

A Methodology for Estimating Small-area Population by Age and Sex Based on Methods of Spatial Interpolation and Statistical Inference

Qiang Cai

Department of Geography, the University of Iowa, Iowa City, IA 52242

qiang-cai@uiowa.edu

Abstract

The objective of this research is to estimate population data for small areas whose boundaries are different from those for which the data were originally observed. The problem belongs to the class of “change of spatial basis” problems in geographic studies. Solutions to this problem involve spatial interpolation techniques augmented by statistical principles. From Census population data, three different continuous distributions of age-sex weights are computed and then used to estimate population detail for small areas. Three spatial interpolation methods: the areal weighting method, the inverse distance weighting method (IDW) and a modification of Tobler’s (1979) pycnophylactic method are used to create the three distributions. A statistical model based on the central limit theorem is used to compute statistical bounds for the population estimates of small areas. The three methods were implemented on age-sex population data for Census Tracts in Iowa together with population totals for 90 meter squares (LandScan U.S.A. data). The methods were then used to compare age-sex population estimates for paired Block Groups that straddled Census Tract and therefore were spatially misaligned. In this test, the pycnophylactic method and the areal weighting method were more accurate than the IDW method. The method is general and can be used to transform data values from observed spatial entities to unobserved spatial entities where change of spatial basis is required. The application that motivated this study was the mapping of cancer rates which are known to be sensitive to the spatial scale of the population data on which they are based.

1. Introduction

This paper was motivated by my work related to a research contract with the National Cancer Institute (NCI). In that project, age-sex populations for small-areas in Iowa are critical values for many consequent analysis tasks such as computing age-sex adjusted cancer incidence rates for flexibly defined areas. The method developed here is capable of transforming data values from observed spatial entities to unobserved spatial entities where change of spatial basis is required.

The central concept for understanding the problem in this paper is ‘spatial basis’, which is defined as a set of discrete spatial objects (points, lines, and areas) used to describe the spatial variation of variables (Goodchild et al. 1993). It explains how attributes are associated with space- by area (Census data, for example) or by location (like environmental monitoring data), small scale or large scale. Many socioeconomic spatial data, however, are usually collected and reported by different agencies using spatial zoning systems for their own convenience or other concerns (like confidentiality). For a variety of reasons agencies find themselves working with

spatially incompatible units for different socioeconomic spatial data. Examples of spatially incompatible spatial basis include counties, metropolitan statistical areas, labor-market areas, hydrologic study areas, Zip Code areas etc. The direct use of these data associated with usually arbitrarily defined spatial basis in regional or spatial analysis leads to intractable errors, which is known as the modifiable areal unit problem --MAUP (Openshaw, 1984). There is the need to transfer spatially distributed variables from a set of source zones (or points) to target zones (or points) so that they can have the same spatial basis for further analysis. The process involves change of spatial basis. Methods for making such transfers are generally called spatial interpolation. Examples of spatial interpolation methods include Inverse Distance Weighting (IDW), kriging, areal smoothing, areal weighting, dasymetric mapping, Bayesian mapping etc (Mugglin et al. 1999).

The research problem, which this paper tackles, is to find an approach to compute estimates of small-area populations by age and sex with reliable quality based on population data from the Census and other sources. Obviously, this belongs to the problem of change of spatial basis (from Census zoning systems to custom-defined small-areas on which our research interests focus). It is natural, therefore, to look for solutions from the broad area of spatial interpolation. In addition to what traditionally spatial interpolations do, where attribute estimates for new spatial basis are often single optimal values based on the specific interpolation method, we would like to incorporate suitable statistic models into the interpolation process to compute error bounds for the estimates as well. Through this approach, we can compute the range that our estimates are likely to fall in given that assumptions of the statistical model have been satisfied. Such results will serve further analysis better.

For the rest of this paper, Section Two reviews various methods in spatial interpolation, Section Three explains my research scenario and how spatial interpolation methods are selected and implemented in this project. Estimation results are described as well. In the Fourth section, estimation errors are discussed and results from a verification test are described. Section Five evaluates the main contribution of this paper and points out problems for further studies.

2. Methodology Review

There are many spatial interpolation methods available and this section mainly focuses on methods that are closely related to the problem of making population estimates for small-areas.

2.1 Surface-Oriented approaches

A straightforward strategy in spatial interpolation is to develop a continuous surface representation from attributes associated with source zones so that interpolated attribute values change gradually across the space. For attributes that can be portaged as densities per unit area (like population), the task of computing estimates for target zones is then simplified to an integration of the “density” surface over target zones. This view of population as a continuously varying phenomena across space is widely accepted (Unwin, 1981). Methods belonging to this surface approach in the literature are generally classified into two categories: point-based and area-based.

The class of point-based methods create surfaces from centroids using the general idea that unknown values at locations are determined by known values at centroids (or points at locations) surrounding them according to their relative distance and spatial configuration. Main point-based methods include Inverse Distance Weighting interpolation --IDW (Bracken and Martin, 1989; Bracken, 1993; Mennis, 2003) and kriging (Cressie, 1993; Gotway, 2002; Gribov, 2004).

The IDW method uses a distance-decay weighting function to determine weights of known centroids for predicting values at positions where observations are not available. It has been widely used to create smooth surface representation and is also chosen as one of the three interpolation methods used in this paper. However, IDW is not capable for accurate estimation due to several problems: no mass-preservation property; distortion in edge areas; unable to handle direction information.

The kriging method computes an empirical semivariogram based on observed data to estimate spatial autocorrelation of the interested variable. Based on the spatial autocorrelation evaluated, estimates at unobserved positions can be predicted with minimal kriging error. However, for socioeconomic variables like population which is discrete and often measured by region, kriging is rarely used though the spatial autocorrelation function that is a property of the kriging method is often used.

Polygon-based surface interpolation methods begin from observations associated with a set of disjointed polygons rather than a set of points; so the size, shape and spatial arrangement of source regions affect interpolation results. Tobler's (1979) pycnophylactic interpolation method is the most well-known polygon-based surface interpolation method. The major advantage of this method is that it can smooth the abrupt changes along region boundaries which are quite common and misleading in choropleth maps while keeping the mass-preserving property. Here smooth means gradual change of values across space and mass-preserving means aggregated count value within a source region does not change before and after the interpolation process. Smoothness is achieved by minimize the sum of the squares of all the partial derivatives across the whole interpolation area. The mass-preserving property is achieved by redistributing the known difference between the observed mass of the source region and its interpolated counterpart by the area of source region after the smooth step. When the phenomenon represented changes gradually over space in nature, this method seems appropriate. Later in this paper, we show that Tobler's method requires modification in order to meet defined secondary conditions as well as the one-dimensional mass property.

