutton-Marion, K. E. (1988). Principles of interpolation procedures in the display and analysis of spatial data: a comparative analysis of conceptual and computer contouring. Unpublished Ph.D. thesis, Department of Geography, University of Calgary, Calgary, Alberta.

Environmental Systems Research Institute (1991a). *ARC/INFO command references: grid^{TM} command references.* Redlands, CA.

Environmental Systems Research Institute (1991b). *ARC/INFO user's guide: cell-based modeling with grid^{TM}.* Redlands, CA.

Flowerdew, R. (1988). Statistical methods for areal interpolation: predicting count data from a binary variable. Research Report No 6, North West Regional Research Laboratory, Lancaster University.

Flowerdew, R. & Green, M. (1989). Statistical methods for inference between incompatible zonal systems. In M. F. Goodchild and S. Gopal (Eds),*Accuracy of Spatial Databases* (pp. 21–34). London: Taylor and Francis Ltd.

Flowerdew, R. & Green, M. (1990). Inference between incompatible zonal systems using the EM algorithm. Research Report No 6, North West Regional Research Laboratory, Lancaster University.

Flowerdew, R., Green, M., & Kehris, E. (1991). Using areal interpolation methods in geographic information systems. *Papers in Regional Science, 70*, 303–315.

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

Fotheringham, A. S., & Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A, 23*, 1025–1044.

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

Goodchild, M. F., & Hosage, C. (1983). On enumerating all feasible solutions to polygon aggregation problems. *Modeling and Simulation 14, Proceedings of the 14th Annual Pittsburgh Conference on Modeling and Simulation* (pp. 591–595.

Goodchild, M. F., & Gopal, S. (1989). *Accuracy of Spatial Databases.* London: Taylor and Francis Ltd.

Goodchild, M. F., & Kemp, K. K. (1990). *Technical Issues in GIS: NCGIA Core Curriculum.* Santa Barbara, CA: NCGIA, University of California.

Green, M. (1989). Statistical methods for areal interpolation: the EM algorithm for count data. *North West Regional Research Laboratory Lancaster University: Research Report No 3.*

Klosterman, R. E., & Xie, Y. (1992). TIGER: a primer. *MicroSofterware News*, *9*, 1–5.

Lam, N. (1983). Spatial interpolation methods: a review. *The American Cartographer*, *10*, 129–149.

Lovett, A., & Flowerdew, R. (1989). Analysis of count data using Poisson regression. *Professional Geographer*, *41*.

MacDougall, E. B. (1976). *Computer programming for spatial problems*. London: Arnold.

McBratney, A. B., & Webster, R. (1986). Choosing functions for semi-variograms of soil properties and fitting them to sampling. *Journal of Soil Science*, *37*, 617–639.

Oliver, M. A. (1990). Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, *4*, 313–332.

Openshaw, S. (1984). *Concepts and techniques in modern geography, Number 38, the modifiable areal unit problem.* Norwich: Geo Books.

Openshaw, S. (1978). An empirical study of some zone-design criteria. *Environment and Planning A*, *10*, 781–794.

Openshaw, S. (1977). Optimal zoning systems for spatial interaction models. *Environment and Planning A*, *9*, 169–184.

Openshaw, S., Wymer, C., & Charlton, M. (1986). A geographical information and mapping system for the BBC domesday optical discs. *Transactions, Institute of British Geographers*, *11*, 296–304.

Sampson, R. J. (1978). *Surface II, revised edition*. Lawrence, KA: Kansas Geological Survey.

Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, *74*, 519–30.

Veregin, H. (1989). Error modeling for the map overlay operation. In *Accuracy of Spatial Databases*, Goodchild, M. F., and Gopal, S. (eds.), 3–18. London: Taylor and Francis Ltd.