2.2 Zone-Oriented approaches

Zone-oriented interpolation methods view space as a set of disjointed zones that the values of concerned variables are evenly distributed within each disjointed zone. In geographic reality, especially for socioeconomic data, a zoning system with approximately homogeneous in-zone values is possible if the zone is defined properly to follow clearly distinguishable socioeconomic feature (urban buildup area vs. agriculture area, for example). Important zone-oriented interpolation methods include areal weighting method (Goodchild and Lam, 1980;

Flowerdew and Green, 1992), dasymetric mapping (Langford and Unwin, 1994, Eicher and Brewer, 2001; Mennis, 2003), statistical regression modeling (Flowerdew and Green, 1992; Mugglin, 1999, 2000; Gotway and Linda, 2002) etc.

Areal weighting method assumes homogeneous in-zone distribution for source zones or target zones, which makes the interpolation step very straightforward. Since this assumption can be very weak, it is mainly used to test the effectiveness of other interpolation methods.

The technique of dasymetric mapping was initially introduced by Wright (1936). A dasymetric map is a map of zones with zonal boundaries that reflect actual changes of the mapping variable. This usually refers to approximately homogeneous intra-zone distribution and abrupt inter-zone distribution. For the problem of population interpolation, the dasymetric mapping method is widely used and proved effective when high quality dasymetric maps (dasymetric polygons usually does not coincide with source or target zones) of population related variables are available. Langford and Maguire (1991) and Langford and Unwin (1994) discussed advantages of using dasymetric mapping for the purpose of population interpolation. They used a land-use map derived from remotely sensed images as an ancillary data source to redistribute the census population data which were tabulated for census units. Different land-use types were classified into two categories as “residential” and “unoccupied” and population were only assigned to residential classes. The weights assigned to the two types of dasymetric map polygons are relatively arbitrary. Jeremy (2001) refined the weight assigning procedure by using empirical sampling to determine appropriate percentage assignment values to different land-use types.

Statistical regression modeling methods are used to establish a regression relationship between the variable of interest and one or more ancillary variables. Mugglin and Carlin (1999, 2000) introduced Bayesian areal interpolation/estimation as an effective way to incorporate ancillary information for spatial interpolation/estimation. They divided source zones into small regular shaped subregions, and assumed that the distribution of the variable of interest at each sub-regions have conditionally independent distributions given the covariates values. Monte Carlo Markov chain-- MCMC technique is used to estimate the empirical posterior distribution parameters given the observations of covariates.

2.3 Summary of different spatial interpolation methods

Two main views on spatially distributed variables, i.e. continuous space with gradual change vs. disjoint zones with abrupt change, lead to various methods falling into two major categories.

Surface-oriented interpolation methods generally fit the first law of geography and therefore are sensible for many geographically distributed variables. They can create smooth surfaces using observations from source zones (locations), good for representing spatial change of continuous variables. Zone-oriented methods deal with spatial variables that can be organized by zones with interior homogeneous distribution. Many socioeconomic variables fit this concept. When our interest is to show the change of population density across space, a surface representation seems good. When accurate population estimates are needed and highly correlated ancillary data with high spatial resolution are available, it seems better to divide space into small zones

that have approximately homogeneous distributions inside and compute estimates using either a regression model or a dasymetric mapping approach. The interesting thing is, when the zoning division is fine enough –like pixels in a remote sensing image, the result with the form of a high-resolution grid is approximately a continuous surface representation. Also, surface interpolation methods have the potential to smooth results from zone-oriented interpolation. Selection and combination of different spatial interpolation methods should always fit with the characteristics of the source data and the purpose of the study.

3. Problem Setup – Objective, Data, Methodology, and Implementation

3.1 Objective

The objective of this research paper is to develop a spatial interpolation method that is capable of estimating small-area, age-sex populations for Iowa based on available population data. The methodology developed should follow or satisfy the following rules or properties:

1. It should be based on well established spatial interpolation methods.
2. It should be able to fully utilize all population related information available.
3. It should be properly calibrated to meet the specific objective.

The three requirements, in some sense, propose an optimal solution given the data and techniques available. We also want the method developed here to handle a class of problems rather than just age-sex population estimation for small-areas, i.e., to have some breadth in application. Finally, it is desirable that our work here provides some original thoughts on the research area of spatial interpolation and consequently contributes to method development in this area.

3.2 Data

In previous discussions, we established the idea that the choice of method is determined by properties of data used. Now it is time to do some analyses on the dataset used.

There are two major data sources for this research work, U.S. Census 2000 from the U.S. Census Bureau, and LandScan USA developed by the GIS research group at oak ridge national lab (ORNL). Without specification, all data used in this research always refer to the same spatial coverage- *the state of Iowa*.

Census 2000 data include:

- 1) Tabulation data for the counts of total population and the population for 42 age-sex subgroups (21 age groups for each gender, *table 1*) at different level of Census statistic zoning systems (State-County-Tract-Block Group).
- 2) Cartographic Boundary Files (which are generalized version of Census 2000 TIGER/Line files) corresponding to these census zoning systems with tabulation data (Fig 1).

Although the Census 2000 tabulation data are the best population data available for the given census statistical units, a problem with the Census 2000 population data, which rarely is

referenced in the current literatures, is the purposeful degrading of data released by the U.S Census due to confidentiality concerns. The Census permuted their original household data elements in some Census zones (Block, Block Group, Census Tract, County) before they release their tabulated data for Census zones to prevent individuals from been identified through the data released (Census 2000 Evaluation C.1, 2003). The probability that a household variable element is exchanged with another (partner) household in another zone is based on the risk that publishing the element in question would risk the disclosure of individual data. The consequence of such a process is that Census data contain unknown errors and might not be appropriate to use directly. In the zoning system of the Census, the higher rank a zone unit is, the less possibility that its attributes have been affected by disclosure limitation procedures. However, the spatial resolution is lower at the same time. We have to be careful in using the Census data to balance reliability and resolution.

Cartographic Boundary Files are generalized extract from the TIGER/Line files. The spatial error varies from 1 to 10 second (<http://www.census.gov/geo/www/cob/scale.html>), which is approximately 20~250 meters in Iowa.

LandScan USA population data include:

- 1) Population grid with spatial resolution of 3 second (approximately 90 meter) (Fig 2)
- 2) Population Centroids for various Census zoning systems (State-County-Tract-Block Group-Block) (Fig 1)

The LandScan USA’s 3 second resolution population grid is generated by ORNL using what they call an “intelligent interpolation” method. It is basically a hybrid of dasymetric mapping and regression methods using multi-source data layers like population counts from Census 2000, roads, land cover, nighttime light etc. It provides much more detail on the spatial distribution of population (total counts and not by age-sex) than the census tabulation data. Detail information on this method can be found at Bhaduri et al. (2004) “Development of High Resolution Population Distribution Data to Enhance Cancer Prevention and Control Research” (http://www.uiowa.edu/~gishlth/UIORN/2_SEER_reportl_ornl_0304.pdf).

According to our preliminary evaluation of this grid population data by comparing it with results from human interpretation of orthophoto map done in one county of Iowa, the quality of this population grid is quite satisfactory. The population centroid data is generated by applying commonly used centroid generation algorithm to the population grid, so it can also be deemed as a valid representation of true population centroid.

Table 1 Division of age groups in Census 2000 tabulation data (from P12, SF1, U.S. Census 2000)

Age groups	Under 5	5 to 9	10 to 14	15 to 17	18 and 19	20	21	22 to 24
Age groups	25 to 29	30 to 34	35 to 39	40 to 44	45 to 49	50 to 54	55 to 59	60 and 61
Age groups	62 to 64	65 and 66	67 to 69	70 to 74	75 to 79	80 to 84	85 and above	

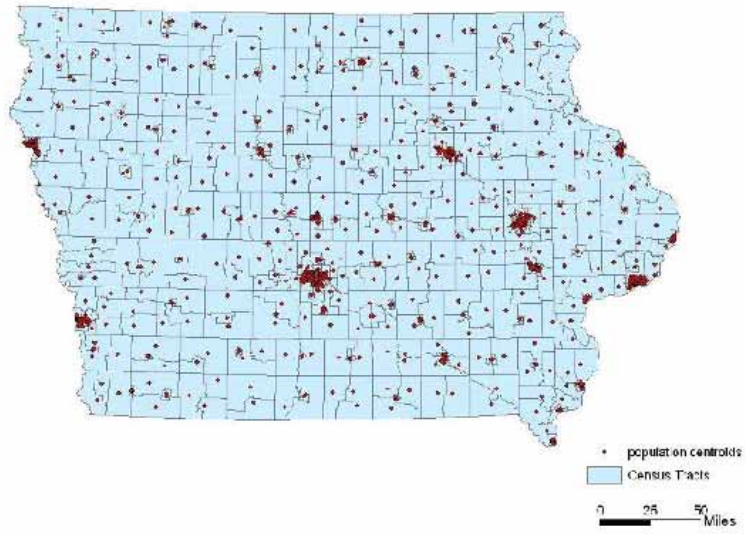


Figure 1 Distribution of Census Tracts and their population centroids in Iowa

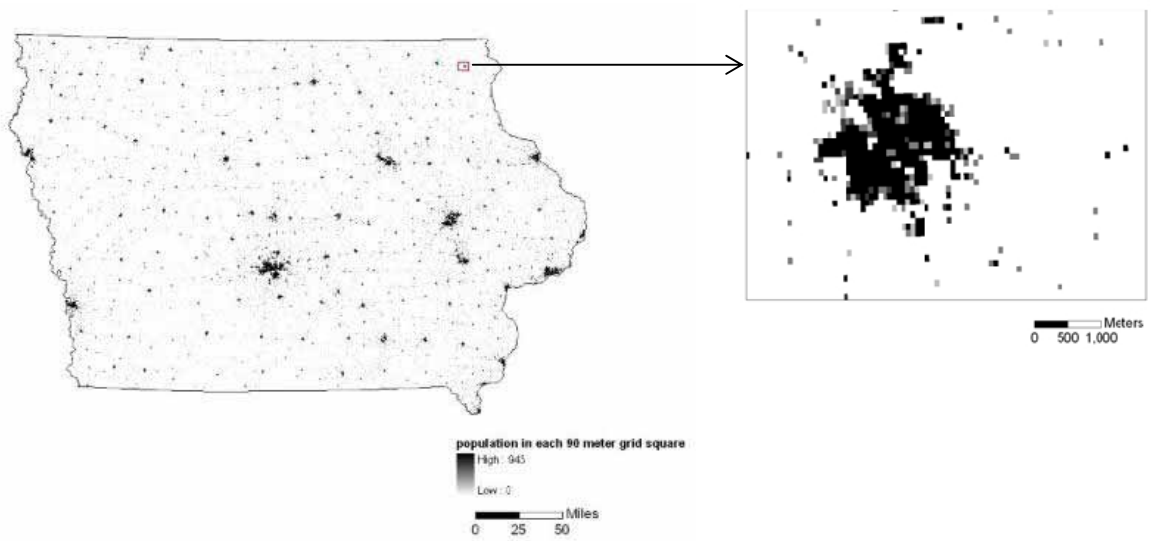


Figure 2 LandScan USA 90 meter population grid, Iowa

3.3 Methodology

We use three spatial interpolation methods to derive age-sex population for small areas in Iowa and test their accuracy against known values for target areas that were not used in any of the estimation methods.

The high resolution population grid from LandScan USA provides very good estimates of total population in the form of an approximately continuous surface. Since population of a certain age-sex group is just a subgroup of total population and can only exist in locations where there are resident people, this population grid can play a critical rule in improving age-sex population estimation. If we can, through some approach, compute the proportion of a specified age-sex population subgroup to total population for each populated cell in the population grid, we can generate an age-sex population grid by simply multiplying the proportion value with population counts at each grid cell. It is then straightforward to derive small-area population estimates from this fine resolution age-sex population grid.

The original data we have on age-sex population distribution are average rates by all levels of census statistic zones. Several spatial interpolation methods are able to create surface representations of the rates from these data. Three spatial interpolation methods are chosen to develop an age-sex rate grid (approximately a surface) with the same resolution and spatial coverage as LandScan USA population grid. They are Areal weighting, IDW, and Pycnophylactic interpolation method. As mentioned in section two, most spatial interpolation methods from geography or cartography (include the three methods) lack the so called “statistical inference” property. Estimates for target zones are just a set of single values, which is kind of arbitrary and there is no way to measure the variability of estimates which obviously will exist. In our research scenario, we fit the age-sex population estimates into a binomial-normal distribution model and both expectation and variance of the population estimate can be derived from this model. A two-step estimation framework is proposed to incorporate the interpolation methods and the statistical model together:

Step one: Create an age-sex “rate” surface for any specified age-sex subgroup (for example, people with ages from 60 to 69) using spatial interpolation. The three interpolation methods selected will create three rate surfaces (fine resolution grids in implementation) separately.

Step two: Assume the rate on each cell of the grids equals the probability that a person resident on that cell (if there are any) belonging to the corresponding age-sex group. For a cell $c(i,j)$ with population N_{ij} and probability Pk_{ij} for the k th age-sex group, the number of people from k th age-sex group in this cell (Xk_{ij}) can be treated as a random variable with a binomial distribution $\text{Bin}(N_{ij}, Pk_{ij})$, just like the classical example of drawing red balls from a bag with fixed proportion of red and blue balls. The probability function of this binomial distribution is:

$$\Pr(Xk_{ij}) = \binom{N_{ij}}{Xk_{ij}} Pk_{ij}^{Xk_{ij}} (1 - Pk_{ij})^{N_{ij} - Xk_{ij}} \quad (3.1)$$

The expectation and variance of Xk_{ij} are:

$$E(Xk_{ij}) = N_{ij} * Pk_{ij} \quad (3.2)$$

$$\text{Var}(Xk_{ij}) = N_{ij} * Pk_{ij} * (1 - Pk_{ij}) \quad (3.3)$$

A target zone can be deemed as the aggregation of many cells with independent binomial distributions. Since the number of 90m cells within a target zone is usually big (one square mile area contains around 400 cells), the sum of these many cells can be approximated to a random

variable with Normal distribution $N(\sum_{s=1}^m E(Xk_{ij}), \sum_{s=1}^m \text{Var}(Xk_{ij}))$ by the Central Limit Theorem.

It is then easy to compute the expectation and confidence interval (CI) for the population of kth age-sex group in any target zones. A 95% CI for this normal distribution will be

$$[\sum_{s=1}^m E(Xk_{ij}) - 1.96 * \text{SQRT}(\sum_{s=1}^m \text{Var}(Xk_{ij})), \sum_{s=1}^m E(Xk_{ij}) + 1.96 * \text{SQRT}(\sum_{s=1}^m \text{Var}(Xk_{ij}))].$$

And since population estimate can only take nonnegative integer values, proper adjustment may be necessary for the resulting CI ranges.

3.4 Implementation

In this case study, census tracts/ population centroids of tracts in Iowa are used as source zones/points since they are fairly small to capture the spatial variations across the state and just big enough to avoid the disclosure limitation problem discussed in section 3.2. There are 794 tracts in Iowa. Two tracts have zero population so no population centroids are applicable for them and their age-sex percentages are zero. Another tract has only 54 people and its population and polygon area are merged into the neighboring tract with the nearest population centroid. So we actually start with 793 (= 794 - 1) tract polygons and 791(= 794 - 2 - 1) tract population centroids (Fig 1). For each tract polygon/centroid, percentage of male with age 55 to 64 in each tract are used as the initial input of the spatial interpolation process as an example, and this process can easily be expanded to any age-sex group of interest. The original percentage value is transformed into integer by first multiplying with 1000 and then rounding into the nearest integer value to facilitate grid computation and map illustration. The values in percentage maps (Fig 3, 4, 5) are these integers rather than actual values.

In Areal weighting, the initial percentage value in each tract polygon is treated as the percentage of every location within that polygon. Then this percentage polygon layer is rasterized into a 3-second cell size rate grid. Cell value ranges from 0 to 73 (Fig 3).

In IDW interpolation, we start from 791 population centroids (from LandScan USA, Fig 1) with respective tract-average percentage value. A standard IDW model from ArcGIS software is applied on these centroids with the constraints of eight nearest points (two from each of the four directions) and the distance decay function has the form of a second-order polynomial equation. The resulting rate surface is then rasterized into a 3-second cell size rate grid (Fig 4).

The pycnophylactic interpolation begins from the 793 tract polygons with associated tract-average percentage values. A revised pycnophylactic interpolation algorithm based on Tobler's article in 1979 and the source code of his original algorithm is developed. The surface we smoothed is the percentage surface, while the mass to be preserved is the counts of age-sex subgroup population. The technique to compromise between the two is to smooth the percentage grid first and then transfer it into the population count grid to estimate the gap between counts after smooth and real counts, and then figure out the respective adjustment on the percentage value and implement that adjustment on the percentage grid to fulfill the mass-preserving property on the population count constraint. Smoothed percentage surface and related algorithm analysis are shown in (Fig 5 - 7)

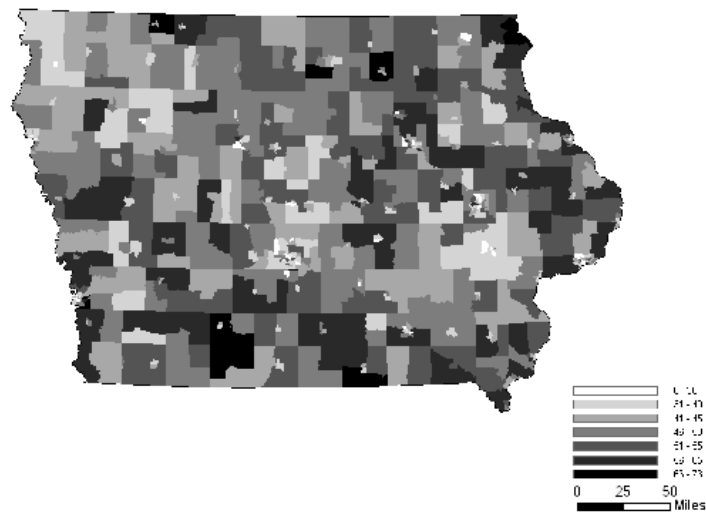


Figure 3 “percentage” surface of males age 55 to 64(Areal weighting interpolation)

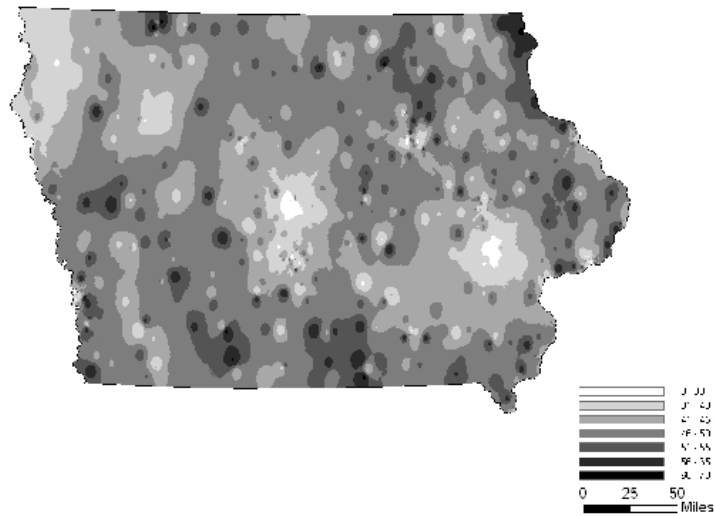


Figure 4 “percentage” surface of males age 55 to 64 (*IDW interpolation from population centroids with average population percentage value of each Census Tract*)



Figure 5 “percentage” surface of males age 55 to 64 (*pycnophylactic interpolation from average population percentage value of each Census Tract – Fig 3*)

It is necessary to do some analysis on results from this revised pycnophylactic interpolation to make sure desired properties like smoothness and mass-preserving are achieved. The area in the box in Fig 5 is zoomed in and illustrated in Fig 6 below to show the smooth effect.

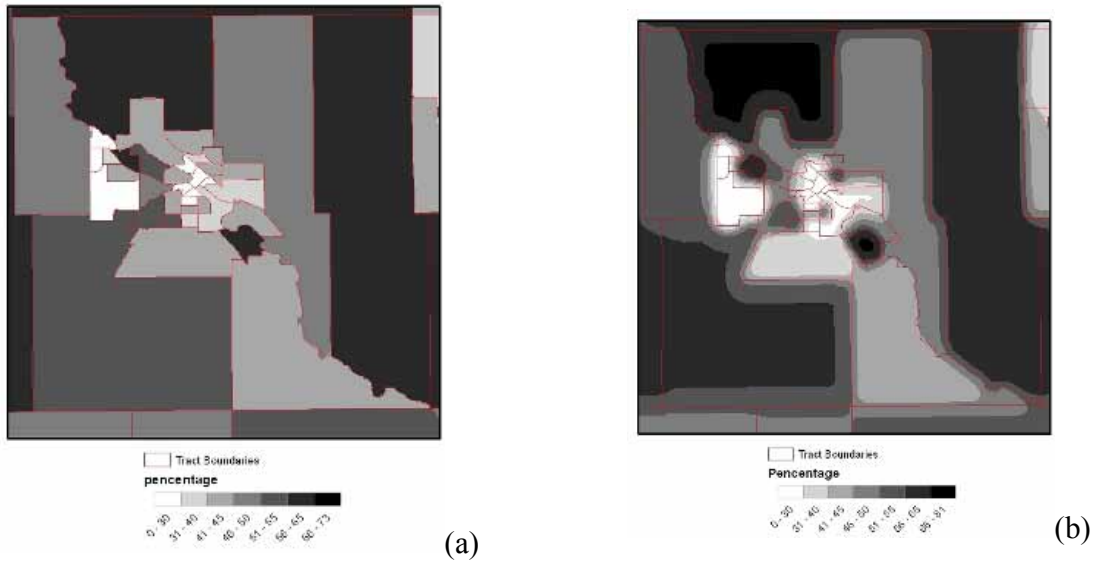


Figure 6 Smooth effect of the Pycnophylactic Interpolation

Fig 6(a) shows values of the percentage surface (male, age 55 to 64) of a sample area in Iowa before pycnophylactic interpolation, while Fig6(b) shows the smoothed percentage surface after interpolation. Quantitatively, smoothness can be shown by a measurement called Relative Square Roughness (Tobler, 1994), which is defined as:

$$\text{Roughness}(i,j) = P_{i,j} - 0.25 (P_{i,j+1} + P_{i,j-1} + P_{i+1,j} + P_{i-1,j}) \quad (3.4)$$

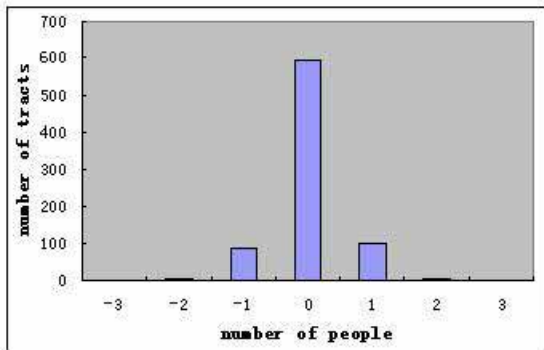
$$\text{Square Roughness} = \text{sum}(\text{Roughness}(i,j) * \text{Roughness}(i,j)) \quad (3.5)$$

$$\text{Relative Square Roughness (after iteration } n) = \text{Square Roughness}(n) / \text{Square Roughness}(0) \quad (3.6)$$

In this example, the relative square roughness is reduced to 0.002 after 267 iterations.

The mass-preserving property is verified by showing the difference between population estimates in each tract from pycnophylactic interpolation and estimates from areal weighting interpolation which is known to have mass-preserving property (Fig 7).

$$\text{Population estimate (tract } n) = \text{Sum} (\text{Population}_{ij} * \text{Percentage}_{ij}) \quad (3.7)$$

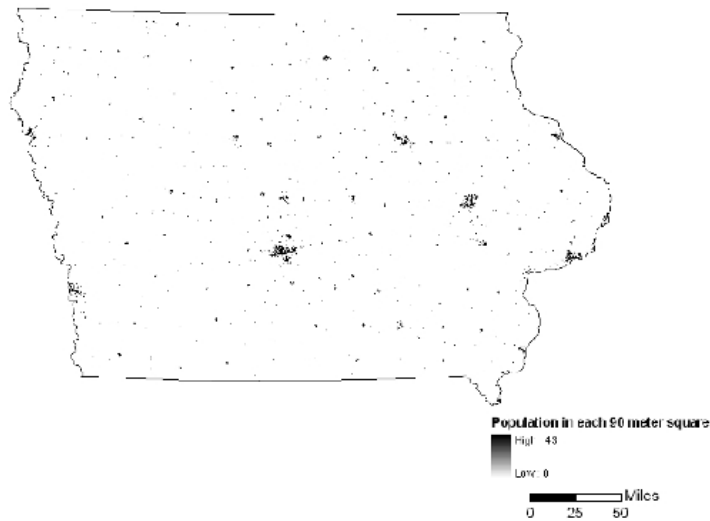


number of people	tracts
-2	4
-1	89
0	593
1	103
2	4

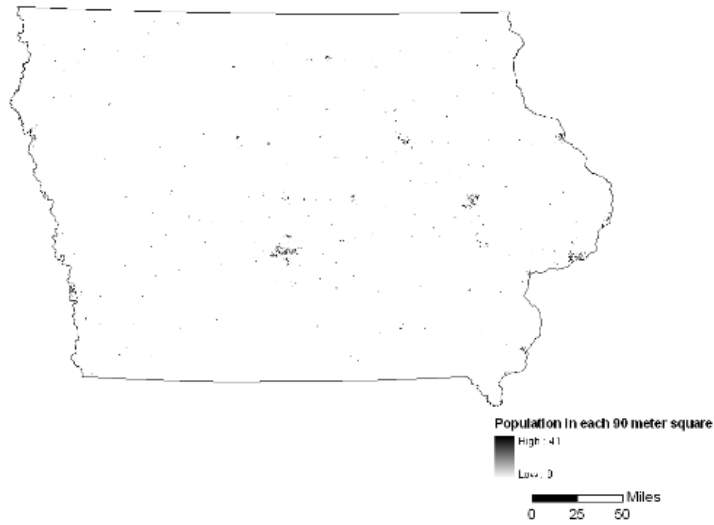
Figure 7 Result of test of mass-preserving property: Difference between population estimates (males age 55 to 64) of the 793 census tracts using Pycnophylactic interpolation and areal weighting interpolation

It is quite impressive that differences of estimates between the two in almost 80% of all tracts are zero and the maximal difference is just two. This proves the mass-preserving property of the revised pycnophylactic interpolation.

So far, three “percentage” surfaces have been created using the three different spatial interpolation methods (Fig 3, 4, 5). Following the two-step inference framework (pp 8 - 9), it is very easy to compute the mean and variance surface of population estimates from them. The values for pycnophylactic interpolation are illustrated in fig 8 below.



(a)



(b)

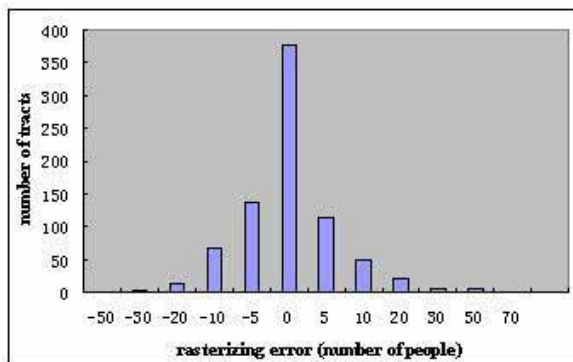
Figure 8 Mean (a) and Variance (b) population grids of males between 55 and 64 using Percentage surface from pycnophylactic interpolation -- Fig 6

4. Discussions

4.1 Sources of errors

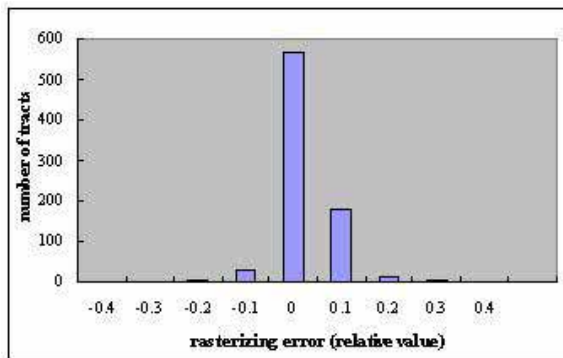
The goal of this research is to compute population estimates with least estimation error. There are two types of error sources: the rasterizing process and the interpolation process.

Error from the rasterizing process happens whenever vector data are transformed into raster data. The cell size determines the rasterizing error. To examine the scale of this type of error in our case where cell size is 3 seconds (approximately 90 meters), population estimates (male between 55 and 64) of tracts using areal weighting method, which is known to have mass-preserving property are computed. Rasterizing errors of population (absolute error equals areal weighting population estimates minus actual population of the same age-sex group) by tracts are plotted in Fig 9 below.



(a)

number of people	tracts
-30	3
-20	14
-10	68
-5	136
0	377
5	114
10	50
20	22
30	4
50	4
70	1



(b)

relative error	tracts
-0.4	1
-0.3	1
-0.2	2
-0.1	29
0	565
0.1	179
0.2	10
0.3	5
0.4	1

Figure 9 Illustration of rasterizing errors of population by tracts -- absolute errors (a); relative errors (absolute error divided by actual tract population) (b)

Fig 9(a) indicates that absolute rasterizing error in population estimation of around 89% of all tracts are within [-5, 5], and in about 6% tracts are over 10. Fig 9(b) supplies some complimentary information from another perspective. An interesting thing about it is that positive error dominates negative error when represented as relative error. In summary, the rasterizing error has some contribution to total error but generally will not be too serious.

Error from the interpolation process is determined by factors like assumptions and interpolation techniques of the interpolation method, characteristic of the spatial variables, etc. In our case, spatial variation of age-sex proportion is the unknown variable to be estimated. The areal weighting (AW) method assumes homogeneous age-sex proportion within each census unit, the IDW and pycnophylactic interpolation assumes gradual transition of age-sex proportion across space. The actual spatial distribution of age-sex population might be affected by multi-factors. Additional information like local resident type may be useful, like we expect there will be a lot of people between 18 and 28 in a university town. In our study, such information is not incorporated due to the limitation of available data and technology. So far, what we can do to address the interpolation error is to pull out some test areas with actual population values available within the spatial extent of the state and compare these values with our population estimates from the three interpolation methods.

4.2 Verification: Comparison of results from the three interpolation methods

To verify each of the three methods, we need to find some target zones with known age-sex population values that do not have the same boundaries as the source zones -- that is they are spatially misaligned. Since tracts are source zones and counties are too big, block group is the only suitable choice. To make them spatially misaligned with tracts, we use block group pair which is defined as two adjacent block group polygons that belong to two different tracts. 34 block group pairs are randomly chosen as the test area (Fig 10). Population estimates and upper and lower bound of the 95% Confidential Interval (CI) in these block group pairs are then computed by applying the equations in section 3.3 on the mean and variance grids produced before (see Fig 8).

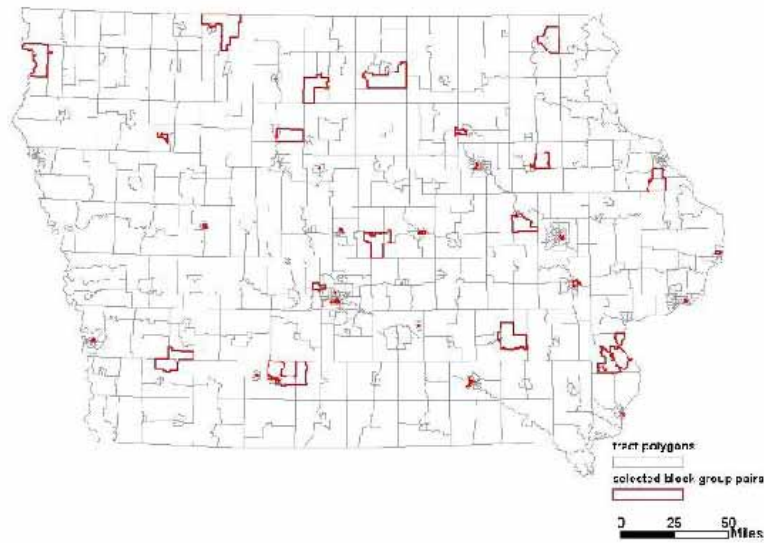
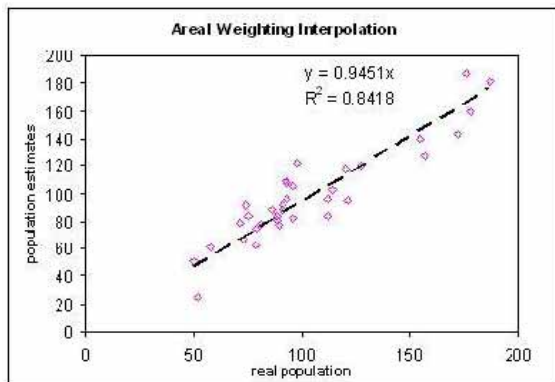
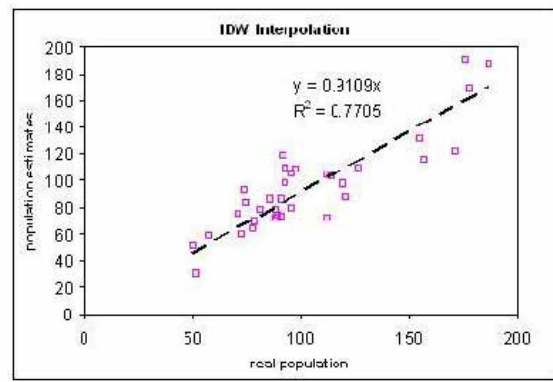


Figure 10 Map of the 34 selected block group pairs (the area of each pair is found in two adjacent tracts)

For each of the three interpolation method, we have a separate population estimate and corresponding confidential interval for each block group pair. Their relationship with actual block group pair population from census tabulation data are illustrated in the following figures (Fig 11, 12).



(a)



(b)

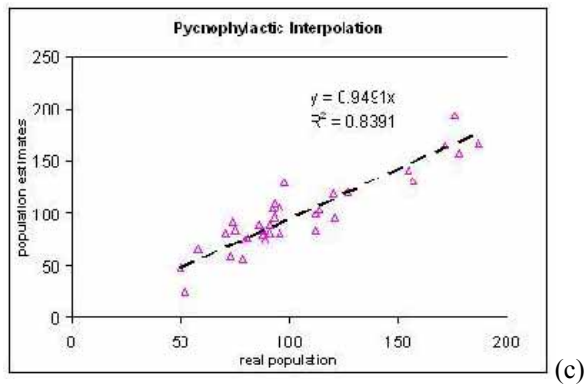
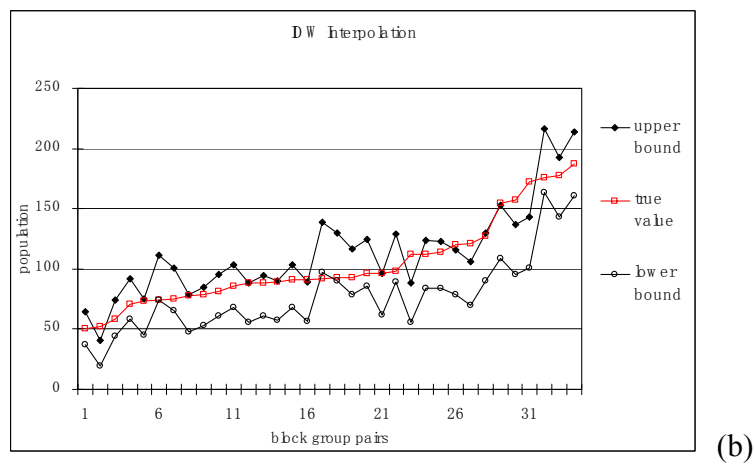
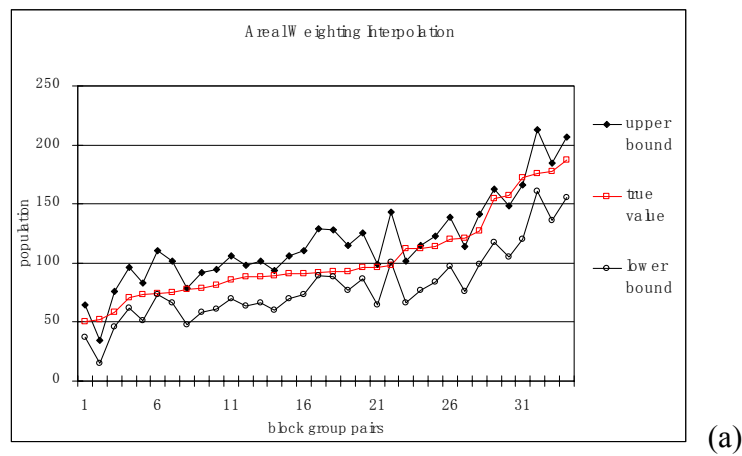
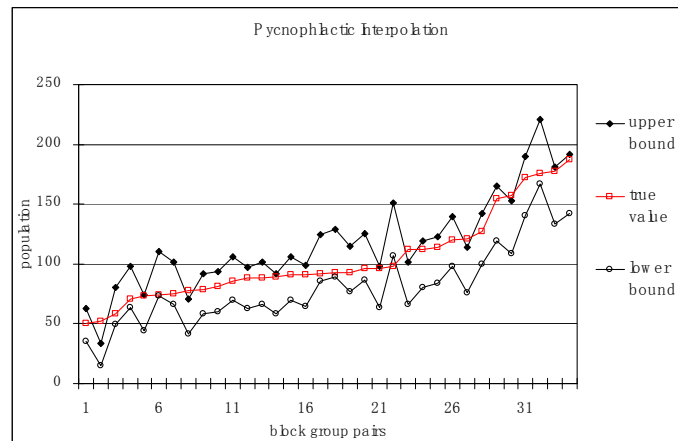


Figure 11 Correlation between actual population in the 34 block group pairs and population estimates from the three spatial interpolation methods - (a) areal weighting interpolation; (b) IDW interpolation; (c) Pycnophylactic Interpolation





(c)

Figure 12 Relationship between actual populations in the 34 block group pairs (ascending order by true population) and 95% confidential intervals of population estimates from the three spatial interpolation methods - (a) areal weighting interpolation; (b) IDW interpolation; (c) Pycnophylactic Interpolation

Fig 11 and 12 collectively imply several interesting things about the effectiveness of the three interpolation methods in small area population estimation within the framework proposed in this paper. (1) Population estimates from pycnophylactic interpolation are closest to actual population according to Fig 11 where the regression coefficient is biggest among the three (0.9491 compare with 0.9451 from areal weighting and 0.9109 from IDW). Notice that all the three are smaller than one, which indicates underestimation. (2) The R-squared values in Fig 11 indicate that Expectations of population estimates pycnophylactic interpolation and areal weighting interpolation have better correlation with actual population than result from IDW interpolation.(0.8391, 0.8418 vs. 0.7705) (3) In Fig 12, 28 out of 34 samples from areal weighting interpolation approach, 24 out of 34 samples from IDW interpolation approach, and 28 out of 34 samples from pycnophylactic interpolation approach have 95% CIs of population estimations that include the corresponding actual population values. (4) The average absolute error between estimates and actual population are 12, 15, and 13 for AW, IDW, and pycnophylactic method. We conclude that areal weighting method and pycnophylactic performs better than the IDW approach. These results may imply that for source areas as small as census tract, the homogeneity assumption is enough to generate fairly good estimates. However, since target zones in this case are quite big compared with source zones (a block group pair is about half the size of a tract on average), part of the error from areal weighting interpolation cancels. When smaller areas are used as target zones, pycnophylactic method might perform better than the areal weighting method.

5 Conclusions

Contributions

- Developed a framework for the problem of small-area estimation that incorporates spatial interpolation method and statistical inference. With this method we can compute composite form of estimates (expectation and variance as well) for a class of problems that depend on rate values that can be interpreted as probabilities of events.
- Revised the Pycnophylactic algorithm to enable it do age-sex population estimation (through the incorporation of age-sex information and total population information).
- Error analysis and a reasonable verification process.

Limitations and problems

- The validity of assumption that rate values can be interpreted as probabilities in some cases might be questioned, yet such assumption is necessary to implement the binomial-normal model developed.
- None of the three interpolation methods are perfect. The experiment shows that without further information on age-sex distribution, the other two methods did not perform better than the areal weighting method in general.
- The approach used here is particularly suitable where population totals for very small areas are available as was the case here with the LandScan USA data.

References

- Bracken, I. and Martin, D. (1989), "The generation of spatial population distributions from census centroid data", *Environment and Planning A*, Vol 21, 537 – 43.
- Bracken, I. (1993), "An extensive surface model database for population-related information: concept and application", *Environment and Planning B: Planning and Design*, Vol 20, 13 – 27.
- Cressie, N. (1993), *Statistics for Spatial data*, Wiley: New York.
- DeGroot, M. H. and Schervish, M. J. (2002), *Probability and Statistics: Third Edition*, Addison – Wesley.
- Eicher, C. L. and Brewer, C. A. (2001), "Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation", *Cartography and Geographic Information Science*, Vol 28, No 2, 125 – 38.
- Flowerdew, R. and Green, M. (1992), "Developments in areal interpolation methods and GIS", *The Annals of Regional Science*, Vol 26, 67 – 78.
- Goodchild, M. F., Anselin, L. and Deichmann, U. (1993), "A framework for the areal interpolation of socioeconomic data", *Environment and Planning A*, Vol 25, 383 – 97.
- Goodchild, M F. and Lam, N. S-N. (1980). "Areal interpolation: a variant of the traditional spatial problem", *Geo-Processing*, Vol 1, 297 – 312.
- Gotway, C. A. and Young, L. J. (2002), "Combining Incompatible Spatial Data", *Journal of the American Statistical Association*, Vol 94, No 458, 632 – 48.
- Gribov, A., Krivoruchko, K. and Ver Hoef, J. M. (2004). "Modified weighted least squares semivariogram and covariance model fitting algorithm", *Stochastic Modelling and Geostatistics. AAPG Computer Applications in Geology*, Vol 2, In press.
- Langford, M. Maguire, D. J. and Unwin, D. J. (1991), "The areal interpolation problem; estimating population using remote sensing in a GIS framework", *Handling Geographical Information: Methodology and Potential Applications*, Longman: London, 55 – 77.
- Langford, M. and Unwin, D. J. (1994), "Generating and mapping population density surfaces within a geographical information system", *The Cartographic Journal*, Vol 31, 21 – 26.
- Mennis, J. (2003), "Generating Surface Models of Population Using Dasymetric Mapping", *The Professional Geographer*, Vol 55, No 1, 31 – 42.
- Mugglin, A. S., Carlin, B. P., Zhu, L., and Conlon, E. (1999), "Bayesian areal interpolation, estimation, and smoothing: an inferential approach for geographic information systems",

- Environment and Planning A*, Vol 31, 1337 – 52.
- Mugglin, A. S., Carlin, B. P. and Gelfand, A. E. (2000), “Fully Model-Based Approaches for Spatially Misaligned Data”, *Journal of the American Statistical Association*, Vol 95, No 451, 877 – 87.
- Openshaw, S. (1984), *The Modifiable Areal Unit Problem*, Norwich, U.K.: Geobooks.
- Peuquet, D. J. (1988), “Representations of geographic space: toward a conceptual synthesis”, *Annals of the Association of American Geographers*, Vol 78, 375 – 94.
- Rase, W. D. (2001), “Volume-preserving interpolation of a smooth surface from polygon-related data”, *Journal of Geographical Systems*, Vol 3, 199 – 213.
- Robinson, A. (1971), “The Genealogy of the Isopleth”, *The Cartographic Journal*, Vol 8, 49 – 53.
- Tobler, W. R. (1979), “Smooth Pycnophylactic Interpolation for Geographic Regions”, *Journal of the American Statistical Association*, Vol 74, No 367, 519 – 30.
- Tobler, W. R. (1994), Source code for pycnophylactic spatial interpolation, MS QuickBasic version.
- Wackernagel, H. (2003), *Multivariate Geostatistics: An Introduction with Applications Third Edition*. Springer: Berlin.
- SEER Special Project #08: Development of High Resolution Population Distribution Data to Enhance Cancer Prevention and Control Research
http://www.uiowa.edu/~gishlth/UIORN/2_SEER_reportl_ornl_0304.pdf
- Census 2000, Evaluation C.1, 2003
http://www.census.gov/pred/www/rpts/eval_top_rpts.htm
- ArcInfo 8.20 online Help
- MSDN online library